United States Academic Decathlon®

Indian Hills High School - Oakland, NJ

ECONOMICS

AN INTRODUCTION TO ECONOMICS AND TECHNOLOGY, INNOVATION, AND THE ECONOMY

RESOURCE GUIDE

2023-2024

The vision of the United States Academic Decathlon® is to provide students the opportunity to excel academically through team competition.

Toll Free: 866-511-USAD (8723) · Direct: 712-326-9589 · Fax: 712-366-3701 · Email: info@usad.org · Website: www.usad.org

Table of Contents

Introduction

For well over two hundred years, the field of economics has studied how human societies organize themselves to transform their available resources into the goods and services that their members wish to consume. The outlines of modern economic analysis were already apparent in Adam's Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations*, published in 1776, but discussion of topics relevant to economics can be found even earlier in the writings of Aristotle.

At its core, economics is concerned with how individuals make choices and how these individual decisions and actions interact with one another to determine what happens at the level of the entire economy. Modern economics approaches this problem from several directions. Whereas microeconomics begins with the analysis of individual decisions and then explores how these individual decisions are coordinated through market transactions, macroeconomics begins by considering aspects of the behavior of entire economies and develops models that help make sense of these observed phenomena. Although these two branches of economic analysis start from different points, they are unified by a set of fundamental assumptions about human behavior.

This resource guide begins by describing the basic assumptions on which all economic analysis rests. The list of these assumptions is relatively short, and, as you will see, they are not terribly controversial. Yet, these assumptions provide the basis for the development of an extremely rich and flexible set of theories that can account for a wide range of observed phenomena.

In the second and third sections of the resource guide, we describe some of the most important themes in economics. The second section provides a description of microeconomics. This section starts with the model of perfectly competitive markets. Although the assumptions of this model apply precisely to only a small subset of economic activity, it is a crucial starting point. In the remainder of the section, we show how relaxing the assumptions of the perfectly competitive model allows us to analyze a much broader range of phenomena, and how this analysis in turn leads to important insights about public policy and individual actions.

The third section of the resource guide turns to the subject of macroeconomics. It begins by describing important characteristics of aggregate economic performance and how these characteristics are measured. It then lays out a framework for understanding differences over time and across countries in the quantity of output produced by economies and for understanding short-run fluctuations in economic activity.

In the fourth and final section of this resource guide, we employ some of the conceptual tools developed in the first three sections to examine the economics of technology and innovation.

NOTE TO STUDENTS: You will notice as you read through the resource guide that some key terms and phrases are boldfaced. While many of these terms are defined and/or explained in the text of the guide, you can also find explanations of these terms in the glossary at the end of the resource guide.

SECTION 1
Fundamental Economic Concepts

> *It is not from the benevolence of the butcher, the brewer or the baker, that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages.*
>
> —*Adam Smith,* An Inquiry into the Nature and Causes of the Wealth of Nations

Economics is about everyday life, about the choices each of us makes, and how these choices affect our neighbors, our community, our nation, and our world. Looking at these choices from the perspective of economics helps to illuminate hidden wonders in the everyday world around us. For example, the next time you stop at the supermarket to pick up a loaf of bread on your way home, pause for a minute to reflect on your surroundings. If your supermarket is like most, there will be rows of fresh produce, aisles of baked goods, shelves full of laundry detergent, cases of frozen foods and dairy products, and many other items. In fact, the average supermarket carries more than 33,000 different items.[1]

That each of these items is on the shelf is the result of a complicated chain of decisions by an almost uncountable number of different people. For example, for a loaf of bread to reach the store, a farmer had to decide to grow the wheat, a milling company had to purchase the wheat and grind it into flour, a bakery had to purchase the flour along with other ingredients and then combine them to produce the loaf, and finally this perishable product had to be delivered in a timely fashion to the store. Each product has a similar story.

When you go to the store, you expect to be able to find the bread and all the other products your supermarket carries; but what insures that all of them will be there, as they almost always are? No one ordered the farmer to grow wheat, or the baker to bake bread; they didn't take these actions so that you could stop to pick up a loaf of bread on the way home; they did what they did because it was in their own best interest. Yet somehow, almost magically, all of these individual choices were coordinated so that when you arrive at the store there is an entire aisle of different types of bread available for you to choose from.

Now step back and consider the fact that the store you are in is only one of thousands of supermarkets across the country, and that the supermarket is only one of the many millions of businesses that make up our economy. Many people take all of this for granted, but as the example of less developed countries around the world makes clear, there is nothing automatic or inevitable about how well our economy functions. Economics can help us to understand both why our economy functions smoothly most of the time, and why it occasionally breaks down.

BASIC ASSUMPTIONS OF ECONOMICS

Economics is the study of how individuals make choices about how to allocate scarce resources in order to satisfy virtually unlimited human wants and about how individuals interact with one another. While economists study a vast range of different behaviors, their work is unified by their reliance on a few seemingly simple, yet remarkably powerful assumptions.

Scarcity

Scarcity is an inescapable fact of human existence. There are only twenty-four hours in the day to devote to work, study, play, sleep, and other essential activities. No matter how wealthy a society is, the amount of work, energy, knowledge, and **capital** available to produce the goods and services people wish to consume is limited. On the other hand, our

desires are insatiable. Just as families must choose how much income to spend on food, clothing, vacation travel, and **savings** for retirement, societies face choices about how much of their resources to devote to healthcare, national defense, and education.

Trade-offs

Scarcity implies that every choice we make requires us to give up something to get something else. If you decide to spend an hour watching television, then that is one less hour you have available to study. Similarly, if you choose to spend $10 to go to a movie, then you have $10 less to spend on video games or to save for college expenses.

Opportunity Cost

The cost of what you choose is what you have to give up to get it. Economists call what you give up the "**opportunity cost**" of your choice. It is important to note that the opportunity cost is not necessarily the same as the monetary price you pay. For example, suppose a friend offers you a free ticket to a baseball game. You may not have to pay for the ticket, but the opportunity cost of attending the game is the value of what you would have been doing during that time if you had not gone to the game. For example, if you had been planning to work mowing lawns, the opportunity cost of this choice is the income from mowing that you would forego by attending the game.

Opportunity cost is a seemingly simple concept but applying it can sometimes be rather tricky. Consider the cost of attending college. It might seem obvious that the cost of attending college is the sum of the price of tuition, books, room and board. But this answer excludes an important cost of attending college. For most people, the biggest cost of attending college is the value of their time. By choosing to attend class and do homework, you are giving up time that could otherwise be spent working for pay. At the same time, the explicit monetary costs of attending college may overstate the true expense. Even if you did not attend college, you would still need to eat and have someplace to live. Thus, the costs of room and board are not really part of the cost of college.

Rationality

Economics assumes that people make choices by comparing the benefits of each action with the opportunity costs of that action and then select the action that produces the greatest benefit. It is important to note that the benefits can be interpreted broadly. Many people care a great deal about social issues—such as reducing pollution or helping those less fortunate than themselves. Such concerns are entirely consistent with rational decision-making or **rationality**.

Most of the time, people perform this cost-benefit calculation intuitively and approximately. In the same way that a basketball player does not stop to calculate the physics behind a perfect three-point shot, rational people acquire a feel for what the costs and benefits of their actions will be. Just as some of us are better at hitting three-point shots than others, we are not born with the ability to infallibly calculate costs and benefits. One of the rewards of studying economics is that it helps us to become better decision-makers.

Gains from Trade

Individuals differ in their abilities, interests, and resources. As a result, we all are better at and get more pleasure from some activities than others. By specializing in the things we like and do the best, and then trading with other people who have different abilities, both we and they can then be better off. As long as the exchange is voluntary, then the benefits must outweigh the costs for both of the people involved.

MODELS AND ECONOMIC THEORY

Economic analysis relies on careful observation, description, and measurement of economic activity. But it also relies on theory. To understand how the economic phenomena we observe fit together, it is necessary to build theoretical models that capture the essential features of these interactions while stripping away the unnecessary details. Models come in a wide variety of forms and can be expressed in many different ways.

In economics, models most often consist of diagrams or mathematical formulas. At first glance, many of these models may appear hopelessly simplistic. But the test of a model is in how well it captures the aspects of reality that we are seeking to understand. The simplicity and lack of realism of many of these models is what allows us to identify so clearly what assumptions and characteristics are important.

POSITIVE AND NORMATIVE ECONOMICS

The insights that economics offers about individual and social decisions can be used in two ways. **Positive economics** uses the tools of economic analysis to describe and explain economic phenomena and to make predictions about what will happen under particular circumstances. It focuses on identifying cause-and-effect relationships and measuring their size. For example, positive economics tells us how much we might expect the **consumption** of gasoline to decrease when the price of gasoline increases. In this sense, positive economics is essentially value free. It does not require that the economic analyst express any opinion about the relative merits of different choices.

Normative economics is the term used to describe the use of economic analysis to guide decisions about what *should* be as opposed to what *is* the case. Normative economic statements combine economic analysis with value judgments about the relative merits of different possible economic outcomes. The tools of economic analysis, such as cost-benefit comparisons, can help to structure a discussion of different possible outcomes. But, choices between these outcomes usually require us to refer to criteria beyond the scope of economic theory to justify our particular choices.

To better illustrate this, let's consider the debate about whether to increase the minimum wage. Positive economics can help identify the way in which such an increase would affect different groups as well as provide estimates of their size. In addition to recognizing that a hike in the minimum wage would increase the incomes of those workers who hold minimum-wage jobs, it is important to also note that higher wages may result in some minimum-wage workers losing their jobs. Moreover, others who are seeking employment in jobs covered by the law may be unable to find employment. Finally, employers who have to pay higher wages may see their profits diminish, and they may pass some of the costs on to consumers, who will see the prices of goods and services that depend on minimum-wage workers increase. As this list suggests, an increase in the minimum wage will benefit some people and hurt others. To decide whether the benefits outweigh the costs requires a value judgment about the relative ranking of these effects on the different groups affected by the legislation.

EFFICIENCY AS A GOAL

An important criterion that economists often apply in evaluating a society's use of scarce resources is the efficiency of the resulting allocation. Given any particular outcome, economists would say that it was efficient if there is no way to improve at least one person's well-being without reducing the well-being of someone else. This criterion is called **Pareto efficiency**, after the Italian economist Vilfredo Pareto (1848–1923), who was the first to make use of this concept.

Notice that Pareto efficiency can characterize a wide range of different economic outcomes. Consider, for example, an economy with ten people that produces $100 worth of goods and services. If each citizen receives $9 of benefits and $10 of production is wasted, then this outcome is not Pareto efficient. Redistributing the $10 would make at least some of the citizens better off without making any of them worse off. On the other hand, a situation in which each citizen receives $10 is Pareto efficient; there is no way to increase the well-being of any citizen without reducing the benefits of another.

However, an outcome in which one citizen receives $91 of income and each of the other nine citizens receives $1 is also efficient by the Pareto criterion. The only way to make anyone better off is through redistribution. Pareto efficiency does not provide a basis for choosing between these alternative efficient distributions of benefits. Which distribution is best is, from the perspective of economic analysis, a normative judgment that rests on criteria outside the realm of positive economics. While economic theory does not provide a basis for such choices, economists often offer such value judgments along with their positive analysis.

Despite this limitation, efficiency is an important first step in maximizing overall well-being. When we make decisions about how to allocate resources, it is important that we do so in a way that does not waste any of them.

MICROECONOMICS AND MACROECONOMICS

The tools of economic analysis can be used to study a wide array of phenomena, ranging from how individuals and businesses make decisions, to how they interact in markets, on up to the factors that determine the overall level of production, employment, and the price level of national economies.

The field of economics is traditionally divided into two broad subfields: microeconomics and macroeconomics. Microeconomics concentrates on individual behavior and the operation of particular markets. Macroeconomics concentrates on the overall performance of the national economy.

Clearly microeconomics and macroeconomics are closely linked. They share common assumptions about the basic features of human behavior. But, because they focus on economic activity on different scales, different aspects of this behavior are important. And, their modes of analysis are sufficiently different, so it is useful to consider them separately.

SECTION I SUMMARY

- Economics is the study of how individuals make choices about how to allocate and distribute scarce resources and how they interact with one another.
- Scarcity is inescapable because resources are limited and human desires are insatiable.
- Every choice we make involves trade-offs. The opportunity cost of what we choose is what we must give up by making that choice.
- Economics assumes that people make choices rationally by comparing the benefits and opportunity costs of each action and selecting the action that yields the greatest net benefit.
- Trade makes everyone involved better off.
- Economic models help us to understand economic phenomena by capturing essential details and eliminating unnecessary details.
- Positive economics uses the tools of economic analysis to describe economic phenomena and make predictions about what will happen under particular circumstances.
- Normative economics uses the tools of economic analysis to evaluate the relative merits of different situations.
- Pareto efficiency is an important criterion in economics. It describes a situation in which the only way that anyone can be made better off is by reducing the well-being of one or more other people.
- The two main branches of economics are microeconomics and macroeconomics.

SECTION II
Microeconomics

As the example of the supermarket discussed earlier illustrates, our modern economy achieves a high degree of coordination. The mechanism that produces this coordination is the interaction of supply and demand within markets. Within markets, the actions of buyers and sellers determine the price at which each product or service sells and the quantity that changes hands. Individual buyers and sellers respond to market prices in predictable ways.

The interaction of supply and demand in markets is the central topic of microeconomics. Our starting point is to develop an understanding of the behavior of perfectly competitive markets. We will begin by defining what we mean by a market, and then we will describe in more detail how supply and demand are determined by the self-interested choices of individual market participants. Although the assumptions of perfect competition may seem unrealistic at first, the resulting model is an essential building block for economic analysis. It is approximately true in many situations and provides an important benchmark against which to compare many other more complicated models.

After developing the model of perfect competition, we will illustrate its usefulness in analyzing a range of important topics, including the effects of taxation and other types of government policies, as well as the costs and benefits of trade. Having explored these applications, we will then begin to introduce additional features necessary to capture a wider range of economic phenomena. In this segment of the resource guide, we will examine a number of different ways in which markets may "fail" to be economically efficient.

We will conclude our discussion of microeconomics with a closer look at the role of government and other forms of collective choice.

PERFECTLY COMPETITIVE MARKETS

Markets

A market is comprised of all of the buyers and sellers of a particular good or service. Some markets, such as the New York Stock Exchange or the Chicago Mercantile Exchange, are highly organized. Buyers and sellers in such markets come together at a single location, and an auctioneer helps to set a price at which exchanges take place.[2]

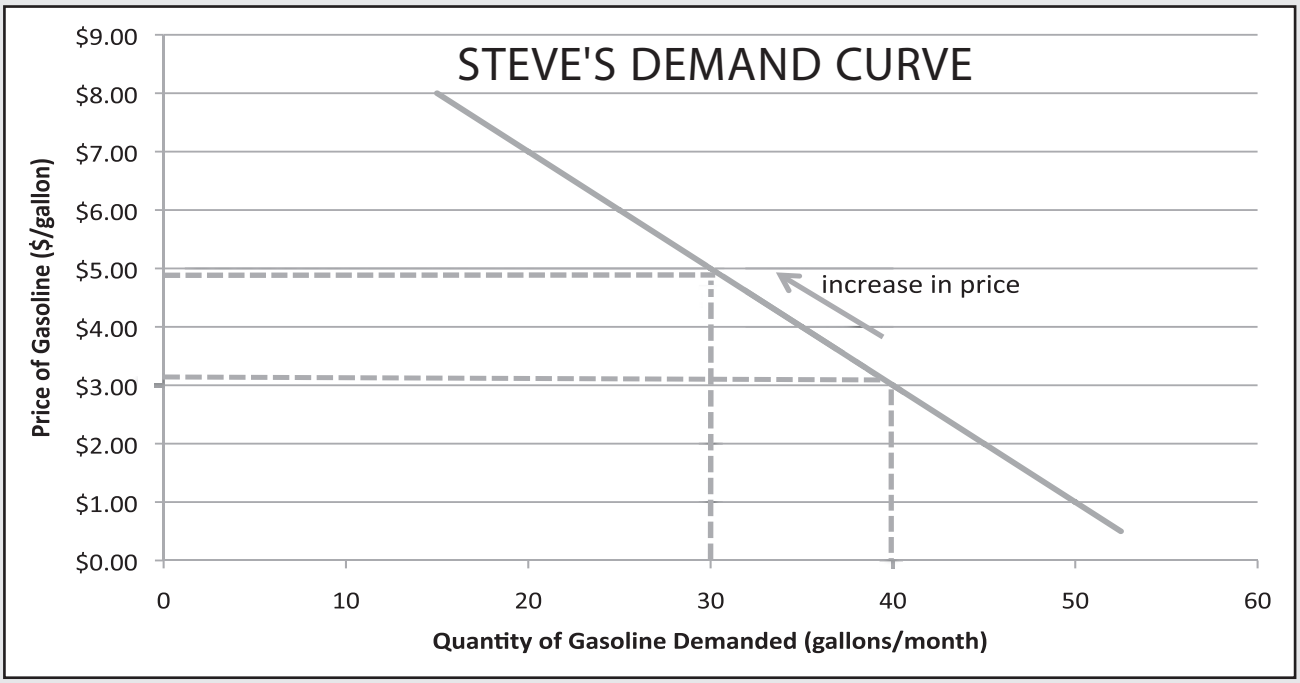
More often, markets are less formal. Nevertheless, we can think of the interaction between buyers and sellers as constituting a market. For example, consider the market for gasoline in your community. The sellers in this market are all the local gas stations in town, while the buyers consist of all the vehicle owners in the community or passing through it. Each of the sellers in this market posts the prices at which they will sell a gallon of gasoline, and buyers will select where to fill their tanks based on price and convenience. The buyers of gasoline are likely to be well informed about prices because gas prices are continually posted at all of the different stations.

The market for gasoline is highly competitive. There are many buyers and sellers even in a relatively small community, and none of these market participants trades more than a small fraction of the gasoline that changes hands. As a result, no one buyer or seller influences the price of gasoline, or the quantity sold. Rather, the price and quantity sold are determined by the combined actions of all the buyers and sellers in the market. The owner of each gas station knows that there are other stations selling a very similar product, so if the owner raises their price above the going price, then customers will go elsewhere. On the other hand, the owner has no reason to lower the price significantly below the going price because this will simply reduce their income.

FIGURE 1

STEVE'S DEMAND SCHEDULE

PRICE	QUANTITY OF GASOLINE DEMANDED
$0.50	52.5
$1.00	50
$2.00	45
$3.00	40
$4.00	35
$5.00	30
$6.00	25
$7.00	20
$8.00	15
$9.00	10



Steve's Demand Schedule and Demand Curve for Gasoline

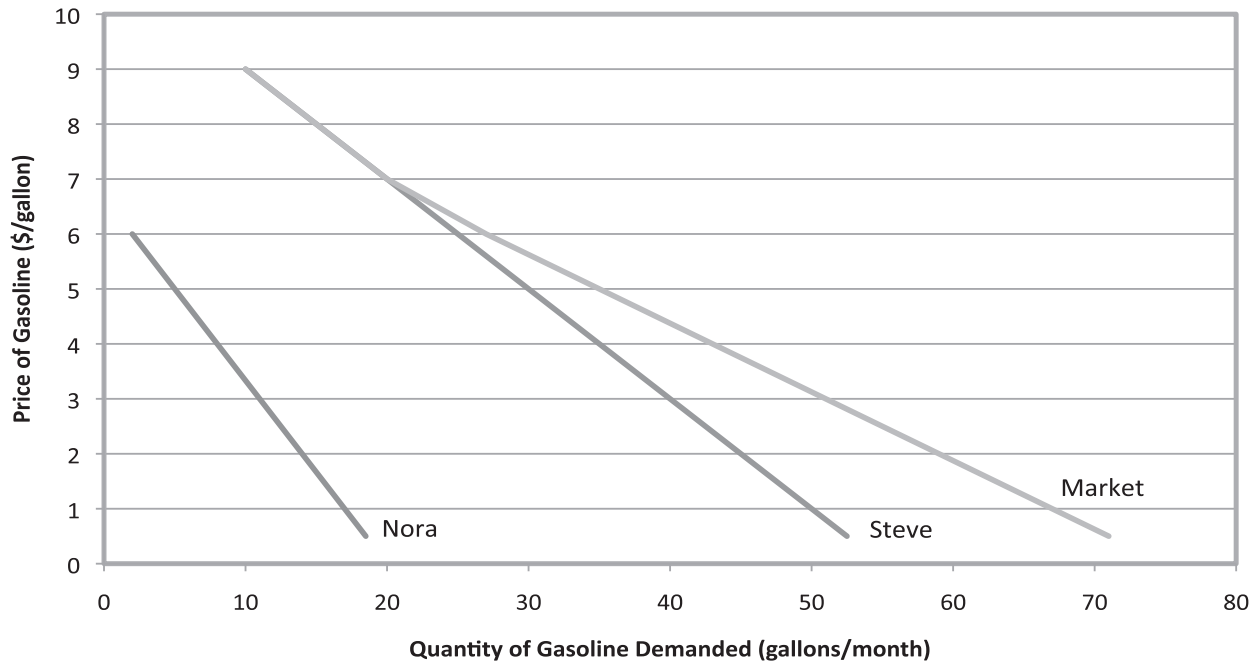
In much the same way, because each buyer purchases only a small amount of gasoline compared to the total market, no one buyer can influence the price.

We say that a market is *perfectly competitive* if the good or service being bought and sold is highly standardized, the number of buyers and sellers is large, and all of the participants are well informed about the market price. In such a market, buyers and sellers know that they can buy or sell as much as they wish without influencing the market price.

While only a few markets precisely conform to the assumptions of perfect competition, many real world markets are characterized by a high degree of competition and can usefully be described in terms of the perfect competition assumption. The market for gasoline is a good example of a nearly competitive

FIGURE 2

PRICE	STEVE		NORA		MARKET
0.5	52.5	+	18.5	=	71
1	50		17		67
2	45		14		59
3	40		11		51
4	35		8		43
5	30		5		35
6	25		2		27
7	20		0		20
8	15		0		15
9	10		0		10



Derivation of Market Demand for Gasoline

market. Unless you live in a very small town, you have probably noticed that the price of gasoline is not precisely the same at different stations. But, the differences in prices are never very large. As a result, many of the lessons we learn from analyzing perfectly competitive markets can be applied to less than perfectly competitive markets. Our analysis of perfect competition will also provide a useful benchmark against which to compare the outcomes of other types of markets.

Demand

The quantity demanded of any good is the amount of that good buyers are willing and able to purchase. This quantity depends on a wide range of factors. One of the most important is the good's price. If the price of the good is higher, buyers will demand less of the good; if the price is lower, then they will demand more. This negative relationship between a good's price and the quantity demanded is called the **law of demand**.

The law of demand is a result of the cost-benefit analysis that rational decision-makers use when deciding how to allocate their resources. As the price of a good increases, the opportunity cost of consuming that good also increases since consumers must cut back on their consumption of other goods to afford the higher price. If, for example, the price of gasoline rises, people will likely find ways to reduce the amount that they drive. They might do this by planning their trips more carefully or choosing to take the bus or ride a bicycle rather than drive.

The table in Figure 1 illustrates how Steve's purchases of gasoline each month depend on the price per gallon. At $1 per gallon, Steve buys 50 gallons; when the price rises to $2 a gallon, he cuts back to 45 gallons. If the price rises further, to $3 a gallon, he cuts back to 40 gallons. This table is called a **demand schedule**.

The graph in Figure 1 shows another way of representing Steve's demand schedule. The downward-sloping line in this graph is called Steve's **demand curve**. Notice that we plot the points of Steve's demand schedule with the quantity demanded on the horizontal axis and the price on the vertical axis. To read this graph, find a price on the vertical axis (say $3 per gallon) and then draw a line horizontally until it intersects the demand curve. Now draw a line vertically downward from that point until it intersects the horizontal axis. The point at which this line intersects the horizontal axis (40 gallons) is the quantity Steve demands when the price is $3 per gallon.

When the market price changes, we find Steve's quantity demanded by moving up or down along the demand curve until we reach the height corresponding to the new market price. For example if the price were to rise from $3 to $5 a gallon, Steve's quantity demanded would decline from 40 gallons a month to 30 gallons a month. This movement is illustrated in Figure 1 by the arrow pointing up and to the left along the demand curve.

Steve is, of course, just one buyer. To find the market demand schedule, we must add up the quantity that every consumer will purchase at each possible price. Figure 2 illustrates how this process works with two individuals. In addition to Steve, the market now includes Nora. The table in Figure 2 shows that the market quantity demanded is the sum of the quantities that Steve and Nora wish to consume at that price. The graph shows that we add the two demand curves *horizontally* to obtain the market demand.

Shifts in the Demand Curve

The market demand curve depicts the relationship between the quantity demanded and its price, assuming that all other factors that might influence the quantity demanded remain unchanged. But many other things can influence the quantity demanded. If one of these factors changes, it causes the entire demand curve to shift.

For example, if your community creates a new system of bicycle lanes that make it easier to bike from place to place, the quantity of gasoline demanded will decline at every price. As Figure 3 shows, such a change causes the market demand curve to shift to the *left*, indicating that at each price a lower quantity is demanded. Let's consider some of the most important factors affecting the quantity demanded.

Income

Suppose Steve's employer reduces his weekly hours of work, and thus his income. Because Steve has less **money** to spend on all the things he wishes to buy, he will likely reduce his consumption of gasoline. For most goods, demand is *positively* related to income: when income rises, the quantity demanded rises, but when income falls, the quantity demanded falls. Goods for which this is true are called **normal goods**.

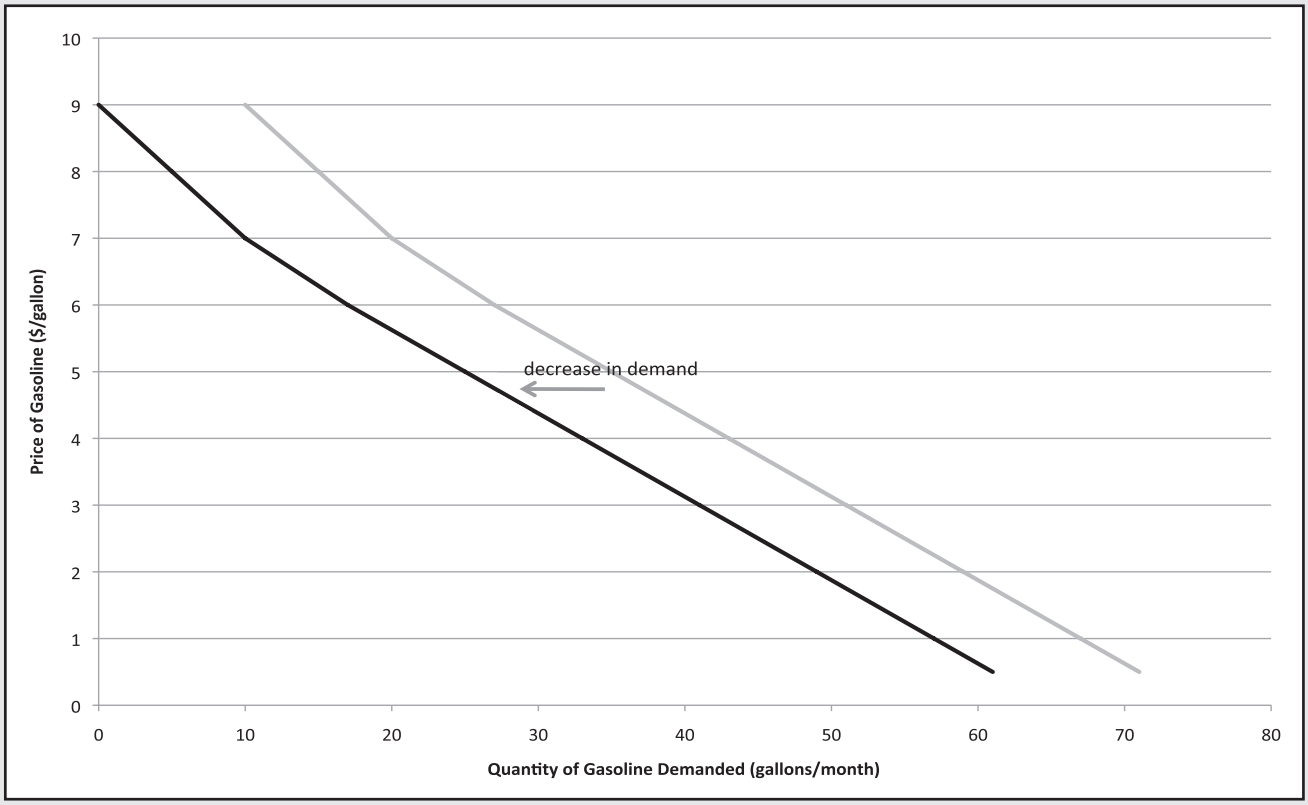
Not all goods are normal goods, however. Goods for which the quantity demanded falls as income rises are called **inferior goods**. Bus rides might be an example of an inferior good. As their income increases, consumers will be more likely to buy a car and drive instead of taking the bus.

The Prices of Related Goods

Suppose that the price of airline tickets falls. The law of demand says that consumers will purchase more airline travel. Because airline travel is to some extent a substitute for travel by car, people will likely reduce the number of miles they drive and hence the quantity of gasoline they demand at any price. When a decline in the price of one good causes a reduction in the quantity demanded of another, we say that these goods are **substitutes**.

Suppose, on the other hand, that the price of automobile insurance falls. Lower insurance costs

FIGURE 3



Effects of a Bike Lane on Demand for Gasoline

make it easier for more people to afford to own automobiles; car ownership will increase and so will the number of miles driven. When a lower price for one good causes demand for another good to increase, we call those two goods **complements**.

Tastes

Remember that the quantity demanded reflects a comparison of the benefits of consumption with the opportunity costs of purchasing the good. If the perceived benefits of consumption change, then so will the quantity demanded. For example, suppose that concerns about the environmental impacts of driving cause people to be more concerned about pollution. The likely impact will be a reduction in the demand for gasoline.

Expectations

Changes that you expect to occur in the future may also affect the quantity demanded. For example, if Steve is afraid that he may lose his job next month, then he might cut back on his driving now in anticipation of this future change in his income.

Number of Buyers

Market demand is derived by adding up the demands of individual consumers. If there are more consumers, then demand will increase. If your community is growing because people and businesses are moving there, then the market demand for gasoline will be increasing with this growing population.

Supply

The quantity supplied of any good is the amount that sellers of that good are willing and able to produce. Many factors influence the quantity supplied, but the most important is the price that suppliers receive. The higher the price is, the greater the quantity that suppliers will want to produce. This *positive* relation between price and quantity supplied is called the **law of supply**.

The positive relationship between price and quantity

supplied reflects the cost-benefit analysis of rational suppliers. Gasoline station owners compare the benefits of each gallon sold to the opportunity cost of their time, effort, and expense to supply that gallon of gasoline. As the price rises, it will be rational to devote more resources to supplying gasoline. So long as the price they receive exceeds their opportunity cost, they will be willing to supply gasoline. At higher prices, they will be willing to work longer hours, hire additional help, and expand the size of their stations to boost sales. At lower prices, they will cut back on the time they spend supplying gasoline, reduce the number of their employees, or shift their efforts toward selling other products.

Figure 4 illustrates the relationship between price and quantity supplied for Shelly. Again, we plot the price of gasoline on the vertical axis and the quantity supplied on the horizontal axis. Shelly's **supply curve** is upward sloping, reflecting the positive relationship between price and quantity supplied.

The market supply curve is obtained by adding the quantities supplied at each price by all of the suppliers in the market. This is illustrated in Figure 5 for the case where there are two suppliers. Again, we obtain the market supply curve by adding the individual supply curves horizontally.

Shifts in the Supply Curve

The market supply curve shows the quantity supplied at each price, assuming that all other things remain unchanged. There are, however, many other factors that will influence the quantity supplied. A change in any of these factors will cause the supply curve to shift. Let's consider some of the most important factors that might cause the supply curve to shift.

Input Prices

Inputs are any of the things that suppliers have to purchase to supply a product. For example, the price that gasoline stations must pay their suppliers for gasoline is a major cost of doing business. If this price falls, the quantity of gasoline supplied will increase, causing the supply curve to shift to the right. But, there are other inputs that are important as well. These include labor costs, the real estate costs for the land on which the gasoline station is located, and utilities such as electricity. If any of these input costs increases, it will decrease the quantity supplied at every price, causing the entire supply curve to shift to the left.

Technology

Changes in **technology** can affect how businesses operate and hence the quantity supplied. In the case of gasoline, the shift from full-service to self-service reduces labor costs and increases the quantity supplied. Similarly, pumps with credit card readers further reduce labor costs and increase the quantity supplied.

Expectations

If suppliers expect prices to rise in the future, then they may reduce the quantity they will supply today and store current inventory in expectation of the higher future prices.

Number of Sellers

As more sellers enter the market, the quantity supplied will increase. On the other hand, if a seller decides to leave the market, then the quantity supplied will be reduced.

Equilibrium

What will the price of gasoline be? How many gallons will be sold? To answer these questions we need to put the information about the market demand and market supply together. There is, as we will see, only one combination of price and quantity at which the market is at **equilibrium**, and it is at this point that the market will settle.

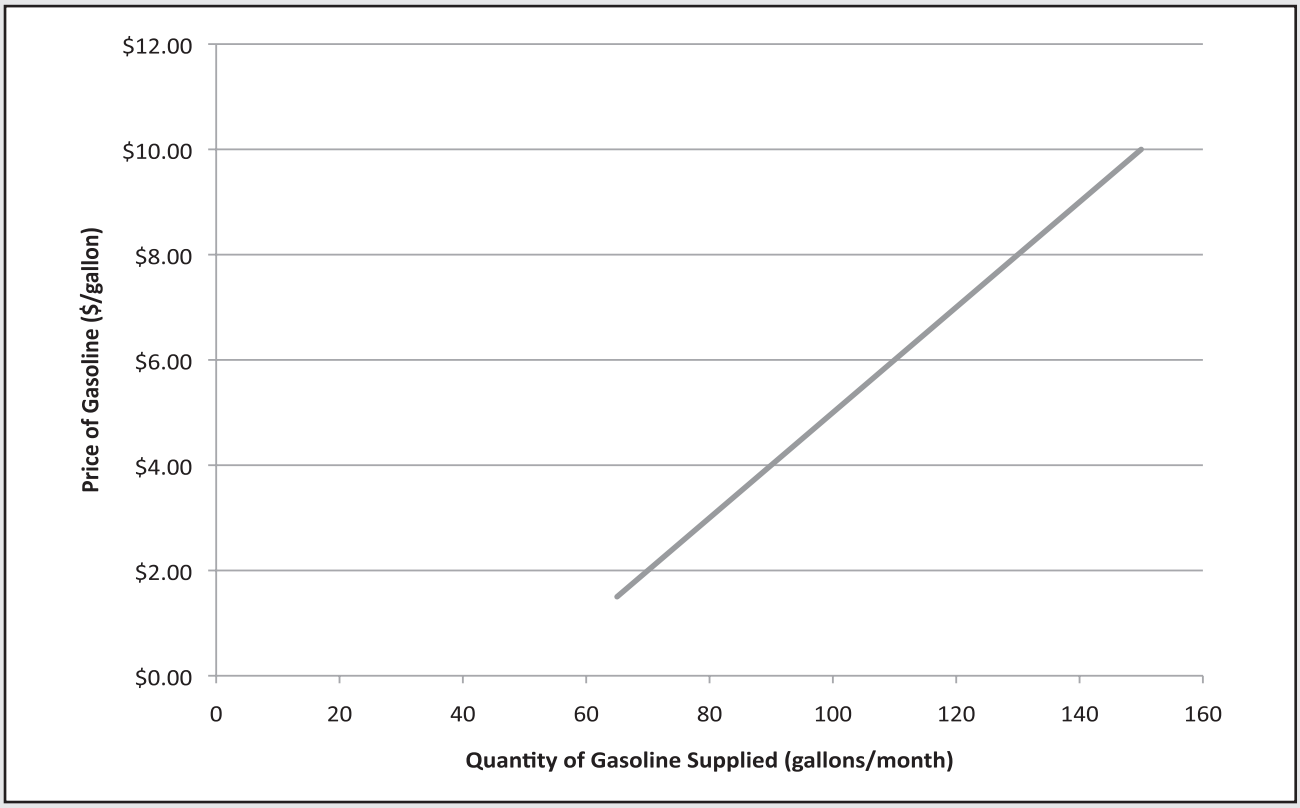
Equilibrium is a widely used concept in both the physical and social sciences. It is defined as a point at which all the forces at work in a system are balanced by other forces, resulting in a stable and unchanging situation. In economics, a market is in equilibrium when no participant in the market has any reason to alter their behavior.

The market equilibrium occurs at the combination of price and quantity where the market supply and demand curves intersect. Because the supply curve is upward sloping and the demand curve is downward sloping, there is only one possible point of intersection. Figure 6 illustrates the market equilibrium for gasoline. In this hypothetical example, the equilibrium price is $2.50, and the equilibrium quantity is 10,000 gallons of gasoline per month.

At this point, we can say that the buyers and sellers in this market are all satisfied, in the sense that buyers are able to purchase as much gasoline as they would like at a price of $2.50 a gallon, and suppliers can sell as

FIGURE 4

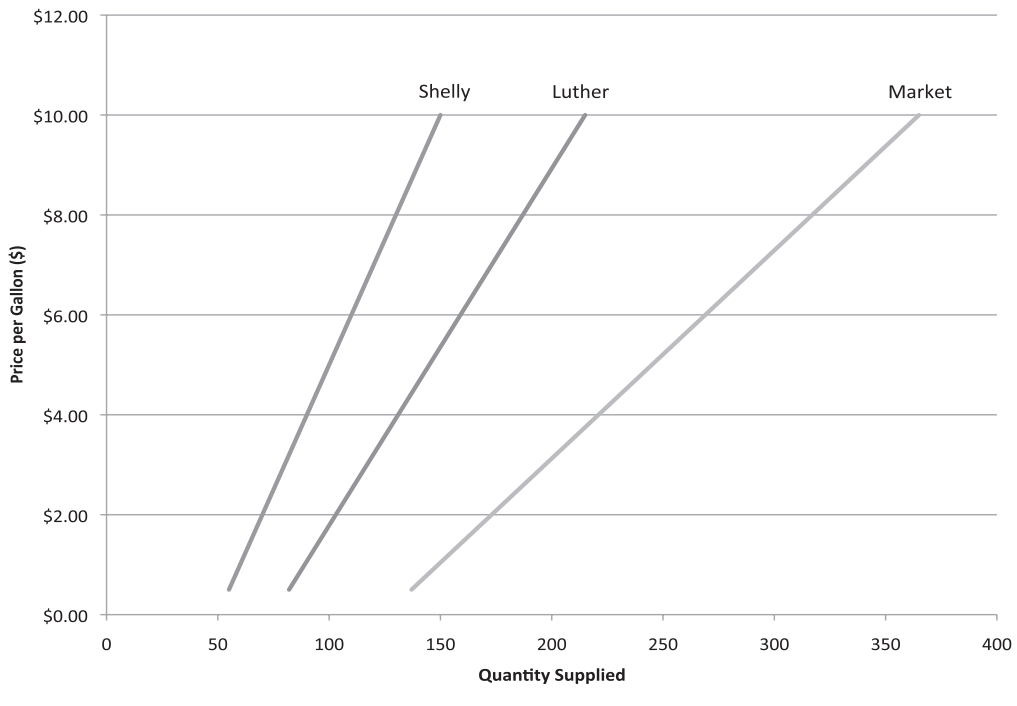
PRICE OF A GALLON OF GASOLINE	QUANTITY OF GASOLINE SUPPLIED
$1.50	65
$2.00	70
$2.50	75
$3.00	80
$3.50	85
$4.00	90
$4.50	95
$5.00	100
$5.50	105
$6.00	110
$6.50	115
$7.00	120
$7.50	125
$8.00	130
$8.50	135
$9.00	140
$9.50	145
$10.00	150



*Shelly's **Supply Schedule** and Supply Curve*

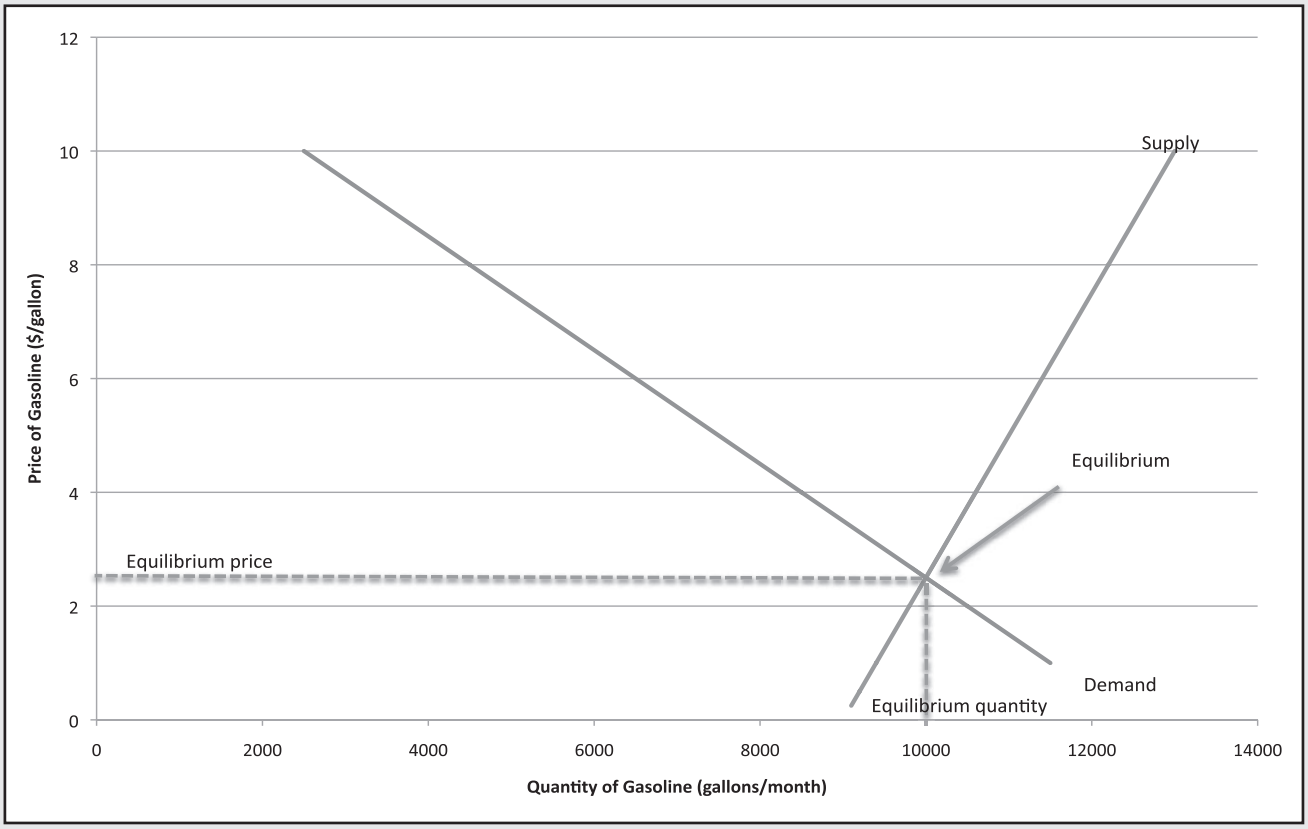
FIGURE 5

PRICE OF A GALLON OF GASOLINE	SHELLY'S QUANTITY SUPPLIED		LUTHER'S QUANTITY SUPPLIED		MARKET QUANTITY SUPPLIED
$0.50	55	+	82	=	137
$1.00	60		89		149
$1.50	65		96		161
$2.00	70		103		173
$2.50	75		110		185
$3.00	80		117		197
$3.50	85		124		209
$4.00	90		131		221
$4.50	95		138		233
$5.00	100		145		245
$5.50	105		152		257
$6.00	110		159		269
$6.50	115		166		281
$7.00	120		173		293
$7.50	125		180		305
$8.00	130		187		317
$8.50	135		194		329
$9.00	140		201		341
$9.50	145		208		353
$10.00	150		215		365



Derivation of the Market Supply Curve

FIGURE 6



Market Equilibrium

much gasoline as they would like at this price. There are, no doubt, buyers who complain that the price of gasoline is too high and would like the price to be lower, and similarly suppliers who complain that the price is too low and would like it to be higher.

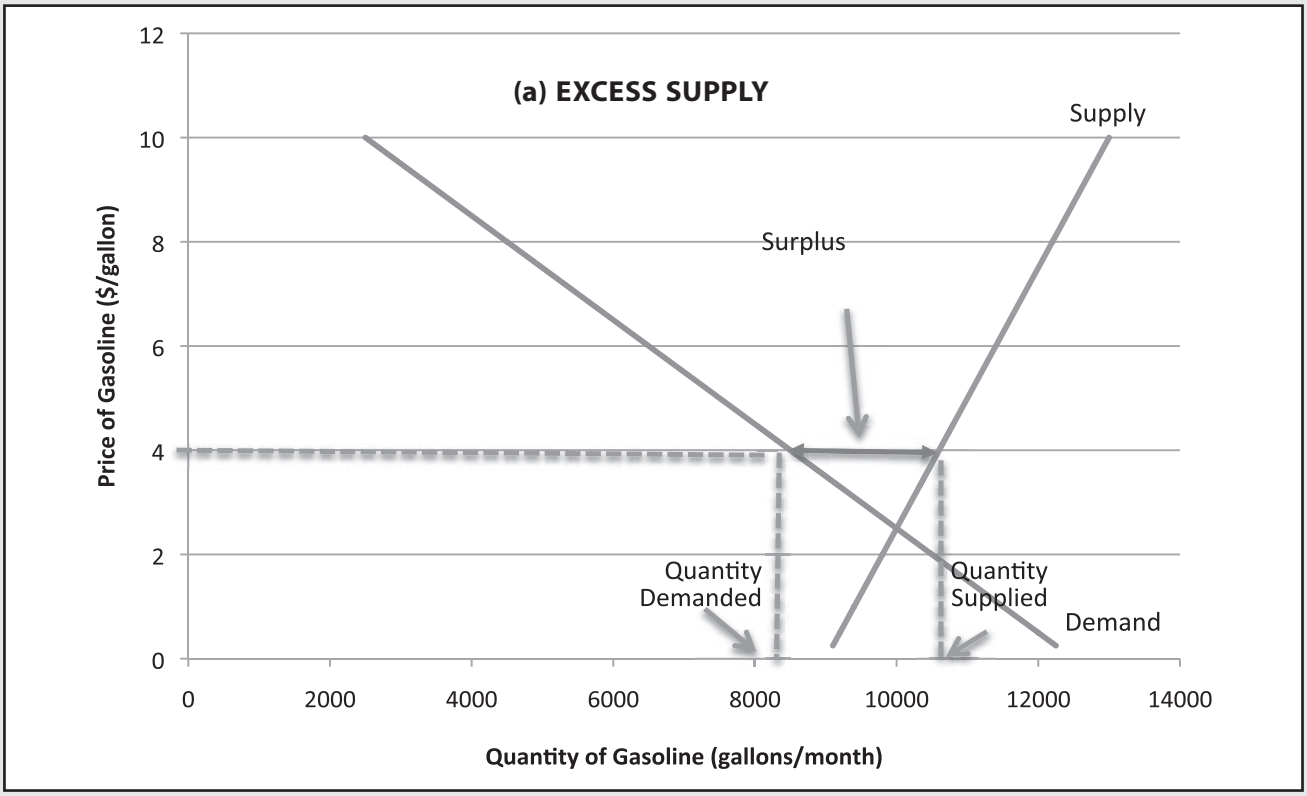
An important feature of market equilibrium is that the market has an automatic tendency to gravitate toward this combination of price and quantity. Figure 7 illustrates this point. We start (Figure 7a) by supposing that the price is higher than $2.50. At a price of $4 a gallon, for example, suppliers would like to sell 10,600 gallons, but buyers only wish to purchase 8,500 gallons a month. In other words, there is an *excess supply.* No one can force people to buy more gasoline than they want. Suppliers will find that they have too much gasoline on hand, their storage tanks are filling up, and they cannot unload their inventory.

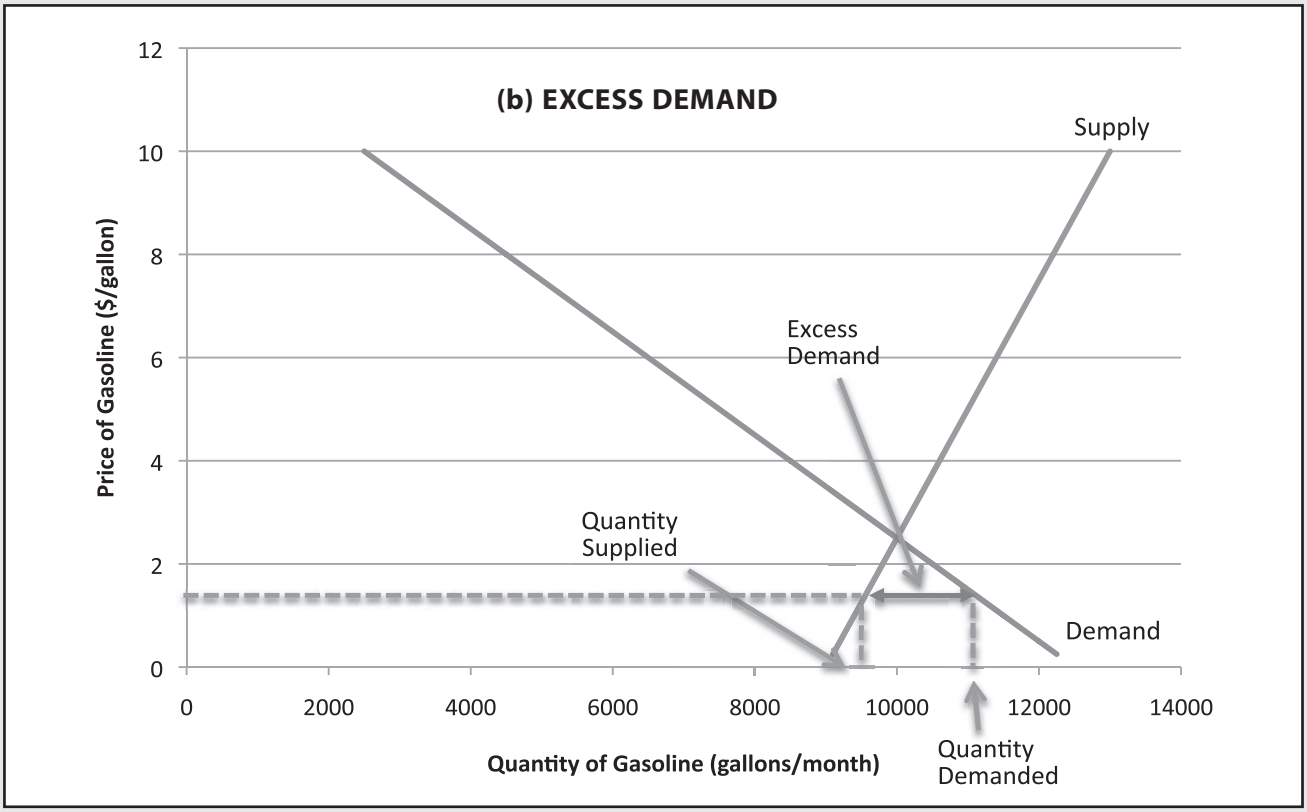
Under these circumstances, suppliers have an incentive to lower their price a little bit. If one station posts a price of $3.90 a gallon, it will attract buyers from other stations, and its surplus will be reduced. But once the other stations see that they are losing customers, they will be forced to lower their prices as well. The pressure to cut prices and attract business will not go away until the price has reached the equilibrium level of $2.50 a gallon.

Now suppose that the price is below the equilibrium price. Figure 7b illustrates this situation. At a price of $1.50 there is an *excess demand* for gasoline. Buyers wish to purchase 11,000 gallons of gasoline, but suppliers are willing to sell only 9,600 gallons. Now there are shortages: some drivers cannot find any gasoline, and others have to wait in long lines to purchase gasoline.

Buyers might be tempted to offer to pay a little bit extra to be sure to get what they need, and sellers will see they can raise prices without sacrificing sales. The pressure to raise prices will continue until the price has reached

FIGURE 7





Markets out of Equilibrium

the equilibrium level. Only at this point will buyers and sellers have no desire to change their behavior.

The Characteristics of Competitive Market Equilibrium

Competitive markets tend to gravitate toward the equilibrium quantity and price. This is a very important feature of markets and has several desirable consequences. First, competitive markets are an extremely effective method of allocating resources. When the market for a good is in equilibrium, the price conveys important information for potential suppliers about the value consumers place on that good. At the same time, the price informs potential demanders about the opportunity cost of supplying the good. This two-way communication is how markets insure that scarce goods and services are produced at the lowest cost and allocated to the buyers who value them the most highly.

The competitive market equilibrium insures that the available supply goes to those buyers who value the good most highly, and that it is provided by those suppliers who have the lowest costs of supplying the good. This fact leads to the second characteristic of the competitive market equilibrium: it maximizes the benefits buyers and sellers receive from exchange.

Let's begin by considering the benefits buyers receive from participating in the market. The important insight is the height of the market demand curve at each point reveals the marginal buyer's willingness to pay. The marginal buyer is the buyer who, at that price, is just indifferent between buying the good in question or not buying it.

To illustrate this, let's consider the highly simplified example presented in Figure 8. The table lists the amount each of four fans would be willing to pay to purchase a ticket to a Bruce Springsteen concert. The table shows that Barb values attending the concert at $100, and at any price less than that she will purchase a ticket. The other potential buyers place a lower value on attending the concert.

If the concert promoter sets the price of tickets at $60, then Steve will not purchase a ticket, since the most he is willing to pay is $50. The other three consumers will all purchase tickets, but the benefit they receive from being able to purchase the ticket for $60 varies. Barb would have paid $100, so attending the concert produces a benefit valued at $40 for her. Since Bob was willing to pay $80, his benefit is $20, and Sharon's benefit is just $10. Adding these amounts together, we see that the three purchasers receive a combined benefit of $70. We call this amount the **consumer surplus** since it is the surplus value that consumers receive.

The demand curve in Figure 8 slopes downward, indicating that as the price falls, more of the fans will be willing to purchase tickets. At any point along this demand curve, its height shows the marginal purchaser's willingness to pay. Because the height of the demand curve measures buyers' willingness to pay, the difference between the height of the demand curve and a horizontal line drawn at the market price measures the consumer surplus for the marginal buyer at each quantity demanded. More generally, we can use the total area below the demand curve and above the market price as a measure of total consumer surplus. This area, then, provides a monetary measure of how much benefit all of the buyers in a particular market receive from participating in that market.

In the same way the height of the demand curve represents buyers' willingness to pay, the height of the supply curve at each quantity supplied measures the willingness to supply of the marginal seller—that is, the seller who would leave the market if the price were any lower. Put somewhat differently, the height of the supply curve measures the opportunity cost to the marginal seller. If the market price exceeds this opportunity cost, the difference is a monetary measure of what is called the **producer surplus**. And we can measure the combined surplus of all suppliers using the area above the supply curve and below the market price as is illustrated in Figure 9.

Combining consumer surplus and producer surplus provides a measure of the total benefits that market participants receive from their transactions. We call this benefit the **total surplus**. One goal of a benevolent social planner should be to maximize this combined surplus, since this is the outcome that produces the greatest overall good. An outcome that maximizes total surplus satisfies the economist's criterion of Pareto efficiency, since at this point there is no way to make anyone better off without reducing the welfare of someone else.

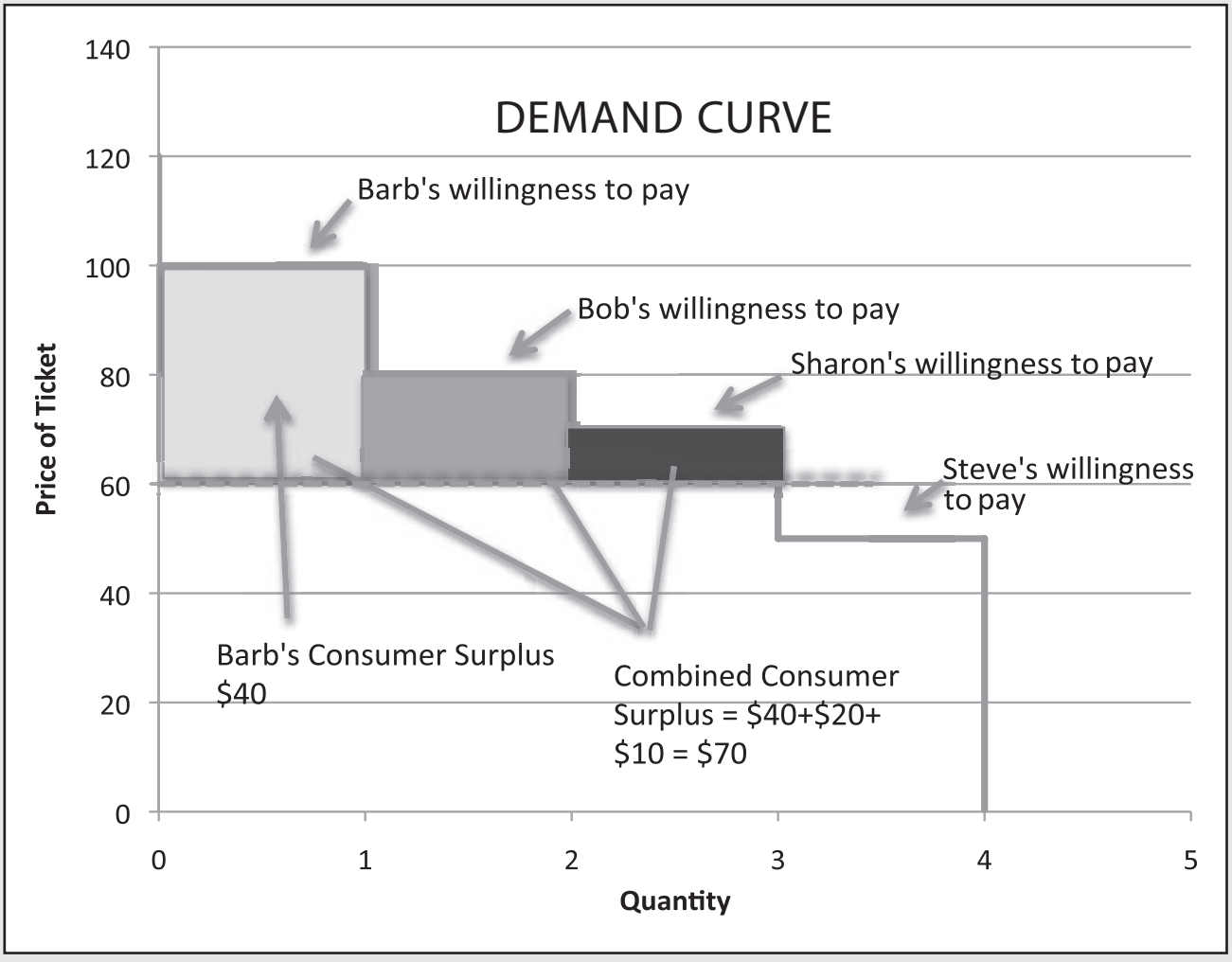
To see that the competitive market equilibrium indeed meets the efficiency criterion and maximizes total surplus, let's consider Figure 10. Suppose first that

FIGURE 8

BUYER	WILLINGNESS TO PAY
Barb	$100
Bob	$80
Sharon	$70
Steve	$50

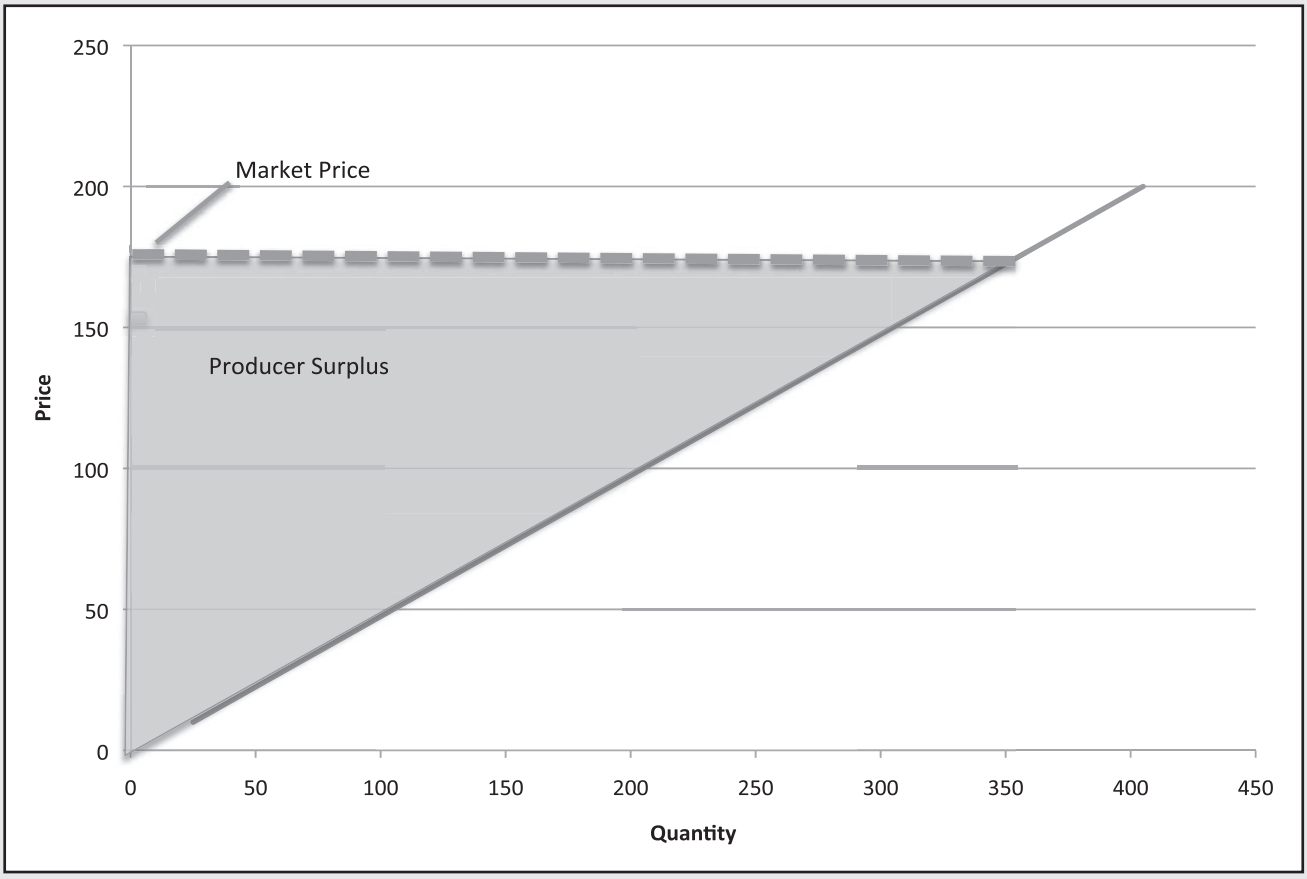
DEMAND SCHEDULE

PRICE	BUYERS	QUANTITY DEMANDED
more than $100	None	0
$80 to $100	Barb	1
$70 to $80	Barb, Bob	2
$50 to $70	Barb, Bob, Sharon	3
$50 or less	Barb, Bob, Sharon, and Steve	4



The Demand Curve Represents Buyers' Willingness to Pay

FIGURE 9



Producer Surplus

a quantity Q_1, which is less than the equilibrium quantity, was exchanged in the market. At this point, the value of the good to buyers exceeds the cost to sellers of supplying the good. A slight increase in the quantity in such a market would yield an increased benefit to both parties. So Q_1, or any other point to the left of the market equilibrium, cannot be efficient. Now, suppose that the quantity traded in the market is Q_2, an amount greater than the equilibrium quantity. At Q_2 the supply curve is above the demand curve, indicating that the cost to producers exceeds the value to consumers. Such an exchange cannot be accomplished voluntarily, but if it did take place, then buyers or sellers would suffer a loss in welfare. Moving to the left would raise overall well-being.

To achieve an efficient outcome, a market planner would need to know the value each consumer places on the good in question, and the cost of producing each unit, and would have to determine how much should be produced, by whom, and to whom it should be given. While such a task would be extremely difficult, a competitive market achieves the same result simply through the self-interested actions of its participants, responding only to the signals provided by the market price.

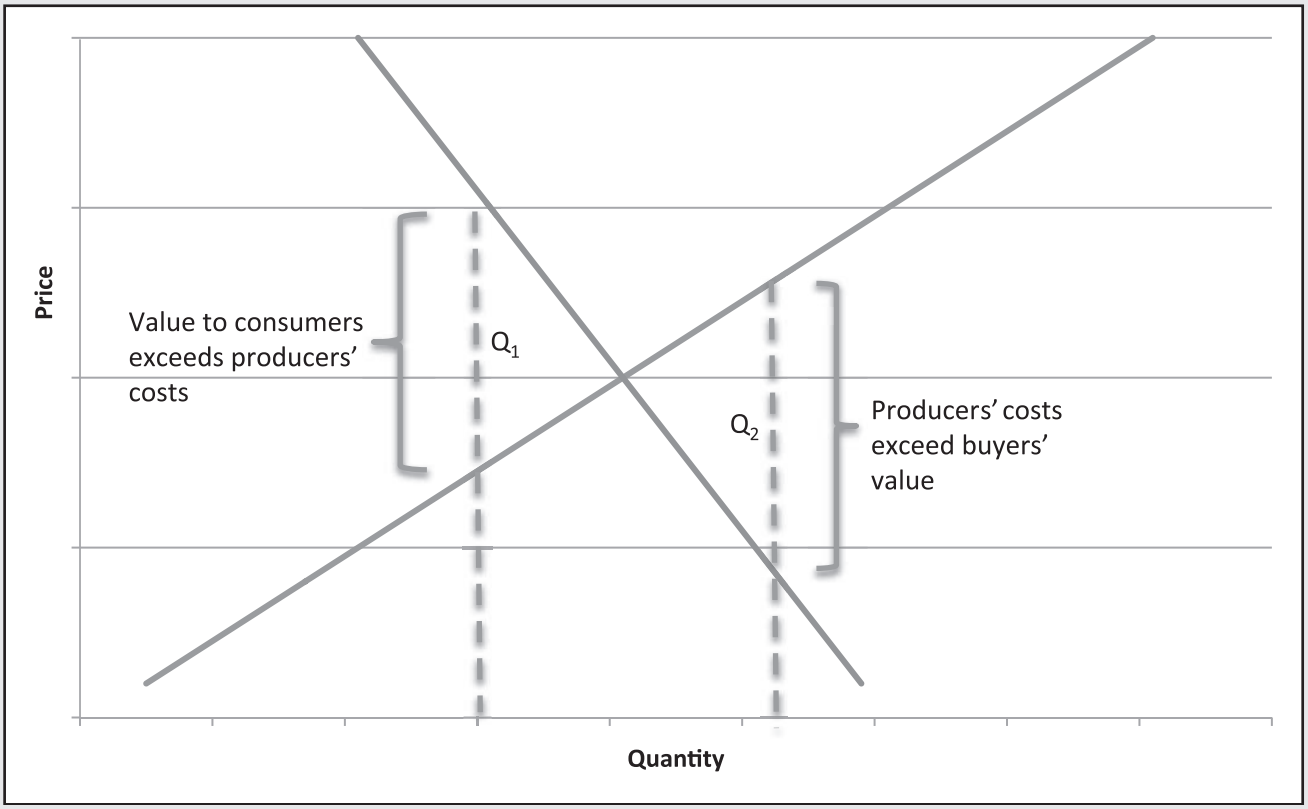
APPLICATIONS OF THE COMPETITIVE MARKET MODEL

Changes in Market Equilibrium

Now that we have seen how to use the concepts of supply and demand to find the equilibrium price and quantity in a competitive market, we can use our market model to make predictions about how shifts in the economy will affect the market. Let's consider some examples illustrating how the competitive market model can be used to analyze important issues.

One of the defining characteristics of our modern

FIGURE 10



Competitive Market Equilibrium Maximizes Social Welfare

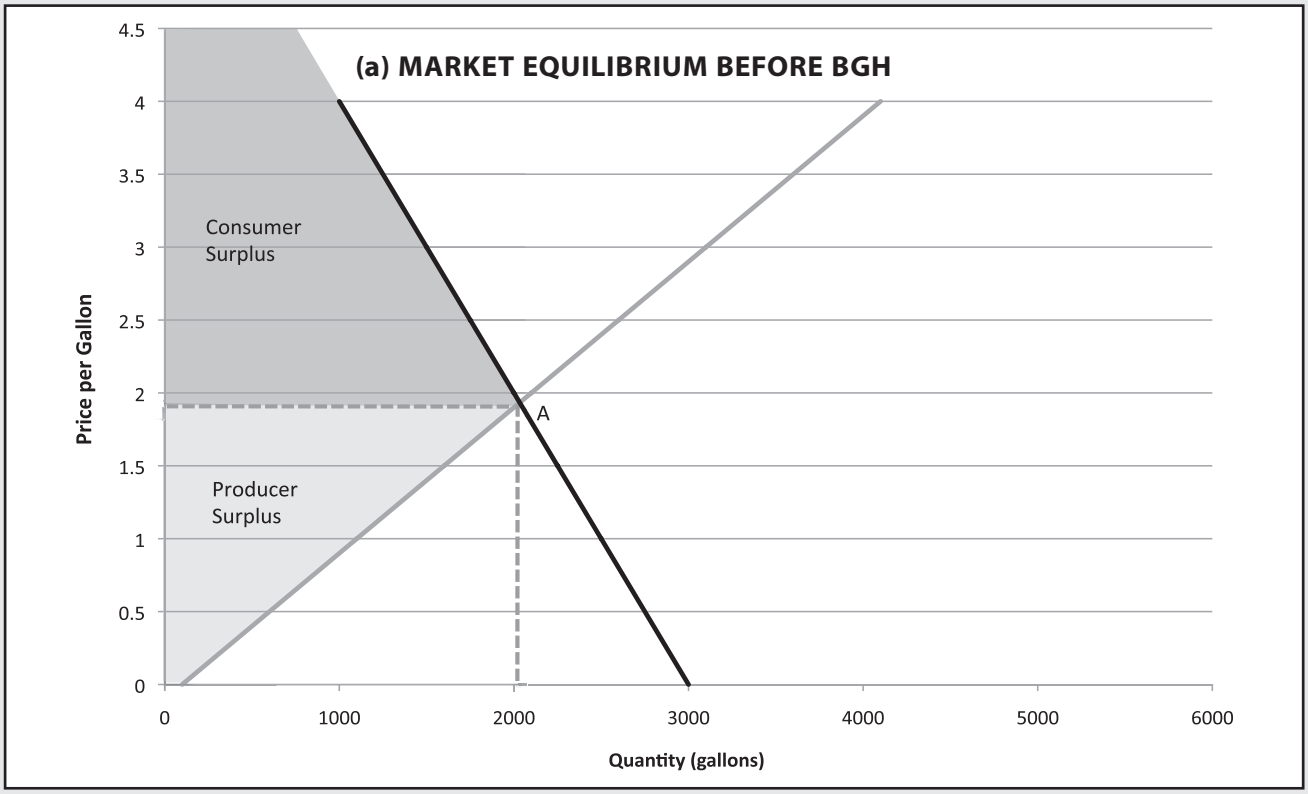
economy is technological progress. New inventions are continually being developed that allow suppliers to produce more at lower costs. One example is the development of synthetic Bovine Growth Hormone (BGH), which allows dairy farmers to increase milk production by between 10 and 15 percent at little additional cost. The direct effects of this innovation are illustrated in Figure 11. As is often the case, the introduction of a new technology has other, more subtle effects, called externalities, that are not immediately obvious from an analysis of the market that is immediately affected.[3] We will discuss how to incorporate externalities into our analysis later in this section of the resource guide.

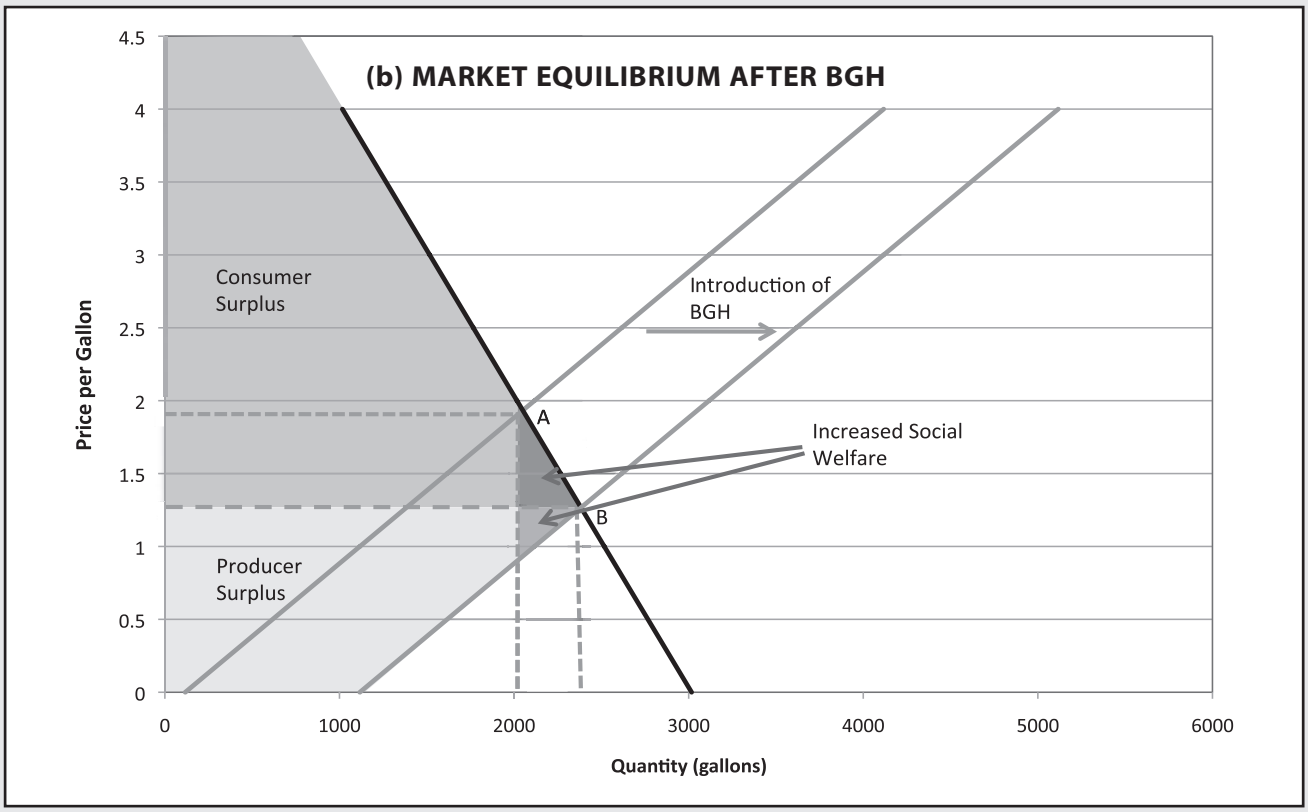
The first panel shows the market equilibrium before the introduction of BGH. The shaded regions indicate the consumer and producer surplus at this equilibrium. The introduction of BGH is illustrated in the second panel of Figure 11. This innovation allows dairy farmers to increase the quantity of milk they supply at any price, so the supply curve for milk shifts to the right. As a result, the point at which supply and demand intersect moves down along the demand curve from point A to point B. In the new equilibrium, the price is lower, and the quantity is higher.

It is clear that the total surplus has increased as well, since the shaded area between the supply and demand curves is now larger. Consumers are unambiguously better off as a result of the innovation. Since the market price is now lower, everyone who previously purchased milk receives a larger surplus. In addition, at the lower price consumers purchase additional quantities of milk. The effect on producers is more ambiguous. The increase in sales causes an increase in producer surplus, but the lower price reduces the producer surplus on the quantity that was previously being sold. Whether producers benefit depends on the balance of these two effects.

Let's consider another example of how shifts in supply

FIGURE 11





Effects of BGH on the Market for Milk

and demand affect market equilibrium. Public health officials have long recognized that cigarette smoking is harmful. As a result, policymakers would like to reduce smoking. One approach is to reduce the demand for cigarettes through public education campaigns and the inclusion of warning labels on packages of cigarettes. Assuming that these efforts do in fact cause buyers to demand fewer cigarettes, what is the effect on the market for cigarettes?

The answer can be found by examining Figure 12. To illustrate the effect of public efforts to reduce smoking, Figure 12 shows the demand curve for cigarettes shifting to the left. As a result, the intersection of the supply and demand curves shifts down and to the left along the market supply curve for cigarettes. After this shift, the equilibrium price and quantity both decrease.

Elasticity

The competitive market model we have developed allows us to predict the direction in which equilibrium price and quantity will change in response to changes in market supply or demand. But to fully understand the impact of these changes, it is important to be able to measure the size of the changes in prices and quantities as well as their direction. To do this, we need to introduce the concept of price **elasticity**.

The **price elasticity of demand** measures how much the quantity demanded responds to a change in price. We calculate the price elasticity of demand using the following formula:

Price elasticity of demand =
(Percentage change in quantity demanded) /
(Percentage change in price)

Recall that because of the law of demand, the quantity demanded of a good is negatively related to its price, so this ratio will always be negative. It is conventional to ignore this sign when discussing the elasticity of demand. In other words, in practice, we use the absolute value of the price elasticity of demand.

The price elasticity of demand reflects how responsive consumers are to changes in the price of a good. The greater the elasticity, the greater the proportionate change in the quantity consumers demand due to any given change in the price. Demand is said to be elastic if a one percent change in price results in a greater than one percent change in the quantity demanded. Demand is said to be inelastic if a one percent change in price results in a less than one percent change in the quantity demanded. And demand is said to be unit elastic if a one percent change in price results in a one percent change in the quantity demanded.

Economists use elasticity because it provides a measure of the responsiveness of demand to price changes that is independent of the units of measurement. For example, if we express the quantity of gasoline demanded in liters, then we will find that the demand curve has a different slope from the one that would result if we measured demand in gallons. However, the elasticity will be the same in both cases.

Measuring the actual elasticity of demand for particular products is an important activity of applied economics. Nonetheless, we can state some general guidelines about the factors that influence the price elasticity of demand.

- **Substitutes**. Goods with close substitutes will tend to have relatively high price elasticities of demand because it is easy for consumers to switch from one product to another. For example, the price elasticity of demand for a particular cola drink is likely quite high because consumers can easily switch to a different brand if the price rises. Conversely, when there are no close substitutes, the price elasticity of demand will tend to be lower.
- **Necessities**. Items that are regarded as necessities will generally have lower price elasticities of demand than luxuries. Many people must drive to and from work and use their cars to run important errands. As a result, the demand for gasoline has a low price elasticity of demand.
- **Market Definition**. The price elasticity of demand will depend on how we define the market. The broader the market definition, the fewer close substitutes there will be and the lower the elasticity of demand. The price elasticity of demand for soft drinks will be lower than the price elasticity of demand for any particular brand of cola drink.
- **Time Horizon**. Fully adjusting to changes in prices may take time. Take the example of gasoline prices considered earlier. At first there is not much people can do to reduce their consumption when the price of gasoline rises.

But, over time people will buy more fuel-efficient cars, move closer to their work, and make other changes that will allow them to more significantly reduce their demand.

Elasticity is related to the slope of the demand curve. If two demand curves pass through the same point, the curve that is flatter will have a higher elasticity. It is important to note that as we move down along a linear demand curve, the elasticity will be falling continuously. To see this, note that a linear demand curve must have a constant slope $\Delta P/\Delta Q = e$, (where we use the Greek letter Δ to denote the change in price and quantity along the demand curve). The ratio $\Delta Q/\Delta P = 1/e$, is also a constant.[4] Consequently the elasticity of demand is equal to $(1/e)\cdot(P/Q)$. As we move down and to the right along the demand curve, P is falling and Q is rising, so the ratio P/Q must be decreasing. Since $1/e$ is constant, the elasticity must also be falling.

Figure 13 shows five different possible demand curves illustrating the range of possible elasticities. In the extreme case (a) demand is *perfectly inelastic*; the quantity demanded does not depend on price at all. The remaining panels show progressively more elastic cases: (b) inelastic, (c) unit elastic, (d) elastic, and the other extreme case (e) *perfectly elastic*, where the demand curve is completely flat.

The price elasticity of supply is defined analogously to the price elasticity of demand. It is calculated as:

Price elasticity of supply =
(Percentage change in quantity supplied) /
(Percentage change in price)

The elasticity of supply reflects the ease with which suppliers can alter the quantity of production. We can establish some general guidelines that allow us to identify factors that are likely to affect this responsiveness.

- **Ease of entry and exit**. If it is easy for new businesses to begin supplying a product or for those in the market to leave, then supply will tend to be more elastic. The supply of airline flights on a particular route is quite elastic because airlines can easily shift planes from one route to another to respond to changes in prices.
- **Scarce resources**. If an input required to produce a good is scarce, then the supply will be inelastic. For example, the supply of beachfront vacation homes is highly inelastic because the amount of beachfront property is limited.
- **Time horizon**. The longer the time horizon is, the greater the elasticity of supply will be. Over short time horizons, firms may not be able to hire and train additional workers or add the necessary equipment to increase production. Over a longer horizon, they can do this more easily.

As was the case with the price elasticity of demand, if two supply curves pass through the same point, the flatter curve will be the more elastic one. Figure 14 illustrates the variety of possible supply curves. Again there are five cases. In the extreme case (a) the supply is *perfectly inelastic*, indicating that the quantity supplied will not change at all as the price changes. The supply of Van Gogh sunflower paintings is perfectly inelastic since there is no way to produce more of these. The remaining cases illustrate (b) inelastic supply, (c) unit elastic supply, (d) elastic supply, and the other extreme case (e) *perfectly elastic supply.*

Using Elasticity

To see how measurements of elasticity can be used, let's return to the example of the introduction of Bovine Growth Hormone that we considered earlier. As a starting point, we need to consider how the elasticity of demand affects **total revenues** available to producers in this market.

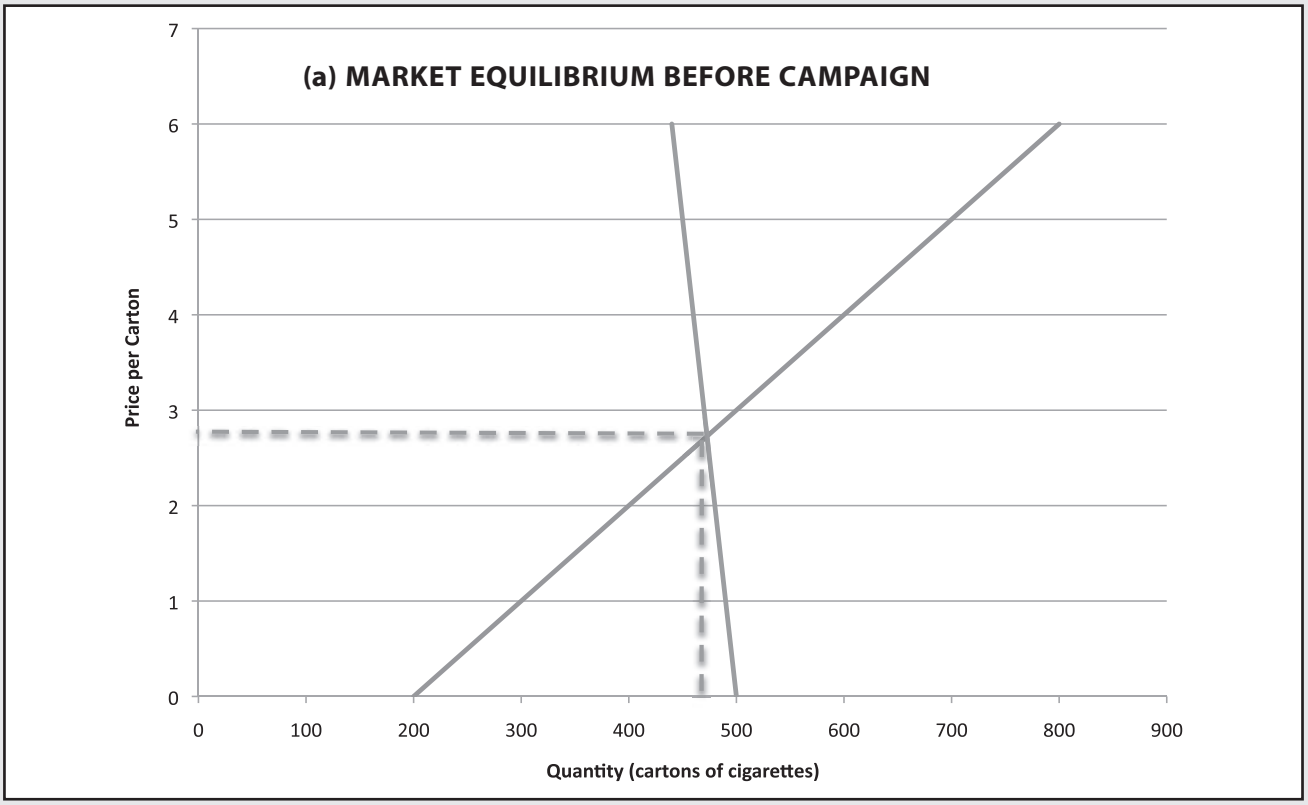
Total revenue is the equilibrium price multiplied by the equilibrium quantity:

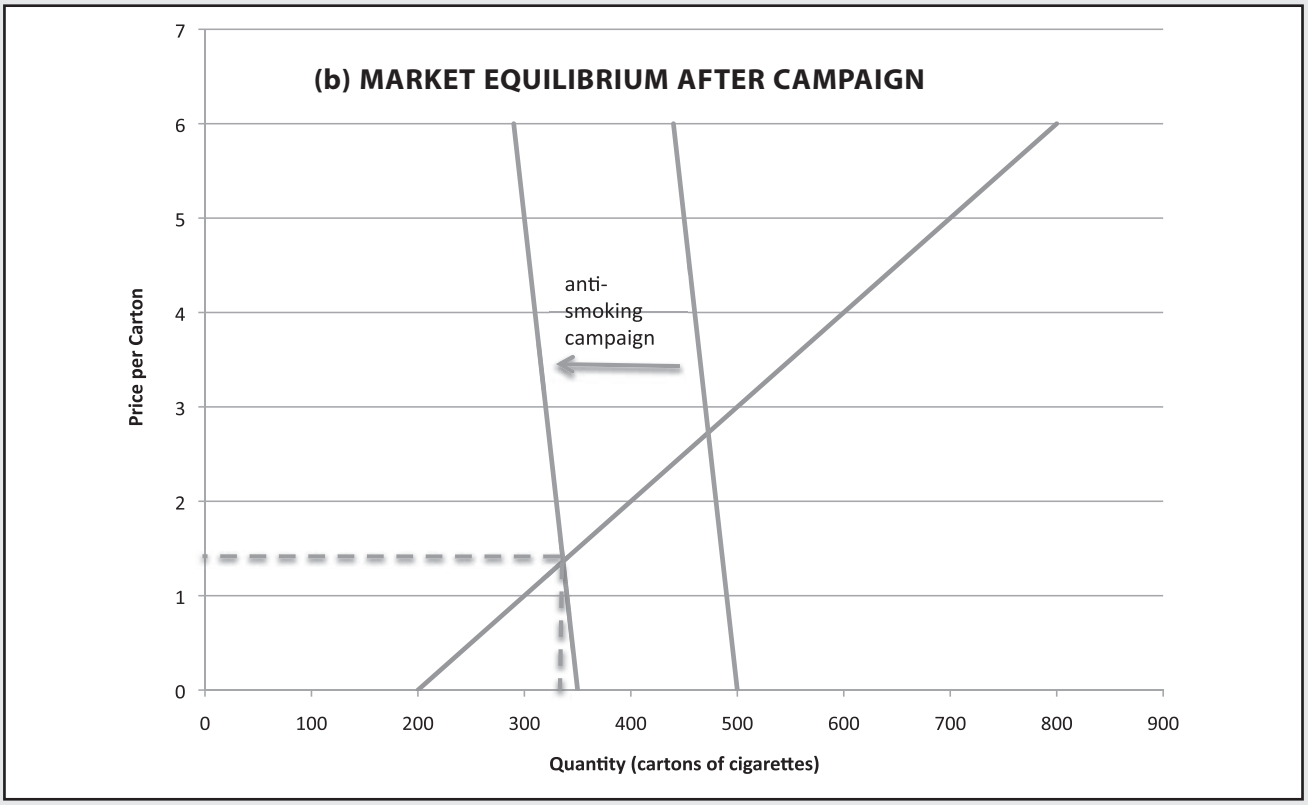
Total Revenue = P × Q

The total revenue can be depicted graphically as in Figure 15. As the price falls, we move down along the demand curve: the height of the box is reduced as its width increases. If the demand is elastic, total revenue will increase since the proportionate change in quantity will be greater than the proportionate increase in the price. But, if demand is inelastic, then total revenue will decrease when prices fall.

Empirical estimates suggest that the demand for milk is relatively inelastic. Milk is a necessity, and it does not have many close substitutes. As a result, declining prices do not induce a large increase in the quantity demanded. On the other hand, the supply of milk is relatively elastic over a time horizon of a year or more. There are a great many dairy farms, and it is easy for

FIGURE 12

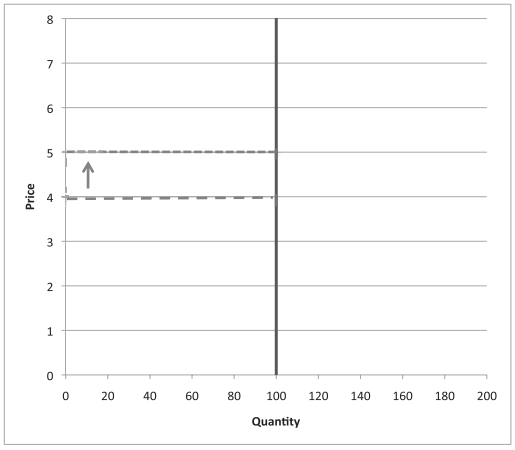




Effects of a Reduction in Market Demand

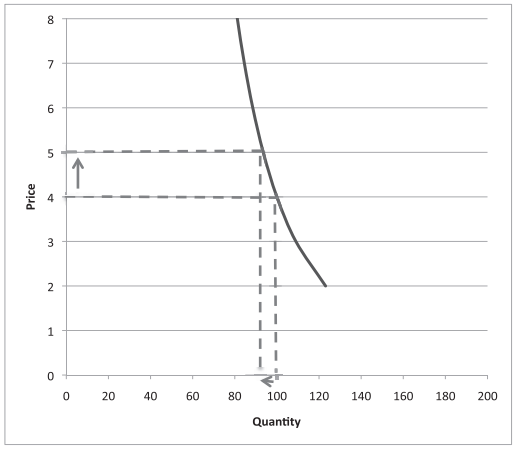
FIGURE 13

(a) PERFECTLY INELASTIC



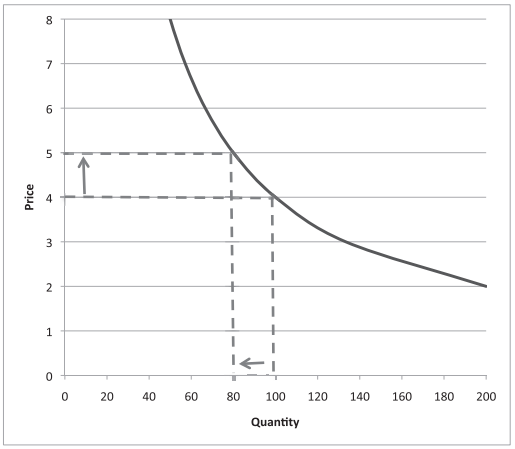
A 22% increase in price and no change in quantity.

(b) INELASTIC DEMAND: ELASTICITY IS LESS THAN 1



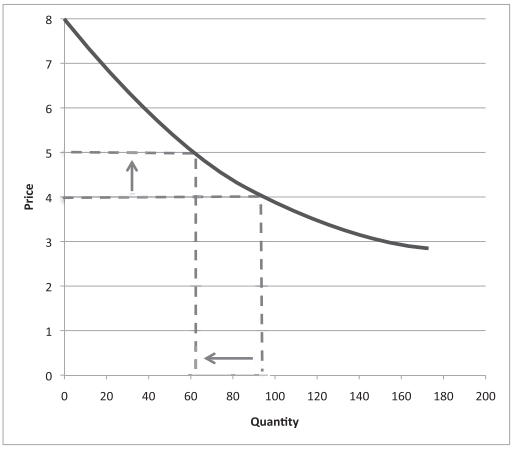
A 22% increase in price leads to a 7% reduction in quantity.

(c) UNIT ELASTIC DEMAND: ELASTICITY EQUALS 1



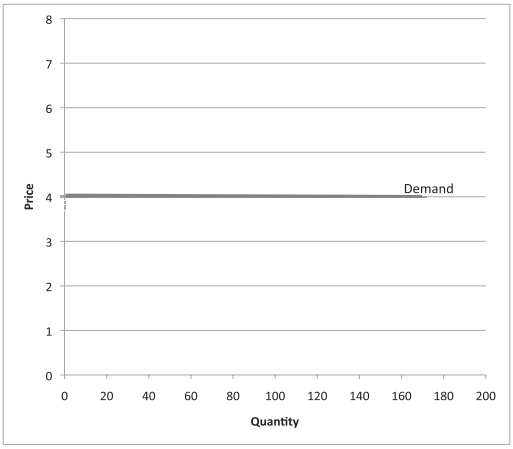
A 22% increase in price leads to a 22% reduction in quantity.

(d) ELASTIC DEMAND: ELASTICITY IS GREATER THAN 1



A 22% increase in price leads to a 44% decrease in quantity.

(e) PERFECTLY ELASTIC DEMAND: ELASTICITY EQUALS INFINITY



At $4, consumers will buy any quantity; if the price is above $4, they will buy none, and if the price is below $4, they will buy an infinite quantity.

The Price Elasticity of Demand

these farms to expand or contract their production.

In Figure 15, the demand curve is drawn as inelastic at the initial price and quantity pair. As the price of milk falls from an initial level of $2 a gallon to $1.50, the quantity demanded per day rises only from 2,000 gallons to 2,250. In this case, the price has fallen by 25 percent, and the quantity demanded has increased by just 12.5 percent, which implies an elasticity of –0.5 (= 12.5 / –25). As a result, total farm revenue falls from $4,000 to $3,375. In aggregate, dairy farmers are now earning significantly less revenue than before.

If using BGH reduces farm income, why do dairy farmers adopt this technology? The answer is that in a competitive market they have no choice. Each farmer supplies only a small amount of the total output, and their choice about whether to use BGH has no effect on the market price. Given the existing market price, each farmer can increase their sales by using BGH. As a result, competition causes them to all adopt the technology, increasing the market supply and driving down prices.[5]

As farm revenue falls, it is likely that some farmers will choose to cease producing, allowing the remaining farmers to maintain or increase their standard of living by producing a greater quantity. This is, in fact, more or less what has happened in the farm sector over the past two hundred years. Successive technological innovations have increased the ability of farmers to produce greater quantities of crops, though this advance has been accompanied by a steady decline in the number of farmers.

EVALUATING GOVERNMENT POLICY: THE IMPACT OF PRICE CONTROLS AND TAXES

So far our discussion has been confined to describing how competitive markets work. The tools we have developed to describe the operation of competitive markets can also be used to analyze several commonly used policy interventions. As we will see, the effects of these policies often diverge from the goals of those who designed them.

Price Controls

Efforts to legislate minimum or maximum prices are a fairly common kind of policy intervention in markets. For many years, U.S. farm policy established minimum prices of major food crops, such as corn and wheat. The federal government and most states have established a minimum wage, and some communities have gone further, seeking to legislate that employers pay a "living wage." Similarly, New York and some other cities have sought to control residential housing costs by establishing rent controls that limit increases in the rates that landlords can charge. In 1979, when Middle Eastern oil supplies were interrupted and heating oil prices shot up, the federal government imposed a ceiling on prices in an effort to protect low-income families.

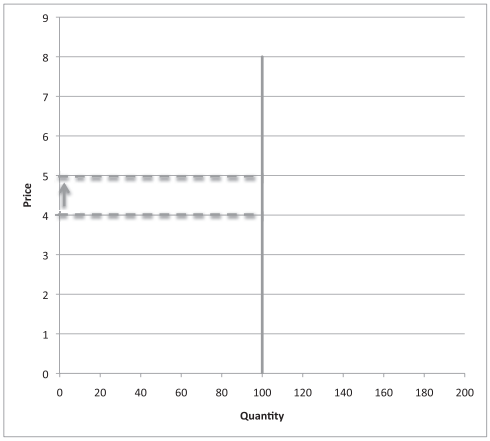
When the market price appears to unfairly hurt either consumers or suppliers, it is tempting to suggest that government intervention could set a better price. But, such efforts create significant, though not always obvious, social costs. To see these, consider a few examples. Figure 16 illustrates the impact of imposing a price ceiling on the housing market. In this example, the competitive market equilibrium occurs at a rent of $400 per month. At this price, consumers rent two thousand apartments each month.

Now suppose that landlords are told they may charge no more than $300. At this price consumers wish to rent 2,100 apartments, but landlords only supply 1,900 apartments. Those tenants lucky enough to be able to find an apartment benefit from the lower rental rates, while landlords find themselves with lower incomes. Meanwhile, other renters lose their apartments. One effect of rent control, then, is to increase the consumer surplus of some renters while reducing the producer surplus of landlords and thus negatively affecting other renters. A second consequence is that total surplus is reduced since rent control prevents some mutually beneficial transactions from taking place. There are landlords who would like to rent their apartments for more than $300 per month, and there are consumers who would be willing to pay a higher price.

Less immediately apparent is the disruptive effect of rent controls on the allocation of apartments. In the competitive market equilibrium, apartments are rationed by price. Everyone who values an apartment as much or more than the market price is able to rent one, and landlords who are willing to supply apartments at or below the market price are able to rent them. Now, however, landlords are in a position to select tenants. They may require people to pay a finders fee, they may choose to rent to their friends, or they may discriminate based on personal characteristics they value or dislike.

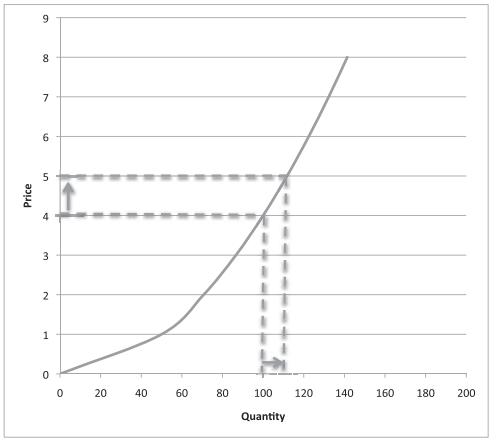
FIGURE 14

(a) PERFECTLY INELASTIC SUPPLY: ELASTICITY EQUALS 0



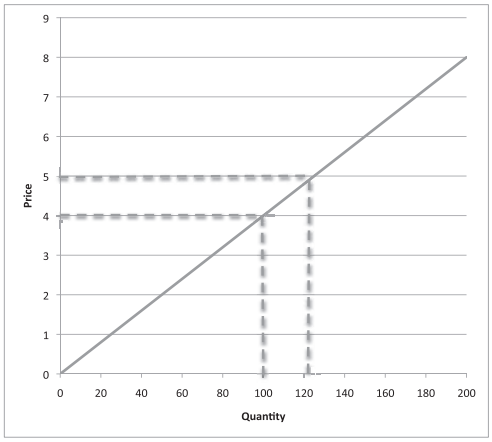
A 22% increase in price leads to no change in quantity.

(b) INELASTIC SUPPLY: ELASTICITY IS LESS THAN 1



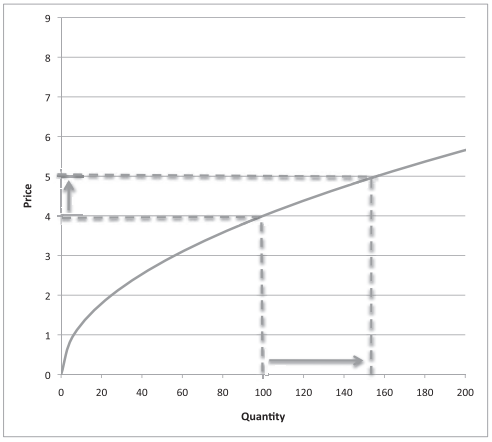
A 22% increase in price leads to an 11% increase in quantity.

(c) UNIT ELASTIC SUPPLY: ELASTICITY EQUALS 1



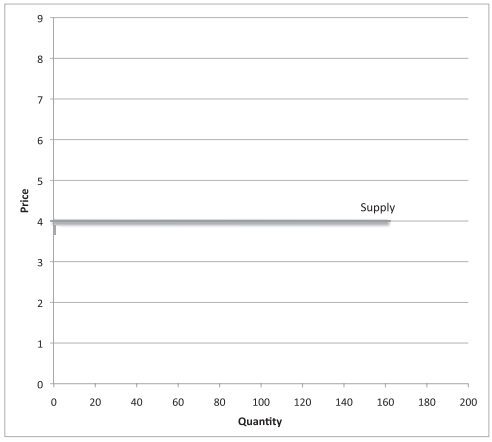
A 22% increase in price leads to a 22% increase in quantity.

(d) ELASTIC SUPPLY: ELASTICITY IS GREATER THAN 1



A 22% increase in price leads to a 44% increase in quantity.

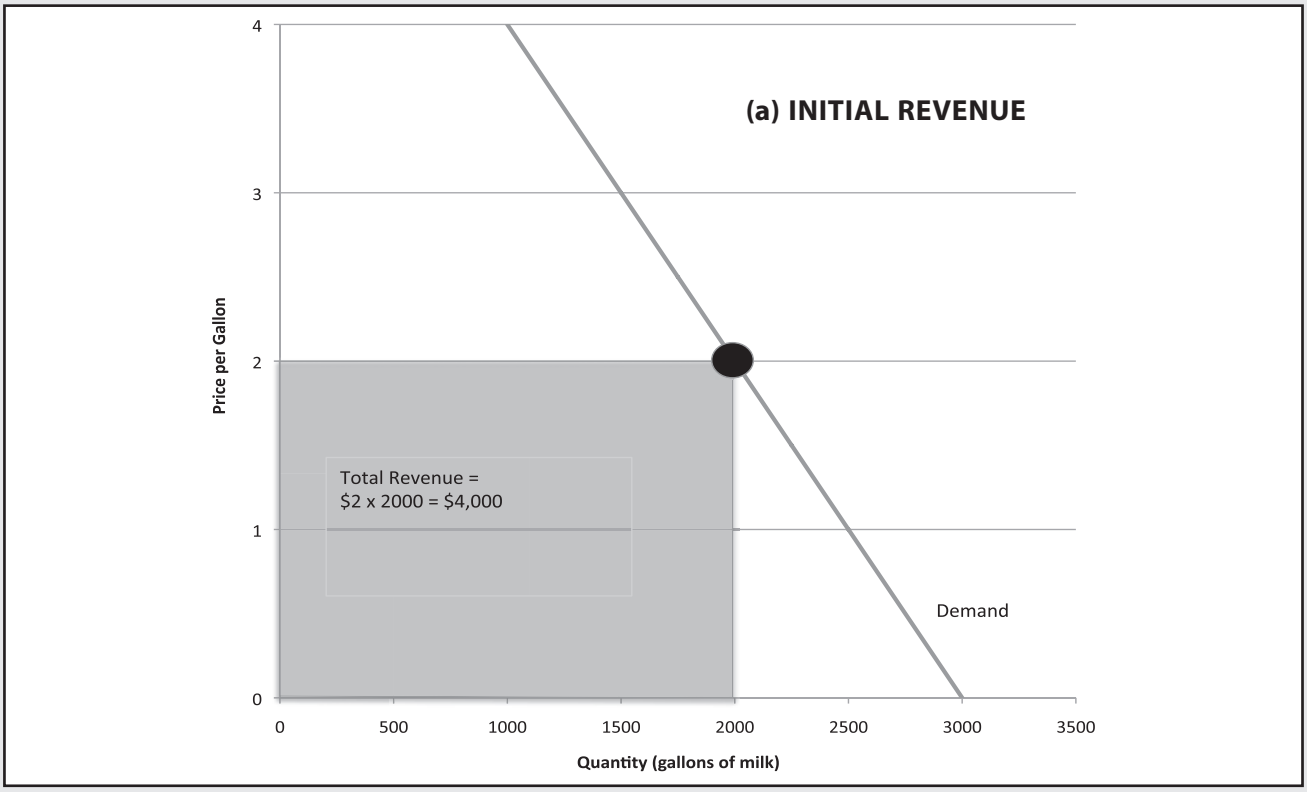
(e) PERFECTLY ELASTIC SUPPLY: ELASTICITY INFINITE

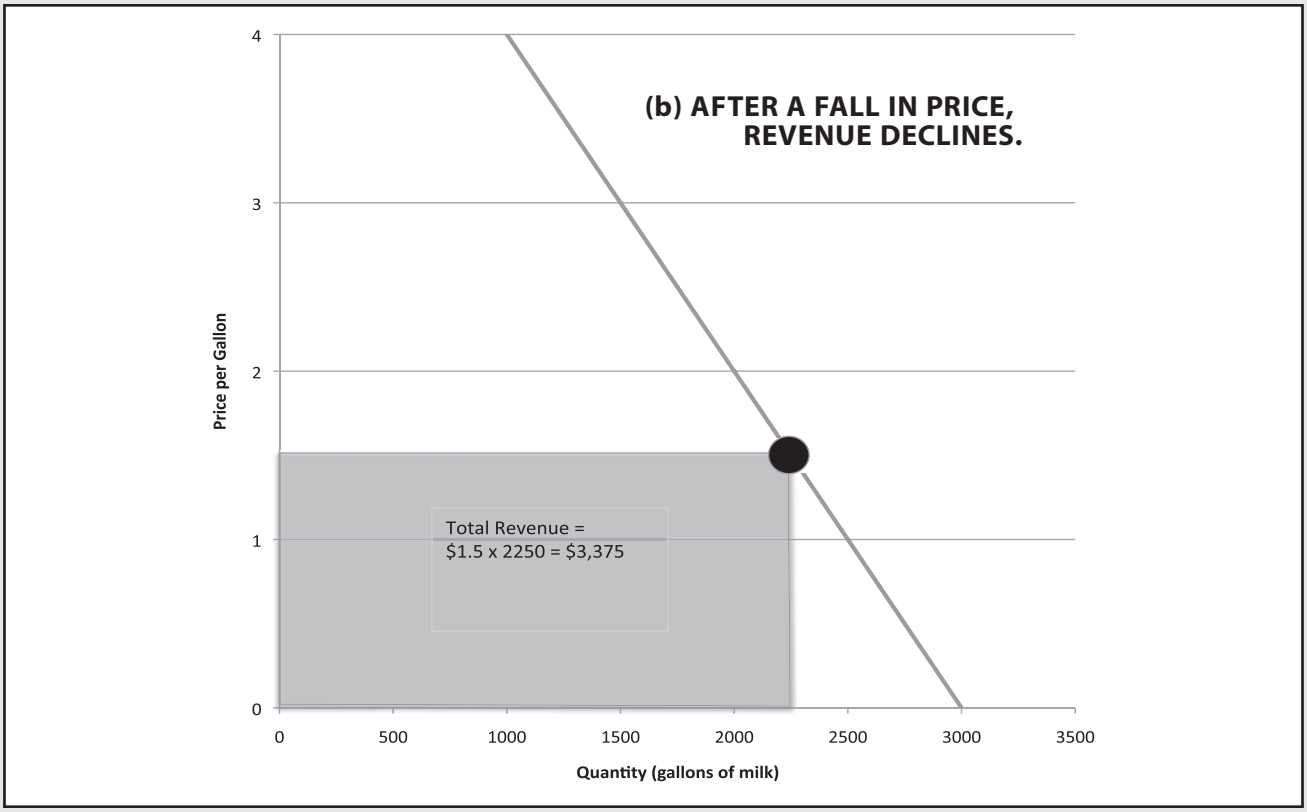


Producers will supply any quantity demanded at $4; if the price is above $4, they will supply an infinite amount; if it is below $4, they will supply zero.

Elasticity of Supply

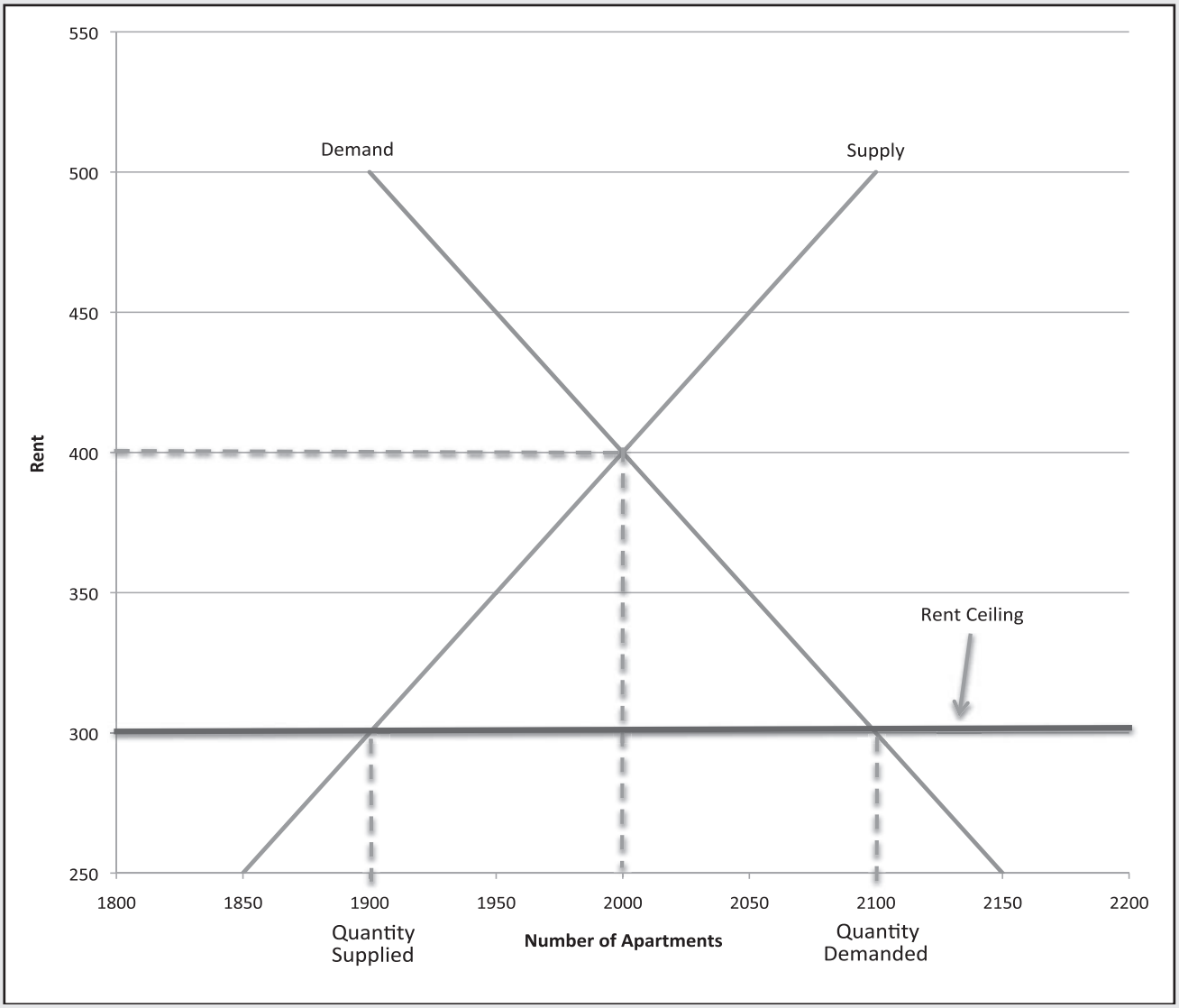
FIGURE 15





Calculation of Total Revenue

FIGURE 16



Effects of a Rent Ceiling

As a result, apartments may no longer go to the individuals who value them most highly, producing a further inefficiency in the market.

Historical experience points to further negative effects of rent controls. In the short run, both the supply of housing in a city and the demand for housing may be highly inelastic. As a result, rent controls mainly lower the price without creating a large excess demand. But, over time, both supply and demand become more elastic. Landlords will cut back on maintenance costs, allowing apartments to deteriorate, and eventually removing them from the available housing stock. Meanwhile, low prices will attract more residents to the city. With these changes, the problem of excess demand and non-market rationing will become increasingly significant.

To illustrate the effect of establishing a price floor, let's consider Figure 17, which shows the market for wheat. The competitive market equilibrium price in this market is $5 a bushel, and the equilibrium quantity is 100 million bushels. Suppose that in an effort to protect family farms, Congress establishes a minimum price of $8 a bushel. Because this price is higher than the current

market equilibrium, it is binding. At this higher price, demand for wheat falls to 80 million bushels, and supply rises to 115 million bushels. Farmers cannot sell all the wheat they are producing on the free market.

Once again this intervention reduces consumer and producer surplus. There are farmers who would be happy to supply wheat at lower prices and consumers who would be willing to buy from them, but they are prohibited from doing so. Those farmers who find buyers at the higher minimum price benefit from the legislation, but others find they are unable to earn any income and will likely be forced to go out of business.

Taxes

All levels of government use taxes of one sort or another to raise revenue that is used to pay for public expenditures. An important issue that often comes up in public discussion of taxes is who bears the burden of paying the tax. It would seem that government could control the distribution of burdens through legislation, but the results suggest that matters are not that simple.

To make matters more concrete, let's consider a tax on mobile phone usage. Figure 18a illustrates the demand and supply in this market before any tax is imposed. In the competitive market equilibrium, cell phone calls cost $0.20 per minute, and consumers make 1 million minutes of calls each day. Suppose the government decides that mobile phones are a luxury and chooses to impose a tax of $0.10 per minute on consumers. From the perspective of mobile phone users, the cost of a minute of talk is now higher than before—it costs them $0.30 (= $0.20 + $0.10).

We can represent the effect of this change in Figure 18b as a downward shift in the market demand curve. Notice that if the market price were $0.10, then consumers would face a cost of $0.20 per minute and would demand the same quantity as they had in the competitive market equilibrium (1 million minutes). So, the demand curve shifts down by $0.10, the amount of the tax. The new market equilibrium occurs at a lower quantity than before, and as a result, the price received by suppliers falls. In this hypothetical example, the new equilibrium price is $0.16 per minute; so suppliers receive $0.16 per minute, and consumers pay $0.26 per minute. Even though the tax is added to the consumers' bill, the actual burden of the tax is divided between suppliers and buyers. At the same time, the tax reduces the equilibrium quantity, lowering total surplus by preventing otherwise mutually beneficial exchange from taking place.

Suppose that instead of taxing consumers, the government imposed the tax on suppliers, charging them $0.10 for every minute of talk they supplied. As a result, the revenue that suppliers receive is reduced by $0.10. The effect of the tax on their behavior can be illustrated by shifting the supply curve upward by $0.10. Because of the tax, suppliers will require a market price of $0.30 per minute to supply the quantity they previously supplied at a price of $0.20 per minute.

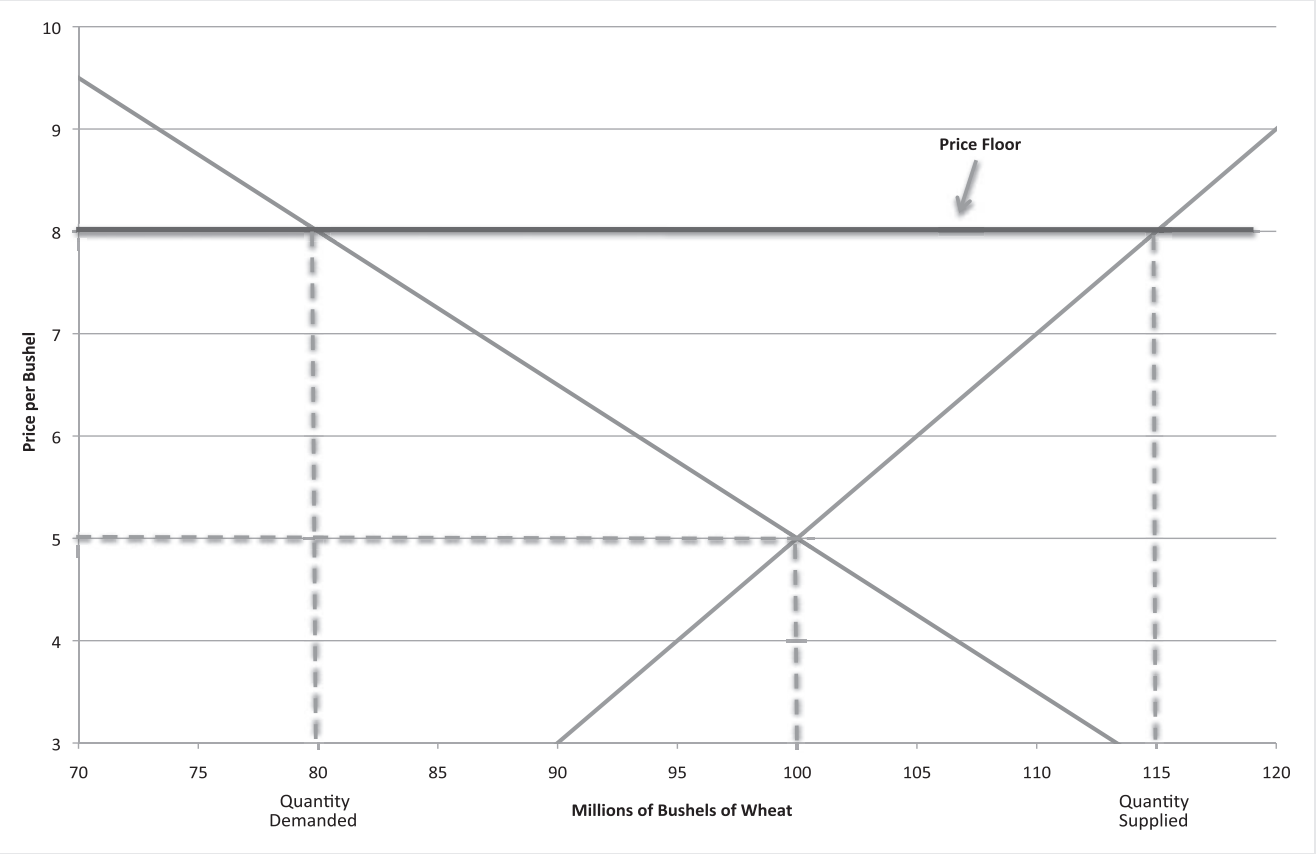
This situation is illustrated in Figure 18c. The new market equilibrium occurs at a market price of $0.26 per minute. At this price, suppliers receive $0.16 per minute. Notice that this is precisely the same outcome as we found when the tax was imposed on consumers.

This example illustrates an important point that is true more generally. A tax creates a price wedge between the amount consumers pay and the amount suppliers receive. This price wedge reduces the market quantity, and regardless of who legally pays the tax, both consumers and suppliers share the cost of the tax. Recognizing this fact, we can depict the impact of the tax in a third way, as is illustrated in Figure 18d. Rather than shifting the supply or demand curve, we search for the point where the vertical distance between the demand and supply curves is equal to the amount of the tax. The heights of the supply and demand curves, respectively, at this point identify the prices that suppliers receive and consumers pay. Extending a vertical line downward from this point to the horizontal axis identifies the equilibrium quantity once the tax has been imposed.

With regard to consumer and producer surplus, the tax has several effects. By introducing a difference between the price buyers pay and that received by suppliers, the tax prevents some otherwise mutually beneficial transactions from taking place. This is indicated by the small triangle to the right of the new equilibrium quantity and between the supply and demand curves. This is a reduction in social welfare and is called the **deadweight loss** of the tax.

The other effect of the tax is to transfer revenue to the government. The government collects $0.10 on every minute purchased at the new equilibrium. The amount of this revenue is illustrated in Figure

FIGURE 17



Effect of a Price Floor for Wheat

18d by the shaded rectangle. Initially people talked 1 million minutes, but notice that a tax of ten cents per minute generates less than $100,000 ($0.10 × $1 million) in revenue to the government. This is because the tax has caused people to demand fewer minutes of calling than before. As this diagram makes clear, the revenue that the government receives reduces the combined consumer and producer surplus from these transactions by an amount equal to the income that the tax produces for the government.

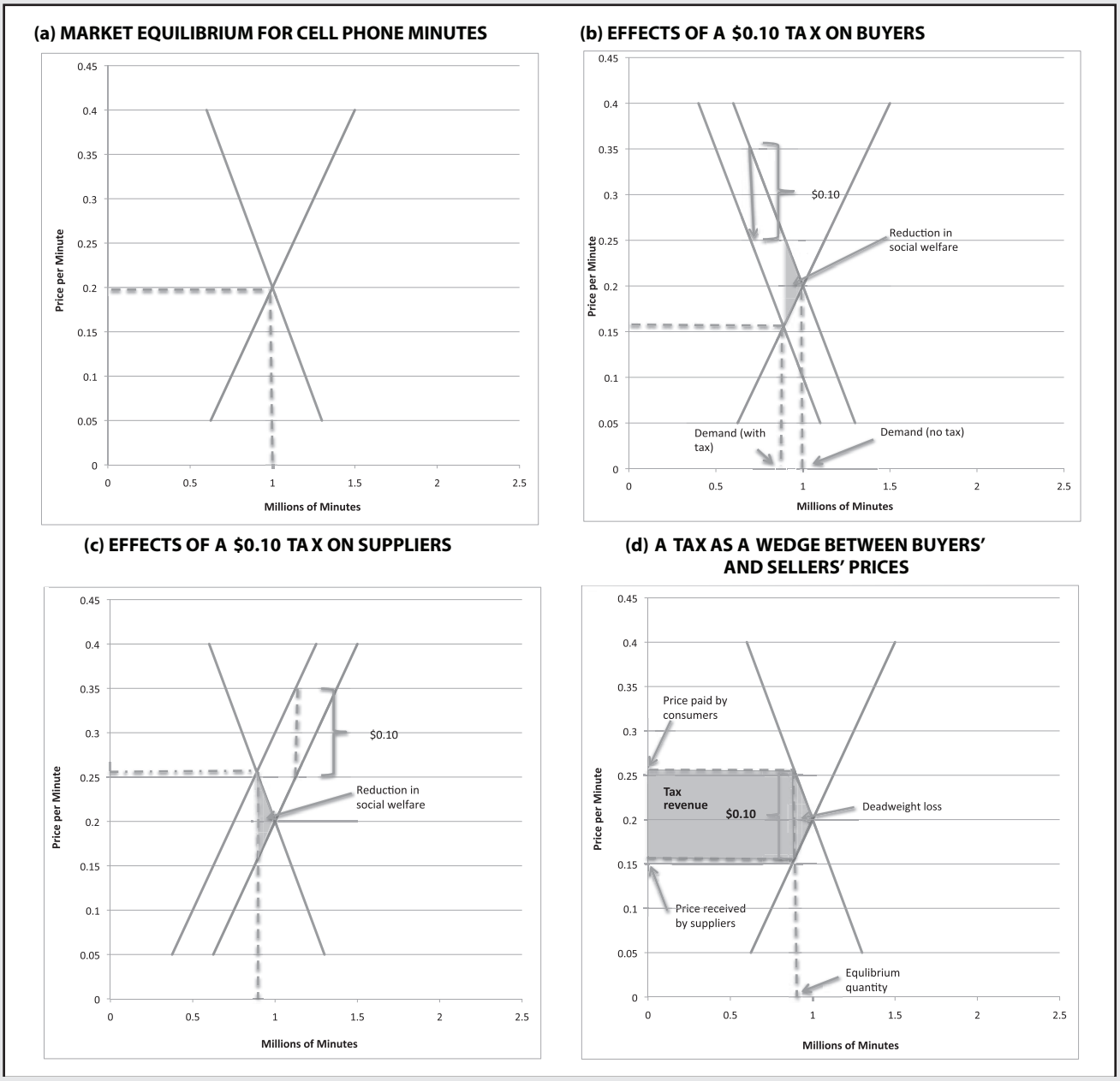
In our hypothetical example, suppliers paid 40 percent of the cost of the tax through reduced revenues, while buyers paid 60 percent of the cost of the tax through an increase in their cost per minute. In general, the distribution of the burden of a tax depends on the relative price elasticities of supply and demand. For any given supply curve, the less elastic the demand is, the greater the share of the tax paid by buyers. This is illustrated for two demand curves in Figure 19a. Figure 19b depicts a similar comparison, showing how the elasticity of supply affects the division of the tax. One final point that emerges from an examination of Figure 19 is that the less elastic the supply and demand curves are, the smaller the effect of the tax on the equilibrium quantity, and therefore the lower the deadweight loss of the tax.

INTERNATIONAL TRADE

One of the fundamental insights of economics is that exchange makes people better off. It does so by encouraging specialization. When individuals or countries specialize in the activity they do the best, the overall economic pie increases. These **gains from trade** are the reason that our modern economy is characterized by such a high degree of interdependence.

To appreciate the gains achieved from trade, we need to begin by considering an isolated economy. Then, we can consider how the opportunity to trade alters well-being.

FIGURE 18



Representing the Effects of a Tax

An Isolated Economy

As a starting point, let's consider a highly simplified economy. Robinson is stranded on a tropical island. Each day he works for eight hours to produce food, which he consumes. He can devote his time either to harvesting coconuts or catching fish. Each hour that Robinson spends gathering coconuts is an hour that he does not spend catching fish. The opportunity cost of the additional coconuts that he gathers is the quantity of fish that he does not catch during that hour.

We can represent the trade-off that Robinson faces in terms of a **production possibility frontier** or PPF like that drawn in Figure 20a. In this diagram, we measure the quantity of coconuts Robinson gathers on the vertical axis and the number of fish he catches on the horizontal axis. The graph shows that if Robinson

spends all eight hours gathering coconuts, he can collect twenty-four coconuts—this is the height of the curve where it intersects the vertical axis. If he spends all of his time catching fish, then he can catch eight fish—this is the distance from the origin to the point where the PPF intersects the horizontal axis.

Robinson can select any point along the PPF, which we have drawn here as a straight line. The slope of this line reflects the opportunity cost of coconuts in terms of fish. Since the PPF has a slope of –3, it indicates that Robinson must give up three coconuts to get one additional fish. All of the points on the PPF are efficient from the perspective of production since along this line there is no way that Robinson can increase the quantity of one good produced without reducing the quantity of the other.

The point that Robinson chooses along the PPF depends on his relative preferences for fish and coconut. He will select the combination of fish and coconuts that maximizes his satisfaction. Suppose that he selects a point like A, where he is consuming fifteen coconuts and three fish. From our discussion of the demand curve, we know that at this point if Robinson is a rational consumer, he will get just as much pleasure from one more fish as from three coconuts. If this were not so, then he could improve his well-being by moving along the PPF. For example, if one fish gave him as much pleasure as two coconuts, he could reduce his consumption of fish by one and increase his consumption of coconuts by three. Since it only takes two coconuts to compensate for the fish he has given up, he would be better off.

Adding the Opportunity to Trade

Crusoe lives on a nearby island, where she too gathers coconuts and catches fish. In Figure 20b we show her PPF. Looking at her production, we can see that Crusoe is better at catching fish than Robinson, and she is better at gathering coconuts. In an eight-hour day, she can catch thirty-six fish or gather thirty-six coconuts. Because Crusoe's PPF is above and to the right of Robinson's at every point, we say that she has an *absolute advantage*.

The slope of her PPF is –1, indicating that the opportunity cost of one fish is one coconut. Crusoe can select any point along her PPF. But by the same logic we used before, we know that at that point she will value one fish the same as one coconut. Let's suppose that Crusoe is initially consuming eighteen fish and eighteen coconuts at point B.

One day, Robinson finds a boat and sails to Crusoe's island. They begin to talk about their respective consumption patterns, and Robinson proposes that if they agree to trade, they can both be better off. Crusoe is skeptical at first since she produces more fish and more coconuts than Robinson, and so she cannot see how they could find an opportunity to trade. But Robinson persists. He points out to her that at the moment they are producing a total of thirty-three coconuts (Robinson's 15 plus Crusoe's 18) and twenty-one fish (3 + 18). But, if Robinson were to devote eight hours to gathering coconuts, he could produce twenty-four. Meanwhile, if Crusoe were to spend two more hours fishing, then she could produce nine coconuts and twenty-seven fish. Together their combined production would be thirty-three coconuts (the same as before) and twenty-seven fish (six more than before). If they split this extra production, they could each increase their consumption by three fish.

Comparative Advantage and the Gains from Trade

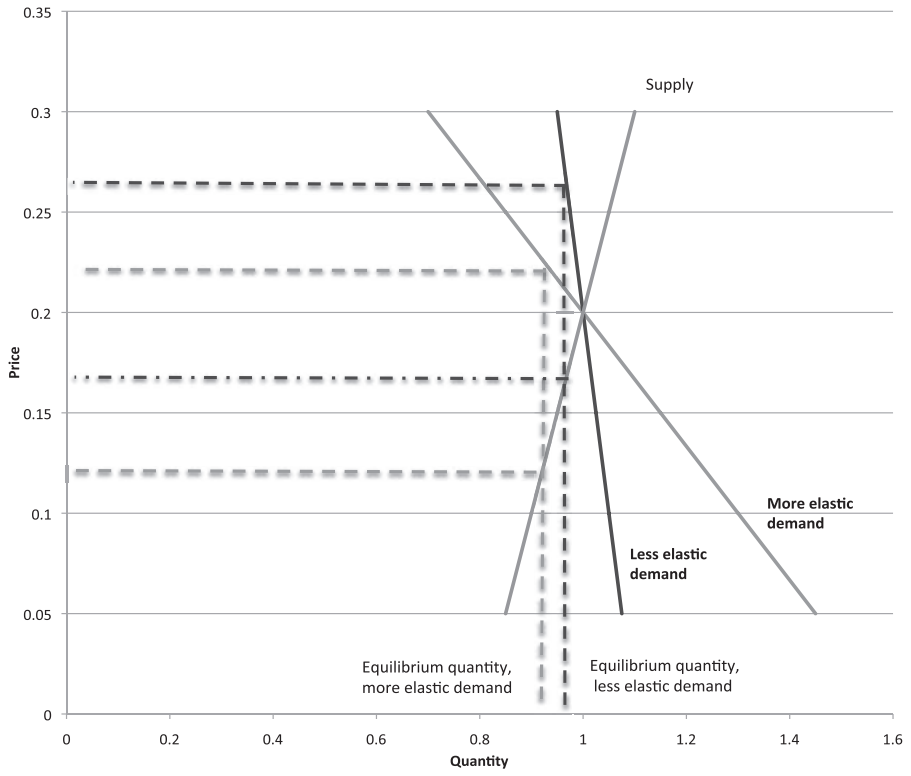
How can it be that Crusoe, who is better at everything, can be made better off by trading with Robinson? The answer to this question lies in the insight that what matters is not the absolute productivity of either Robinson or Crusoe, but rather their respective **comparative advantage**. Even though Robinson produces fewer coconuts per hour than Crusoe, he has a comparative advantage in producing coconuts.

By changing their allocation of time between fishing and gathering coconuts, Robinson and Crusoe in effect "transform" fish into coconuts. Robinson faces a cost of just 1/3 fish per coconut, while it takes Crusoe one fish to produce a coconut. When Robinson specializes in producing coconuts and Crusoe specializes in producing fish, their collective economy can increase its total production.

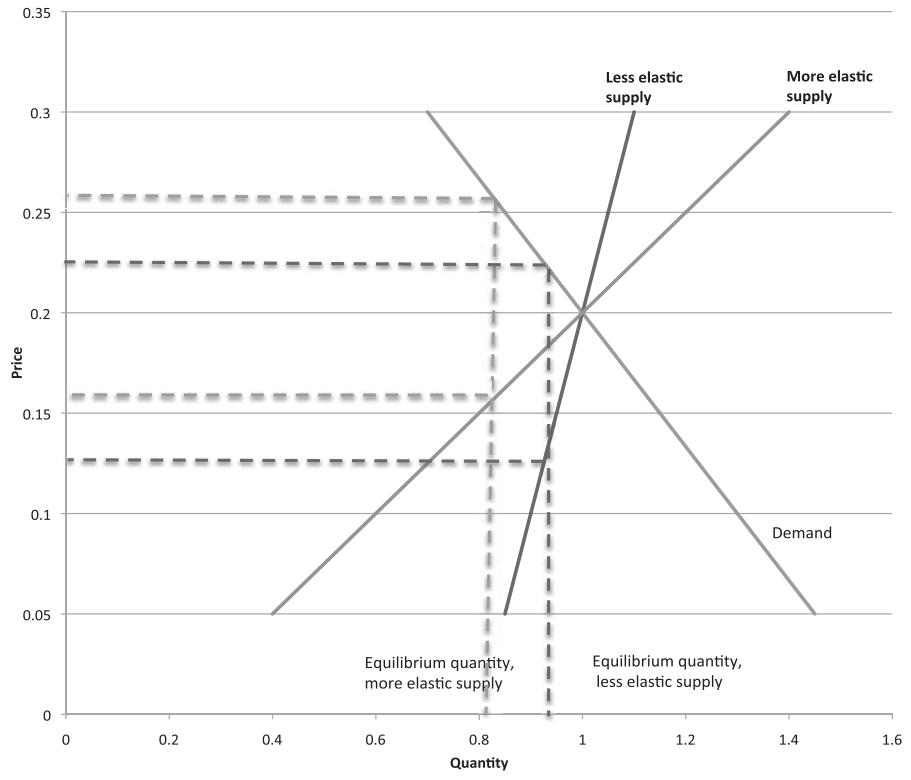
The principle of comparative advantage offers a profound insight about the opportunities for gains from trade that applies equally to individuals and to nations. So long as trading partners differ in their comparative advantage, they can improve their overall well-being by specializing. The more extensive the markets in which they trade are, the greater the opportunities for

FIGURE 19

(a) EFFECTS OF ELASTICITY OF DEMAND ON IMPACT OF A TAX

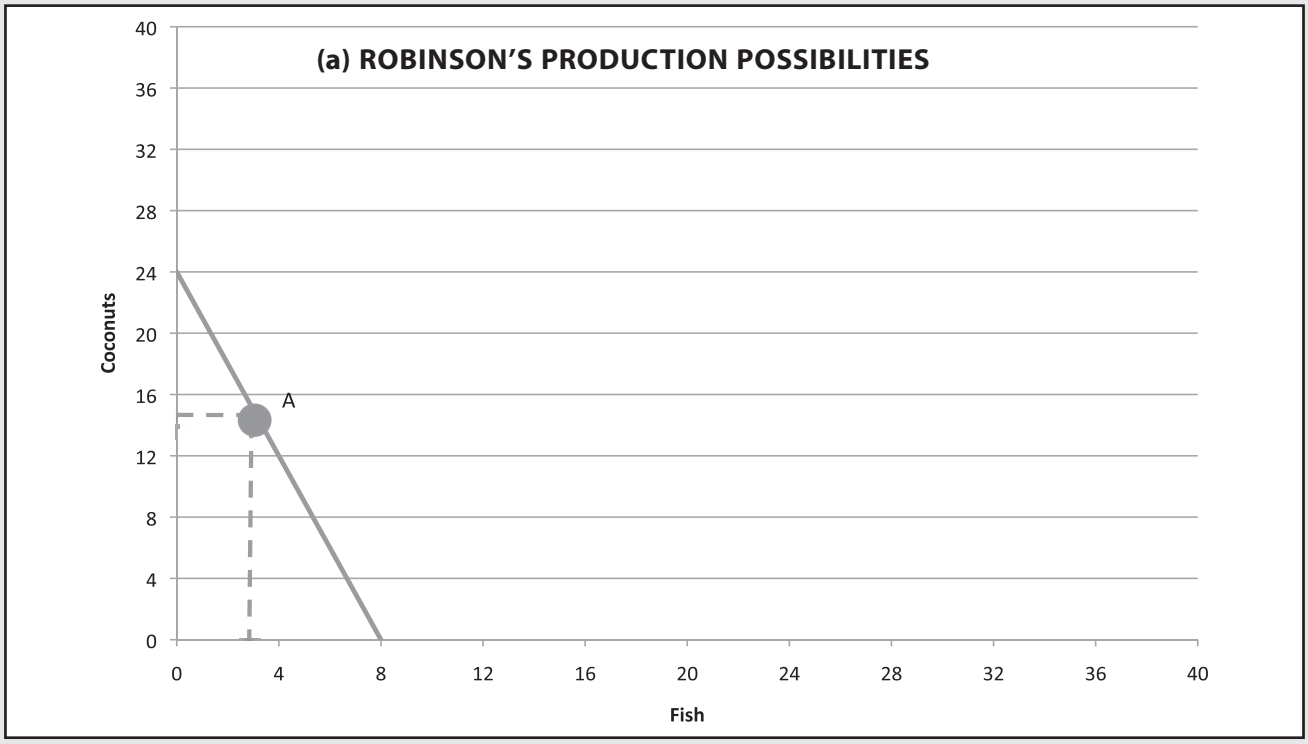


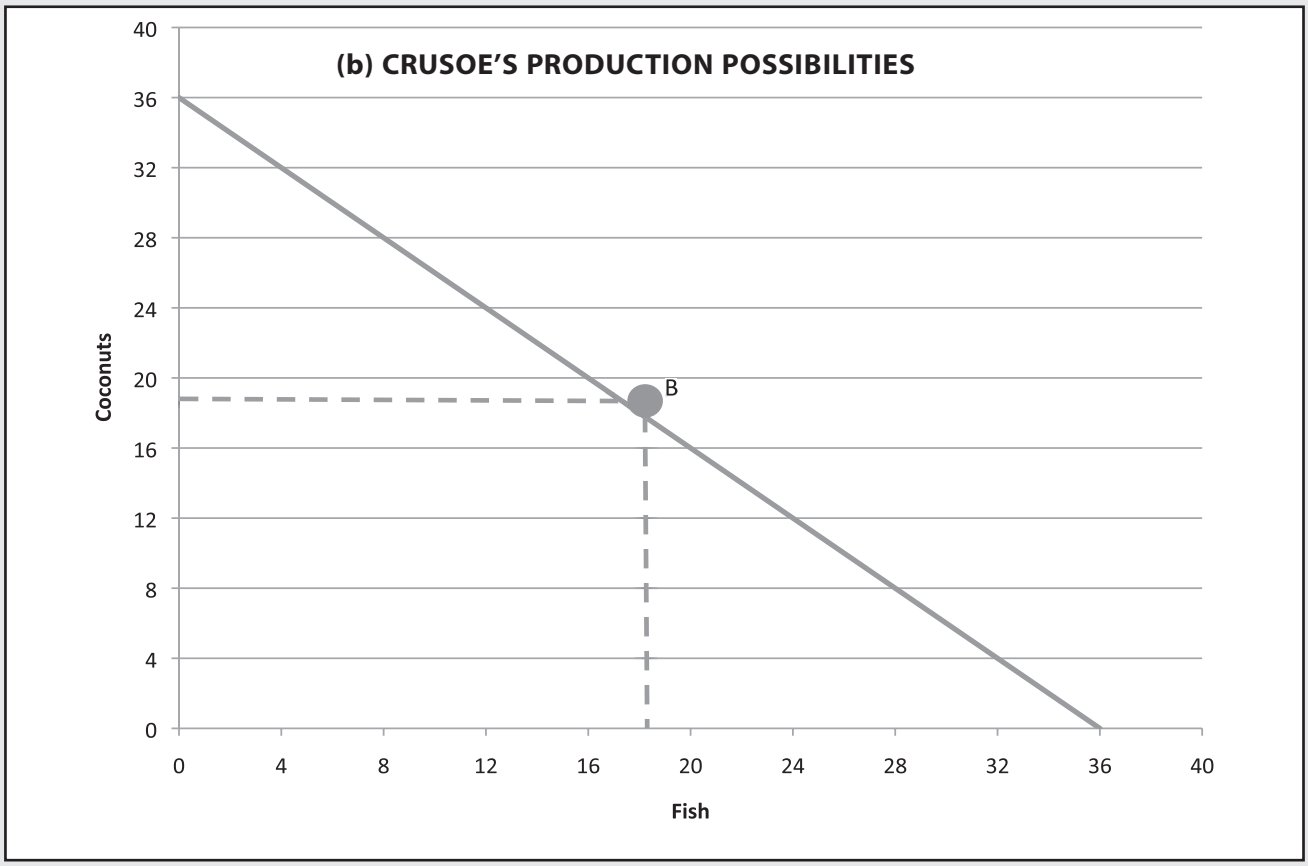
(b) EFFECTS OF ELASTICITY OF SUPPLY ON IMPACT OF A TAX



Effects of Elasticity on the Impact of a Tax

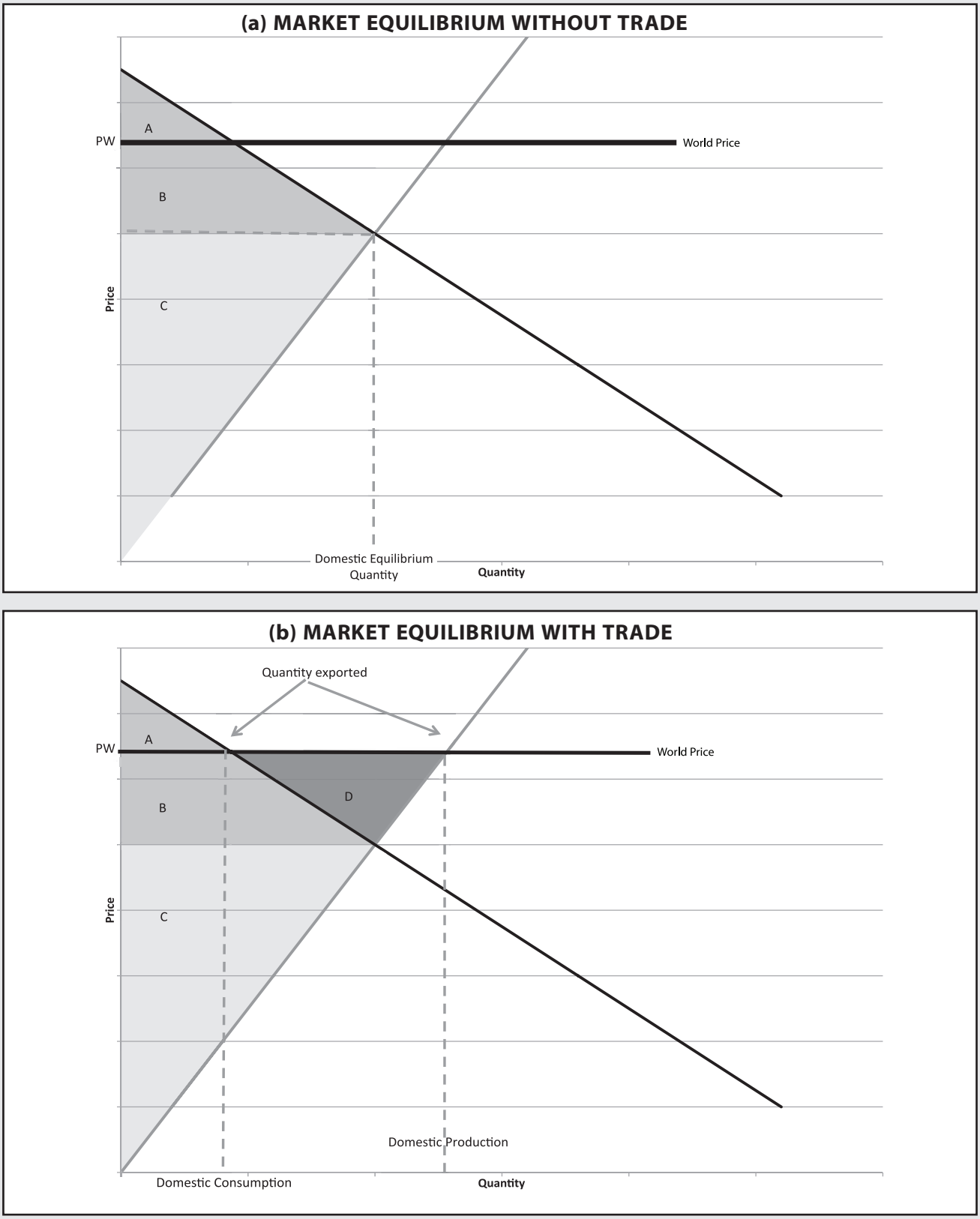
FIGURE 20





Production Possibility Frontiers

FIGURE 21



Welfare Effects of Isolation and Free Trade

specialization and the larger the gains from trade.

The Political Economy of Trade

If trade increases a nation's well-being, then why is there so much public opposition to international agreements designed to promote freer trade? While free trade expands the overall size of the economy, it also implies shifts in the size of different industries. In the previous example, Robinson and Crusoe simply reallocated their time. But when countries become increasingly specialized, the costs and benefits of trade fall on different groups of people. As a result, even though the gains from free trade exceed the losses, those citizens who will experience losses are likely to oppose freer trade.

To see this, let's consider the impact of free trade in more detail. We will begin by considering a small economy that is isolated from international markets because trade is prohibited. As a result, the domestic equilibrium is determined by the intersection of the country's supply and demand curves as depicted in Figure 21a. Suppose that the world price is PW, illustrated by the horizontal line above the domestic equilibrium price. Consumer surplus is equal to the sum of the areas marked A and B; producer surplus is equal to the area C.

If the law prohibiting trade is removed, this country will become an exporter, since its cost of supply is below the world price. To simplify the analysis, we assume that the country is so small relative to the world market that its additional supply will not alter the world price. The equilibrium quantity will occur where the world price intersects the country's market supply curve.

At PW, domestic consumers reduce their consumption. The difference between domestic consumption and the quantity supplied is exported. Consumer surplus falls because the price rises, and consumers purchase less of the good. The value of their surplus is represented by the area labeled A. Producer surplus increases, however. It is equal to the sum of the areas marked B, C, and D. So, producers benefit and consumers suffer when a country becomes an exporter. In total, however, social welfare increases from the area A+B+C to the area A+B+C+D, yielding a net increase equal to the area denoted by D.

For a country that becomes an importer, social welfare again increases, but now it is consumers who benefit, while producers suffer losses. This situation is illustrated in Figure 22, where the domestic equilibrium price is above the world price. When trade is allowed, the domestic price falls to the world price, and the quantity consumed rises. Domestic producers, however, respond to the lower price by reducing their supply. The difference between the quantity produced domestically and the quantity consumed domestically is made up by imports. Producer surplus declines from the areas B+C to B, while consumer surplus increases from A to A+C+D.

THE PROFIT MOTIVE AND THE BEHAVIOR OF FIRMS

Economists use the term "firm" to describe the economic actors who are responsible for supplying goods and services in the economy. Firms combine labor, capital equipment, raw materials, and other inputs to produce the products that we want to consume. Up to now, we have used the supply curve to summarize their actions. According to the law of supply, as the price of a good rises, firms are willing to supply a greater quantity. In many cases, this is all we need to know about the behavior of firms. But, in other instances, we need to look more closely at how firms decide what to produce and how to produce it.

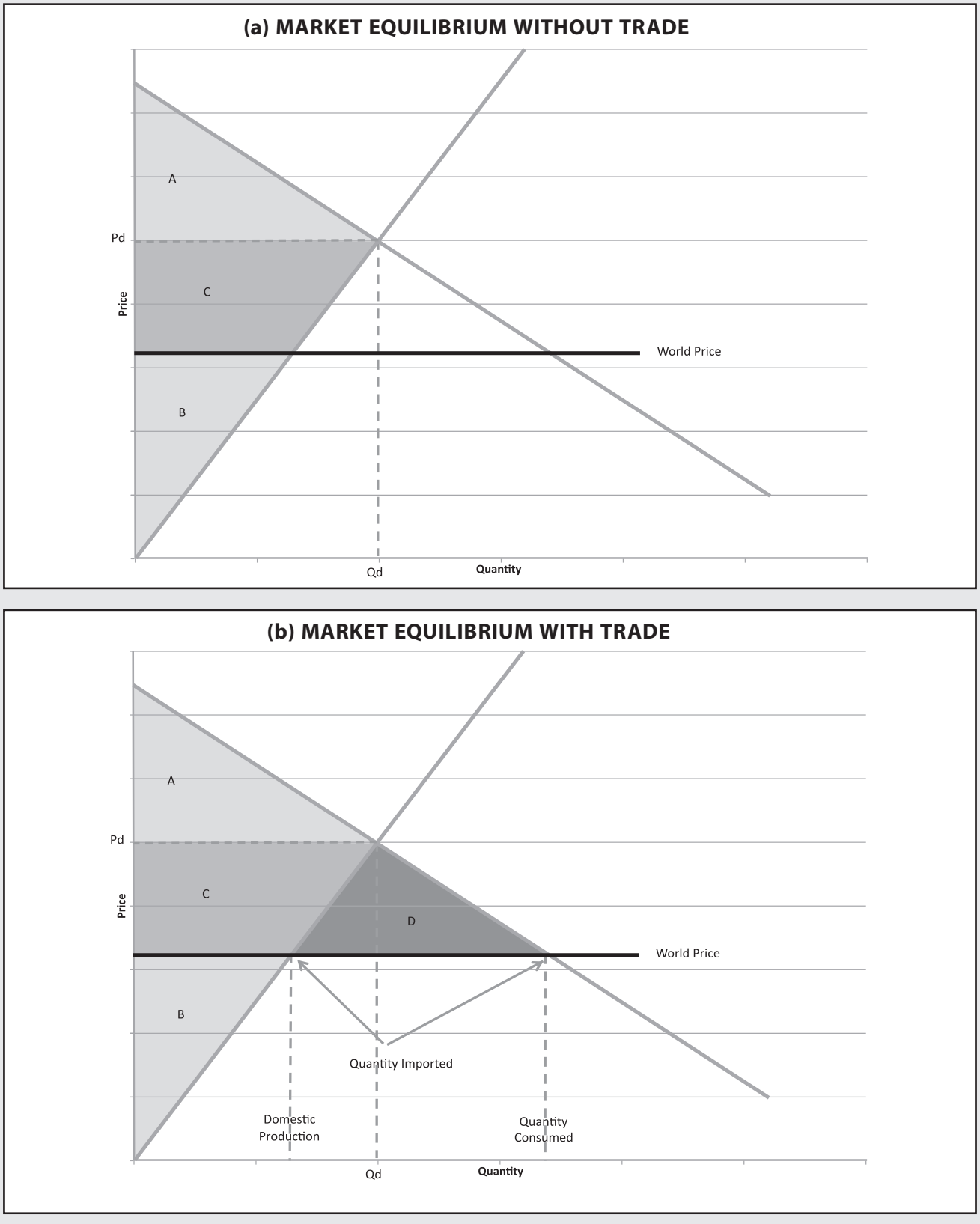
Economic Profits and Accounting Profits

We assume a firm's goal is to maximize profits. Profits are defined as the difference between the firm's *total revenue* and its *total costs*. The meaning of total revenue is fairly clear: it is the total quantity of output the firm produces for sale multiplied by the price it receives. Measuring total costs is a bit more complicated. Economic costs include the opportunity costs of all resources required for production. In contrast, accounting costs will likely include only actual monetary expenditures.

This distinction can be seen more clearly by considering an example. Consider Bob's Bread Company. Bob's is a small bakery that sells a variety of freshly baked breads. All of the baking is done in the back of the store, and Bob operates a retail shop at the front. Suppose that Bob sells 300 loaves of bread a day for $4 each. Total revenues are $1,200 a day.

Bob's explicit costs include purchasing flour and other

FIGURE 22



Welfare Effects of Isolation and Free Trade

ingredients, for which he pays $600 a day; hiring labor to produce the bread costs $300 per day; and renting the shop in which he operates costs Bob $50 per day. These explicit costs total $950, leaving an accounting profit of $250 a day.

But, we have not yet included all of the firm's opportunity costs. Bob is a skilled retailer and, if he were not managing his bread company, he could earn $200 a day managing another store in town. Because Bob gives up this income to manage his own business, we must include this forgone income as part of his economic costs. As a result, the true **economic profit** that Bob's Bread Company earns is $50 per day.

Finding the Firm's Supply Curve

In the example above, Bob is producing 300 loaves of bread each day. How does he choose this level of production? Recall that Bob's objective is to maximize his profits. Some of his costs, such as the opportunity cost of his time and the rent on the building and equipment do not depend on the quantity of bread he produces and cannot be changed in the short run. These are what we call his **fixed costs**. However, the cost of Bob's labor and materials can be varied in the short run. These are called his **variable costs**.

The table in Figure 23 summarizes information about Bob's costs of production. In the second column, we list his fixed costs. Because these do not depend on the quantity of bread Bob chooses to supply, they do not change. In the third column, we show Bob's variable costs of production. Notice that each time we move down a row, output increases by 50 loaves a day, but the additional cost of producing those additional loaves of bread increases from row to row.

The increase in costs that occurs when producing an additional unit of output is referred to as **marginal cost**. The marginal cost is calculated by dividing the increase in total costs by the increase in the quantity of bread produced. This additional cost is referred to as the marginal cost of production. For example, when Bob increases his production from 50 to 100 loaves, his total costs increase from $358 to $483, and thus his marginal cost of producing these additional loaves is ($483 –$358) / (100 – 50) = $2.50. As you go down the rows in the top section of the table, the change in total cost is increasing, implying that marginal costs are increasing as well.

The bottom panel of the table provides the information necessary to determine Bob's profit-maximizing production level. Because Bob can vary his production by amounts smaller than 50 loaves, we have calculated his marginal costs of production for the quantity shown in each row for small changes in the quantity produced at that point. As a result, these values will differ somewhat from the marginal costs you would calculate using the data in the top part of the table.[6]

Increasing marginal costs of the type illustrated in Figure 23 are common in economics. Such a relationship usually arises because some of the factors of production are fixed and cannot be increased in the short run. In this case, Bob cannot increase the number of ovens available. As a result, once the ovens begin to fill up, the addition of more workers produces less and less additional output. This is an example of **diminishing returns to scale**.

Bob's marginal cost of production is the opportunity cost of supplying an additional loaf of bread since it measures the amount that Bob must spend to produce that loaf. The benefit that Bob gets from supplying another loaf of bread is the additional revenue that it will produce. This additional revenue is called the **marginal revenue**.

By assumption, the market for bread is perfectly competitive, meaning that the price Bob receives is not affected by the quantity he chooses to supply. From his perspective, the demand curve is horizontal at the market equilibrium price of $4 a loaf. This means that Bob's marginal revenue is equal to $4 regardless of the quantity he chooses to supply.

Combining the information about Bob's costs with the information about his marginal revenue, we can now find his profit maximizing output. The necessary information is summarized in the bottom panel of Figure 23. So long as Bob's marginal cost of supplying an additional loaf of bread is less than $4, he can increase his profits by producing and selling that loaf. Reading down the marginal cost column, we see that Bob's marginal cost equals $4 when he is producing 300 loaves of bread.

So long as diminishing returns to scale apply, marginal costs will be rising as the firm's output increases. As a result, the profit-maximizing firm's supply curve will be an upward-sloping line.

Entry, Exit, and the Market Supply Curve

Bob is, of course, only one possible supplier of bread. Other potential producers are likely to notice that Bob is making an economic profit of $50 a day. Recall that we have already accounted for the opportunity cost of Bob's time. The opportunity to earn extra profits will induce some of these producers to rent shops and equipment and begin producing bread as well.

The addition of more producers has the effect of shifting the market supply curve outward. And, this in turn will cause the equilibrium price to fall. The entry of additional producers will continue as long as there are positive economic profits to be earned in the market. Only when economic profits have reached zero will entry cease. In the same way, if economic profits were to fall below zero at some point—say because of a shift in preferences that reduced the demand for bread—producers would begin to leave the market, shifting to other activities that offered greater opportunities.

Two points are worth emphasizing about this conclusion. First, in a competitive market business owners earn zero economic profits. They will, however, be content, because they are earning their opportunity wage. In other words, this remains their best alternative.

Second, in addition to their role in rationing scarce goods, prices serve a second important function: they allocate productive resources between different activities. If prices exceed production costs in some activity, then the existence of positive economic profits acts as a signal that additional resources should be deployed to that activity to increase production.

IMPERFECT COMPETITION

Now that we understand how firms behave in perfectly competitive markets, we can begin to develop an understanding of how markets that are not perfectly competitive work. Although perfect competition is a reasonable approximation for many parts of the economy, the markets for many important products are dominated by a small number of very large firms. Examples include the markets for commercial airplanes, automobiles, and ride sharing services. In other cases, such as electricity, water, and cable television, there is only a single supplier in any community. Economists call markets with one or only a few suppliers *imperfectly competitive*.

Firms in imperfectly competitive markets have the same objective as firms in perfectly competitive markets: to maximize their economic profits. But unlike firms in a perfectly competitive market, a firm in an imperfectly competitive market can no longer assume that its decision about how much to supply does not affect the price at which its products can be sold. Rather, the demand curve it faces is downward sloping, meaning that if it chooses to increase its supply, the price it receives will be lower.

Firms facing a downward sloping demand curve are said to possess **market power**, meaning that instead of taking prices as given, they have the ability to choose market prices. Of course, they are not entirely free to choose any price since they are constrained by the combinations of price and quantity determined by the market demand.

Monopoly

There are a wide range of different types of imperfectly competitive markets. The simplest case to consider is the extreme situation of a single supplier, a situation called a **monopoly**. Monopolies arise because there are barriers to entry that prevent competitors from entering the market. The most important sources of **barriers to entry** are:

- **The ownership of a key resource**. The market for residential electricity supply is a monopoly in most communities because a single company owns the retail electricity distribution system. It would not be possible for a competitor to establish another distribution system. Another example is the market for diamonds. Until recently the DeBeers company owned mines from which 80 percent of the world's diamonds are produced. Because diamonds can be mined in only a few places, ownership of these places allows for the establishment of what is effectively a monopoly.
- **Government-created monopolies**. Many monopolies are created when the government gives the rights to supply a product to a single company. Patent and copyright laws are one mechanism through which such exclusive rights are granted. If the government grants a patent to an inventor who has developed a new technology, they are awarded the exclusive right to utilize the technology for twenty years

FIGURE 23

QUANTITY	FIXED COST	+	VARIABLE COST	=	TOTAL COST	CHANGE IN TOTAL COST
50	$250		$108		$358	
100	$250		$233		$483	$125
150	$250		$375		$625	$142
200	$250		$533		$783	$158
250	$250		$708		$958	$175
300	$250		$900		$1,150	$192
350	$250		$1,108		$1,358	$208
400	$250		$1,333		$1,583	$225
450	$250		$1,575		$1,825	$242

QUANTITY	MARGINAL REVENUE	MARGINAL COST	TOTAL REVENUE	–	TOTAL COST	=	PROFITS
50	$4.00	$2.33	$200		$358		–$158
100	$4.00	$2.67	$400		$483		–$83
150	$4.00	$3.00	$600		$625		–$25
200	$4.00	$3.33	$800		$783		$17
250	$4.00	$3.67	$1,000		$958		$42
300	$4.00	$4.00	$1,200		$1,150		$50
350	$4.00	$4.33	$1,400		$1,358		$42
400	$4.00	$4.67	$1,600		$1,583		$17
450	$4.00	$5.00	$1,800		$1,825		–$25

Bob's Bread Company Costs and Revenues

in exchange for revealing the details of his/her innovation. Under copyright law, an author becomes a monopolist over the book they have written.

- **Natural monopolies**. An industry is a natural monopoly when a single firm can supply the market at a lower cost than could two or more firms. This happens when there are large fixed costs that cause the firm's average costs to be falling at a scale of production that can serve the entire market. Railroads, pipelines, and cable television are all examples of markets that are prone to natural monopoly.

Monopoly Supply

To illustrate the supply decision of a monopolist, let's consider the example of the market for cable television services in Smallville, which is served by a single provider Local Media. The table in Figure 24 shows the demand for cable television service is negatively related to the price of a monthly subscription. At a price of $20, no one will purchase the service, but when the price falls to $19 a month, 100 households will subscribe. As the price falls further, demand increases. Local Media can choose to supply at any combination of price and quantity along the demand curve. Its total revenue at that point is equal to the price times the quantity.

What happens to the company's revenues as it selects different points along the demand curve? For example, consider moving from point A in the graph in Figure 24, where price equals $16 and the quantity is 400, to point B where the price is $15. The additional subscribers generate more revenue, but to achieve this, the company must lower its price to existing subscribers.

At point A total revenue is $6,400, and at point B it rises to $7,500. Lowering the price and increasing supply increases total revenue, but the marginal revenue—the incremental increase in revenues produced by each additional subscriber—is less than the price of service. Here the additional 100 subscribers generate just $1,100 in additional revenue, an increase of $11 per subscriber, even though the price of a monthly subscription is $15. The difference is attributable to the fact that Local Media must lower the price it charges its existing subscribers to attract additional customers.

What price should Local Media choose and how much should it supply at this price? The profit-maximizing strategy that we identified for a firm in a competitive market—increase supply until marginal cost equals marginal revenue—still applies for a monopolist. As long as marginal revenue is greater than marginal cost, increasing supply causes economic profits to increase, but increasing supply beyond this point causes profits to begin to decline.

Figure 25 illustrates the application of this strategy. Local Media's marginal cost curve is drawn as upward sloping, reflecting the fact that adding additional subscribers requires the extension of the network, which requires increasingly costly equipment. Local Media's marginal cost curve intersects the marginal revenue curve at a quantity of 700 subscribers. At this quantity, the marginal cost and marginal revenue are both $7, and the height of the demand curve indicates that demand equals 700 when the price is $13 per month. Local Media's profit-maximizing choice is to set the price at $13 and provide 700 subscriptions.

Welfare Consequences of Monopoly

If cable TV service in Smallville had been provided by a competitive market with marginal costs equivalent to Local Media's, then the market equilibrium would occur at a lower price and higher quantity, as can be seen from the location of the intersection of the market demand curve with the marginal cost curve. Compared to this hypothetical competitive outcome, the monopoly supplies a lower quantity at a higher price. It may also earn an economic profit. But, because of barriers to entry, there is no competition to drive these profits toward zero.

From the point of view of social welfare, the fact that Local Media is a monopoly has two effects. First, there is a transfer of consumer surplus to Local Media because those subscribers willing to purchase service at the monopoly price would have been able to purchase this service at a lower price in the competitive case. Second, there is a reduction in social well-being because Local Media restricts supply to be less than the competitive quantity. The additional output would cost less to produce than its value to consumers. But, Local Media will not supply it because to do so would reduce the revenue it gets from subscribers who place a higher value on the service.

Dealing with Monopolies

Because of the negative effects that monopolies create, government policymakers have adopted a variety of responses intended to reduce the impact of monopoly. Beginning with the passage of the Sherman Anti-Trust Act in 1890, the federal government has sought to use legislation to increase market competition. As a result, large mergers and acquisitions must be reviewed by government regulators to insure that they do not reduce competition in key markets. Anti-trust regulators can also break up companies, as happened when AT&T was split up in 1984, or take other steps to restrict anti-competitive practices, such as the requirement that Microsoft unbundle its Internet browser from the Windows operating system.

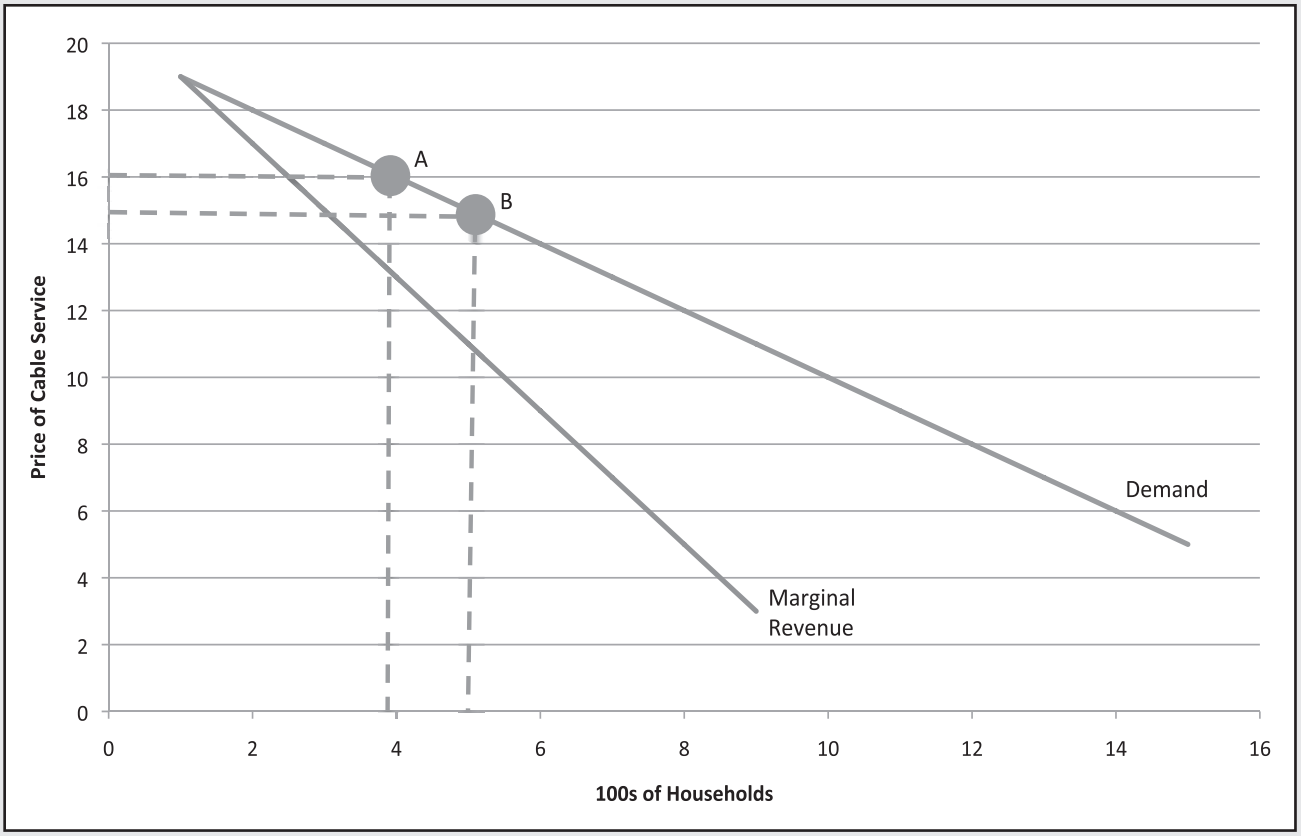
Another widely used approach is regulation. Many natural monopolies are allowed to exist but are closely regulated. Public utilities such as electric power companies and cable television providers cannot freely set prices, but must have rates approved by public oversight agencies. A third approach to the problem of monopoly is public ownership. Local water, sewer, and sanitation services are often operated by municipal governments, for example.

Price Discrimination

In the monopoly example we considered before, we assumed Local Media charged the same price to all of its customers. But, what would happen if it could charge different prices to different customers? If Local Media could charge each customer a price equivalent to the value that customer placed on its service, then it could avoid the negative effect of expanding sales on the revenue earned from existing customers. By charging different prices, Local Media's marginal revenue curve would be identical to the market demand curve, and it would choose to supply a quantity equivalent to the competitive market outcome.

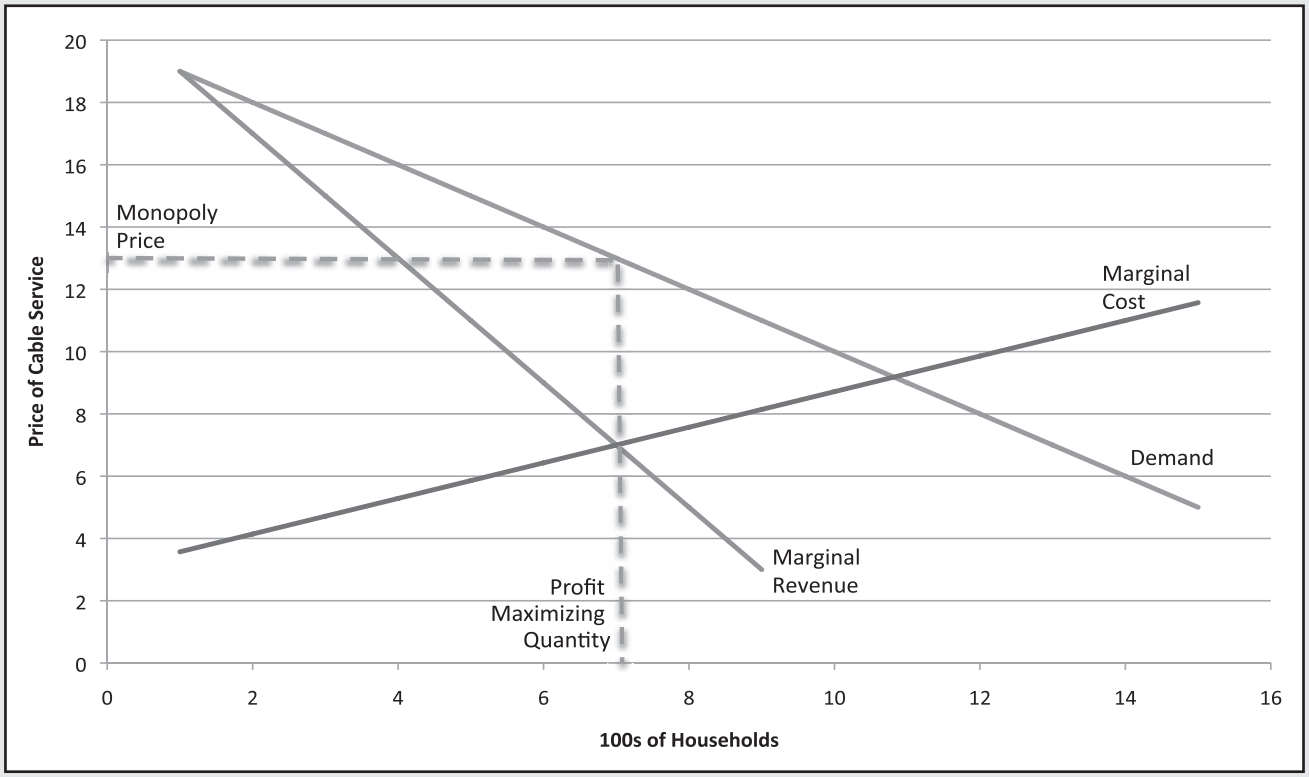
FIGURE 24

PRICE OF SERVICE	QUANTITY DEMANDED (100s OF HOUSEHOLDS)	TOTAL REVENUE = P × Q	MARGINAL REVENUE
20	0	$0	
19	1	$1,900	19
18	2	$3,600	17
17	3	$5,100	15
16	4	$6,400	13
15	5	$7,500	11
14	6	$8,400	9
13	7	$9,100	7
12	8	$9,600	5
11	9	$9,900	3
10	10	$10,000	1
9	11	$9,900	–1
8	12	$9,600	–3
7	13	$9,100	–5
6	14	$8,400	–7
5	15	$7,500	–9



Demand and Marginal Revenue of a Monopoly

FIGURE 25



Monopolist's Quantity Supplied and Price

Such a strategy is called perfect **price discrimination**. While companies can rarely discriminate perfectly between customers, it is easy to identify examples of ways that firms seek to separate customers into groups who value their product differently. One way that cable companies can price discriminate, for example, is by offering different packages of channels. Those who value the service most highly are likely to buy a large package at a higher price.

There are many other examples of price discrimination. Many movie theaters offer lower priced tickets for children and senior citizens, consumers who are likely to have a lower willingness to pay. Airlines typically charge lower prices for travelers who stay over a Saturday night. While leisure travelers will accept this condition in exchange for lower fares, business travelers, whose willingness to pay is higher, will not. College need-based financial aid is another price discrimination strategy.

Price discrimination further increases monopoly profits by allowing the monopoly to capture a greater fraction of the benefits produced by each transaction. But, price discrimination also has the positive effect of increasing social welfare by moving the market closer to the socially efficient quantity.

Oligopoly

Relatively few industries are true monopolies. In many more cases, a small number of producers supplies the bulk of the market. In the United States, the manufacture of tennis balls, breakfast cereals, aircraft, electric light bulbs, washing machines, and cigarettes are all industries in which production is highly concentrated.[7] Economists call a market with only a few sellers an **oligopoly**.

In comparison to monopoly markets, oligopoly markets are much harder to analyze. The reason for this is that in such markets, producers must consider not only the characteristics of the downward-sloping demand curve that they face, but also the choices that other suppliers will make. In other words, there is an opportunity for strategic interaction between the different suppliers.

If the suppliers could agree, for example, to cooperate and behave like a monopolist, total industry profits could be maximized. Such an agreement is called a **cartel**, and it is illegal under U.S. anti-trust law. There are also significant economic forces at work to undermine efforts by the members of potential cartels to collude. If a cartel is successful in restricting output, then marginal revenue will be greater than the marginal cost of production for each firm in the industry, creating a temptation to increase production. Of course, such an increase in supply lowers the market price, but much of this negative effect is felt by the other members of the cartel.

The Organization of Petroleum Exporting Countries (OPEC) provides a good illustration of the problem that cartels face. Because it is an international agreement between sovereign nations, OPEC does not face legal obstacles to its efforts to coordinate production and raise prices. In the 1970s, OPEC played an important role in raising oil prices from $11 a barrel in 1972 to $35 a barrel in 1981. Tempted by the high price of oil, many of its members began to increase production, and by 1986 oil prices had collapsed back to $13 a barrel. In April 2020, in response to share reductions in demand for oil due to the COVID-19 crisis, OPEC initiated a record reduction in oil output. In part due to this significant reduction in the supply of oil, prices increased sharply over the subsequent months.

As these considerations suggest, oligopoly outcomes depend critically on the circumstances of each market. We can nonetheless conclude that the outcome will lie somewhere between the polar cases of monopoly and perfect competition. As a rule then, oligopoly results in some reduction in social welfare, but we cannot easily say how large this reduction will be.

Monopolistic Competition

Perhaps the most common form of **imperfect competition** is **monopolistic competition**. As its name suggests, monopolistically competitive markets combine aspects of the perfectly competitive and monopoly models. Specifically, these are markets in which firms produce similar but differentiated products. An example of such a market is book publishing. Each particular title is unique and distinctive, but there are thousands of titles for you to choose from when you are looking for a book. Other examples include restaurants, clothing, breakfast cereals, and many local service industries.

Because the product of each firm is differentiated—meaning that you can tell the difference between its product and those of other firms—the firm faces a downward-sloping demand curve. As a result, each firm chooses its output in the same way a monopoly firm does, by finding the point at which its marginal revenue equals its marginal cost. Because the firm's demand curve slopes downward, marginal revenue is less than price, so at this point the market price is greater than the marginal cost of production.

We have seen that at the profit-maximizing quantity, a monopolist will earn positive economic profits. In a monopolistically competitive market, however, if firms are earning positive profits this will lead to the entry of new firms supplying similar goods or services. As the range of choices available to consumers expands, existing firms will see their demand curves shift to the left, causing profits to fall.

Because there are no barriers to entry in a monopolistically competitive market, entry will continue until profits have been reduced to zero. If at some point profits fall below zero, there will be exit from the industry, which will continue until the zero economic profit equilibrium is restored.

A full analysis of the welfare properties of monopolistically competitive markets requires more sophisticated mathematical analysis. But there are several points to note about such markets. First, because price exceeds marginal cost, there is some social inefficiency: there are consumers who value the product at more than the cost of increasing production. The failure to complete these transactions is a failure to fully exploit mutually beneficial exchanges. This failure occurs because of the firm's monopoly incentive to restrain production. Second, the diversification of products that results from the efforts of firms to create a distinctive identity for their product creates benefits for consumers by increasing the range of choices available to them.

CREATIVE DESTRUCTION: THE PROFIT MOTIVE AND THE SOURCES OF ECONOMIC CHANGE

When we considered the entry and exit of producers in a competitive market in the previous section, we came to the somewhat surprising conclusion that even

though producers in a perfectly competitive market would earn zero profits, they would be satisfied with this result. In part this is a consequence of our definition of economic profits, which factors in the opportunity cost of all of the resources employed, including the business owner's time.

Economic profits, then, are an additional payment above and beyond the compensation that can be earned in the next best alternative activity. We should not be surprised, then, that self-interested economic agents should seek to identify or create opportunities to earn economic profits. One important way that they can do this is by escaping the constraints of competitive markets. When producers can create barriers to entry, they can create situations of imperfect competition in which they are able to earn economic profits.

As we saw earlier, in comparison to a hypothetical competitive market outcome, imperfectly competitive markets create inefficiencies because producers restrict supply as part of their effort to maximize profits. But, this comparison of different market structures fails to capture an important aspect of the actual way in which economies evolve over time. One of the important routes that firms take to establish market power is innovation.

Entrepreneurs are individuals who take on the risk of attempting to create new products or services, establish new markets, or develop new methods of production. The rewards of entrepreneurship are the economic profits that can be earned by being the first to market with a new product. In the case of scientific innovation, entrepreneurs can obtain a legal monopoly through patents; but in other cases market power arises because of their ability to differentiate the goods or services they produce from other products in the market. Entrepreneurs can differentiate their product by defining the desirable characteristics of their product or by the possession of trade secrets.

At the same time that innovation helps to create barriers to entry that reward the innovator with economic profits, it also serves to break down existing market imperfections because the existence of profits encourages efforts to invent around existing barriers to entry. Examples of this include the development of satellite television in competition with the monopoly of cable television and the efforts of mobile phone manufacturers to imitate the Apple iPhone.

The continued development of new and improved products is one of the key sources of long-run improvements in well-being, a fact that economist Joseph Schumpeter sought to capture when he described the impact of entrepreneurs as a type of "creative destruction." The essential catalyst of creative destruction is the opportunity to earn economic profits. But, the inefficiency in resource allocation that creates these economic profits is—in the view of many economists—small in comparison to the benefits of the innovation to which it gives rise.

MARKET FAILURES

Our study of competitive markets has revealed the remarkable way in which they coordinate the self-interested actions of market participants to produce socially desirable outcomes. Market prices ration scarce goods and services so that they go to those consumers who value them most highly. At the same time, the search for economic profits encourages the allocation of scarce resources toward the production of those goods and services that are valued most highly.

Of course, not all markets fit the ideal of perfect competition. But, in these cases, the opportunities to profit by facilitating mutually beneficial exchange encourages private actors to move closer to the socially efficient outcome. Monopolists, for example, have an incentive to find ways to price discriminate, while the opportunities for private profit tend to break down efforts by cartels to restrict output. Where these forces are not sufficient, economic theory can help us to evaluate possible policy solutions.

There are some circumstances, however, in which competitive markets will fail to produce socially desirable outcomes. These circumstances are called **market failures**. Most instances of market failure can be grouped into two broad categories.

The first type of market failure arises because of externalities. An **externality** arises when the actions of one person affect the well-being of someone else, but neither party pays nor is paid for these effects. When the effect of these actions is beneficial, it is called a *positive externality*; when the effect of these actions causes harm, it is called a *negative externality.* The second type of market failure occurs when the institution of private property breaks down. When it is impossible to establish private property rights in important economic goods or services, we refer to the goods or services in question as **public goods**.

Addressing the problems of externalities and public goods is one of the most compelling roles for government in our economy. Economics allows us to understand more precisely how the characteristics of externalities and public goods affect market outcomes and can provide important guidance when considering the options for policies to correct these market failures.

Externalities

A widely cited example of an externality involves beekeepers and apple growers. In the course of producing honey, the bees pollinate the apple trees, increasing the size and value of the farmer's crop. Since the value of the apple crop does not figure in the beekeeper's costs or benefits, it constitutes an externality. Since the farmer benefits from the beekeeper's actions, it is a positive externality.

One can easily find many other examples of similar types of interactions. For example, when movie studios release movies on Blu-ray discs, they increase consumer demand for Blu-ray disc players, which increases the revenue of their manufacturers. In this instance, the externality operates in the other direction as well because increases in the sale of Blu-ray players increases consumer demand for the studios' movies. When a new highway interchange is built on a busy freeway, it increases traffic on nearby roads, raising their value as business locations. This is a positive externality for the landowners.

Externalities can have negative consequences as well. If one of your neighbors fails to maintain his house, it can have a depressing effect on the value of your home. Pollution is another example of a negative externality. Runoff from farm fields containing traces of fertilizers and pesticides commonly finds its way into nearby rivers. As a result, downstream communities that take their drinking water from these rivers have to spend more money treating this water before distributing it. Concerns about climate change have focused attention on the negative consequences of carbon-dioxide (CO_2) emissions. Again, because the businesses and individuals do not take into account the negative impact of their activities on the global climate, this is an externality.

The Effect of Externalities on Resource Allocation

In general, there will be too little of an activity that generates positive externalities and too much of an activity that generates negative externalities. To see this, let's consider the example of a paper plant. As a by-product of producing paper, the plant also produces polluted waste that it dumps untreated into a nearby river. Figure 26A shows the market for the plant's primary product: paper. The firm's supply curve is upward sloping, reflecting the fact that its marginal costs are increasing as production rises. The demand curve is drawn as downward sloping.

As we have seen, the competitive market equilibrium occurs at the point where the demand and supply curves intersect. This is the quantity the profit-maximizing firm will choose to supply. But this decision does not take account of the social costs that the firm's actions impose on the downstream community. For simplicity's sake, let's assume that the cost of removing the pollutant produced by the paper company is a constant amount of $15 per unit of paper that it produces.

The true social cost of the firm's production is equal to the firm's marginal cost plus the cost of treating the pollution it produces. We can represent the true social cost of production, then, by drawing a new supply curve that is shifted up by $15 at every point. This is illustrated in Figure 26B. Notice that the curve representing total social costs intersects the demand curve above and to the left of the private market equilibrium. The socially optimal level of production is lower than the amount supplied by a profit-maximizing firm because the firm fails to take account of the external costs.

An important implication of the analysis illustrated in Figure 26 is that the optimal level of a negative externality is not zero. Rather, there is likely to be some positive level of the externality that will be consistent with maximizing consumer and producer surplus. This is true because the activity that generates the externality has a positive value, and the cost of reducing this activity too greatly will outweigh the additional benefits of reducing the externality.

We can use a similar approach to analyze a case of positive externalities. Figure 27 illustrates the market for honey that a beekeeper faces. Here the demand curve reflects only the value that consumers place on the honey the beekeeper supplies. But, since each unit of honey also results in an increase in the value of the crop of nearby orchards, the true social value of the activity is shifted up by the amount of this increase. As

this analysis suggests, the resulting equilibrium occurs above and to the right of the equilibrium when the externality is not accounted for.

Private Responses to Externalities

The existence of externalities creates incentives for market participants to attempt to solve the problems they create. In the case of the beekeeper and the apple grower, total revenues would increase if the beekeeper expanded his production. This additional revenue could be divided between the two parties so that both increased their profits. Similarly, in the case of the negative externality caused by the paper company, the downstream community could pay the paper company to produce less or to take other steps to prevent the pollutant from entering the river in the first place. Again, such an arrangement would leave both parties better off.

Another approach to solving the problem of externalities is to internalize them by combining the activities that produce the externality within a single company. For example, Netflix started out by streaming movies and TV shows that had been produced by other companies. In recent years, however, Netflix has started producing its own content; in doing so, Netflix is combining activities that could produce externalities within a single company.

As long as the parties involved can negotiate with each other, the private market should be able to resolve the inefficiencies created by externalities. This insight was first reached by Ronald Coase and is often called the **Coase Theorem**. To illustrate this point, consider the case of two neighbors, Tad and Sue. Tad lets his grass grow long and does not take good care of his yard. Sue must look at the yard from her front porch, which reduces her enjoyment, and it also lowers the value of her house. She can offer to pay Tad to take better care of his yard. So long as the value she places on the appearance of Tad's yard exceeds the cost to him of caring for it, they will be able to negotiate an appropriate payment that makes both of them better off. Of course, if the benefit to Sue is less than the cost to Tad, then they will not reach an agreement, but in this case, that is the efficient solution.

Notice that we have assumed that Tad is under no obligation to maintain his yard. Suppose, however, that a local ordinance requires that he do so, and Sue can compel him to do so by reporting him to city officials. In this circumstance, Tad could negotiate with Sue, offering to pay her to put up with his poorly maintained yard. If the value Sue places on having a well-kept yard to look at is less than the cost to Tad of cleaning it up, they will be able to arrive at a bargain where he pays her to put up with his yard. If his cost of cleaning up the yard is less than the value Sue places on having his yard well maintained, then they will not reach a bargain, and he will be obliged to take care of his yard. But in this case, this is the efficient solution.

As this example illustrates, Tad and Sue will arrive at the efficient solution regardless of whether Tad is free to ignore the upkeep of his yard or is required to keep it neat. One of the important insights of the Coase Theorem is that the initial distribution of rights does not affect the ability of the parties to come to an efficient agreement. So long as the property rights are clearly defined, the parties will arrive at the efficient solution.

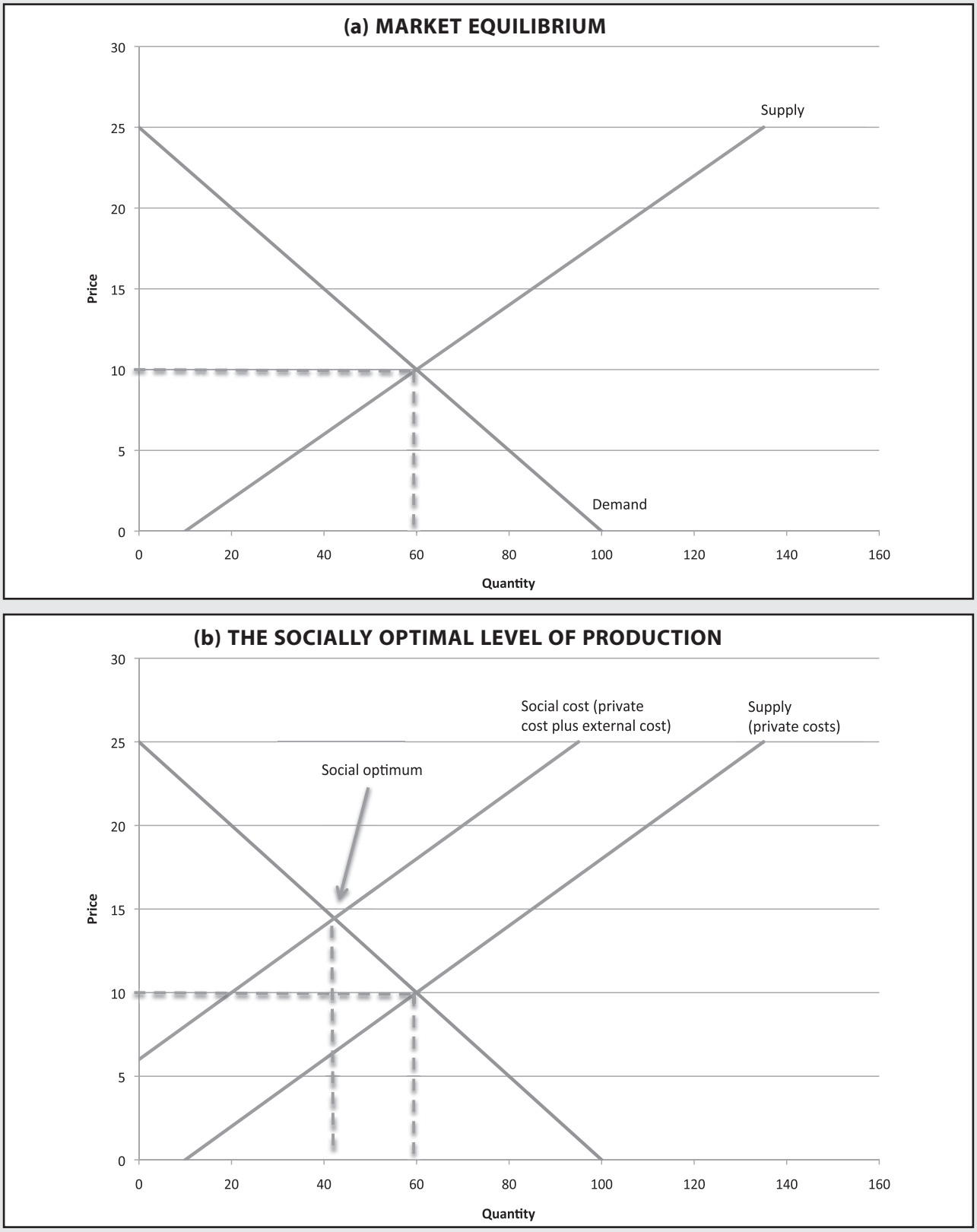
Of course, if matters were this simple, then externalities would be only a minor footnote rather than an important topic in economics. The reason they are often a problem is that in many cases property rights are poorly defined, or nonexistent, and the costs of negotiating between the affected parties are prohibitively high. As an example, consider an oil refinery that pollutes the water of a nearby river used for recreation or fishing. The pollution can cause significant damage even miles downriver from the refinery, so the effects can be extremely diffuse.

Given that the pollution can negatively impact so many people, the total effects are large. However, the individuals affected have little incentive to negotiate with the oil refinery to reduce its pollution—and Coase pointed out that the damages are reciprocal in nature. Prohibiting the pollution would help those who fish in the river while harming the refinery whereas allowing the pollution would help the refinery while harming those who use the river. In any case, the costs of negotiating between the many people who use the river and the oil refinery are prohibitively high, so a negotiated solution is highly unlikely to emerge.

Government Regulation of Externalities

When private bargaining fails, governments can sometimes step in to resolve the matter. Since the problem of externalities arises because the actions of private parties do not fully reflect the social costs or

FIGURE 26



The Impact of a Negative Externality on the Socially Optimal Level of Production

benefits of their actions, one solution is to use taxes or subsidies to correct this problem. An example of the use of taxes to address negative externalities can be seen in cases where some cities have implemented or are considering implementing congestion pricing. Under such a system, drivers in certain parts of cities must pay fees to drive. New York City became the first U.S. city to approve congestion pricing in April 2019. The city plans to hold public hearings on the topic of congestion pricing and will conduct environmental reviews before implementing the plan. For-hire vehicles, including taxis and app-based services like Uber and Lyft, already pay fees in some areas.[8]

Using taxes to remedy the effect of externalities is most effective when it is possible to estimate the value of the externality. In many cases, this information is not readily available. So it may be more effective to reduce a negative externality by establishing a quota limiting the activity that produces the externality. If such an approach were to be used to reduce traffic congestion, then a target number of vehicles would be set and only that many permits would be issued. Of course, a problem with this approach is that the drivers who get permits may not be those who value them most highly. But, this can be resolved by creating a market in which drivers can buy and sell permits.

The United States Environmental Protection Agency (EPA) has used this approach to deal with sulfur dioxide emissions. After establishing a maximum level of emissions, the EPA auctioned off the rights to emit sulfur dioxide to the highest bidders. A similar approach is used in California, which created an emissions trading system in 2013. The California system sets a cap on greenhouse gas emissions each year but allows companies to buy and sell pollution credits in auction markets, which allows them some flexibility in how they achieve the emissions reductions.

Property Rights

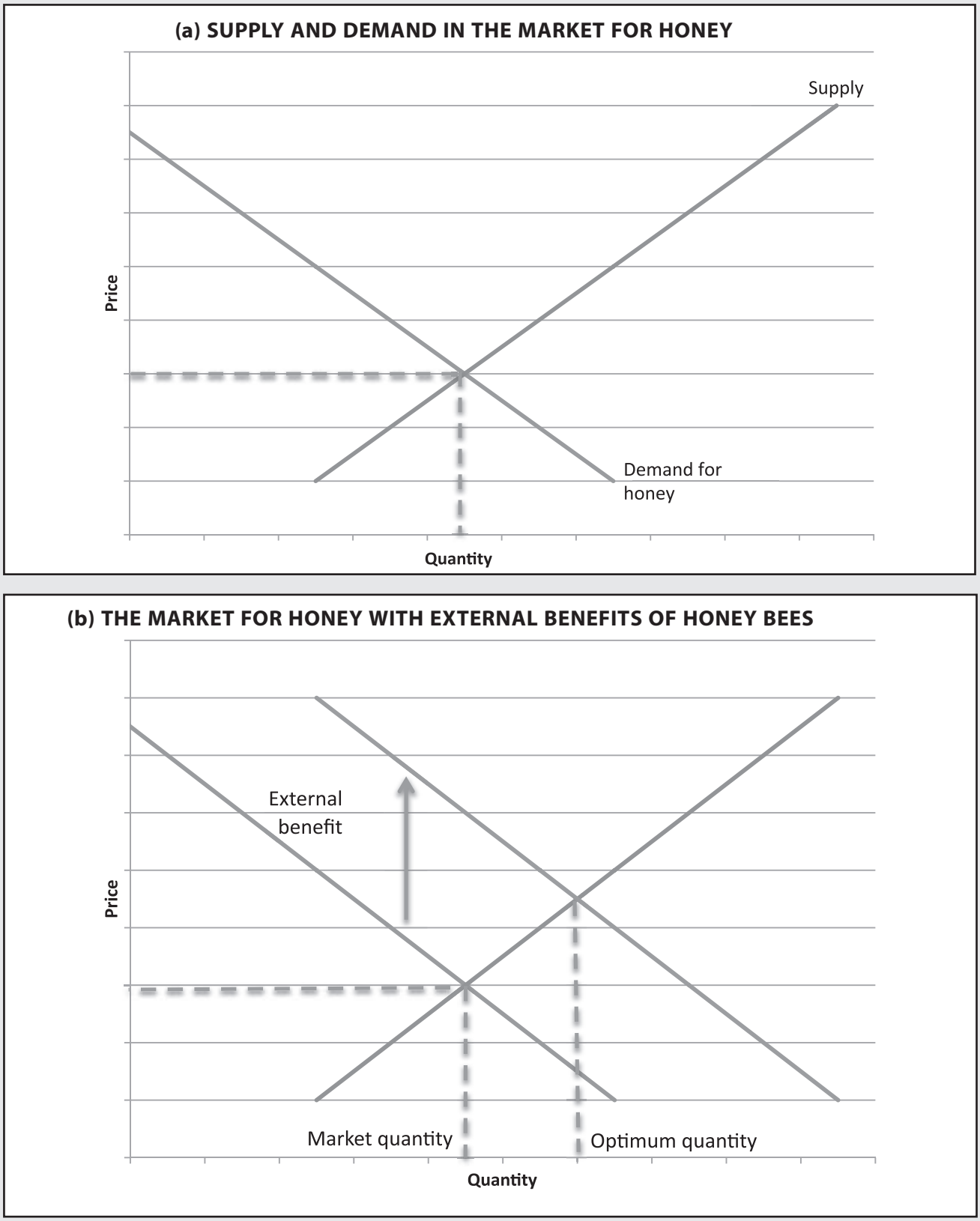
Having grown up in a market economy, the existence of private property seems quite natural to most of us. However, the institution of property rights is not a natural occurrence; it is a social innovation. The importance of this innovation becomes clear when we consider what happens when valuable economic resources have no owner.

To illustrate the importance of private property, let's consider what happens to property that no one owns in this simple example. A village located next to a lake has six residents, each of whom has $100 in savings they can use to either purchase a government bond that pays 15 percent interest, or to purchase a fishing boat necessary to catch fish in the lake. The number of fish each resident can catch depends on the number of residents who catch fish. This relationship is shown in the table in Figure 28. If only one villager purchases a boat, then he/she can catch $130 worth of fish, and his/her net income is $30 ($130 in income minus the $100 cost of the boat). If two villagers buy boats, then they catch $120 worth of fish each, and each earns a net income of $20. The average value of fish caught declines as additional villagers buy boats because they are all fishing in the same lake, and as each one depletes the fish population, it becomes increasingly difficult for others to find fish.

Imagine, first, that the villagers decide one at a time whether to purchase a boat or to invest in the government bond, and that the decisions are public. How many villagers will purchase boats? If a villager purchases the government bond, he/she will earn $15 interest income at the end of the year. He/she should only purchase a boat if his/her income from fishing is $15 or more. From the table, we can see that three villagers will purchase boats. After three boats are purchased, the fourth villager will see that his/her income from fishing will only be $10 and will choose to purchase a government bond. Total income in the village will be $90 per year. Three villagers will earn $15 each from fishing ($3 \times \$15 = \$45$), and three villagers will earn $15 each from bonds ($3 \times \$15 = \$45$).

Is this the socially optimal allocation of resources? Suppose that the villagers got together and decided collectively how to allocate their resources? To maximize village revenue, the villagers should invest in fishing boats only if the marginal contribution to village revenue exceeds the marginal cost. In this case, the cost of purchasing a boat is the opportunity cost of not purchasing the government bond, or $15. The table in Figure 28(b) calculates the marginal income from fishing for each additional fisherman. The marginal revenue generated by the first boat is $30. But the purchase of a second boat raises income from fishing only to $40, so the marginal contribution to village revenue is $10. The villagers should purchase just one boat. Total income will be $30 from fishing, plus $75 = $5 \cdot \$15$ from interest income, or $105.

FIGURE 27



The Effect of External Benefits in the Market for Honey

FIGURE 28

(a) VILLAGE INCOME

NUMBER OF FISHERMEN	AVERAGE VALUE OF FISH CAUGHT	NET INCOME FROM FISHING	+	INTEREST INCOME	=	TOTAL VILLAGE INCOME
1	130	30		75		105
2	120	40		60		100
3	115	45		45		90
4	110	40		30		70
5	105	25		15		40
6	100	0		0		0

(b) MARGINAL REVENUE

NUMBER OF FISHERMEN	MARGINAL REVENUE FROM FISHING
1	30
2	10
3	5
4	–5
5	–15
6	–20

When Some Resources Are Not Private Property

When the villagers make their choices independently, they fail to account for the external effects of their fishing on the income of other boat owners. Because the fish in the lake are a common resource, one villager's decision to purchase a boat and catch fish reduces the income that others can earn from fishing. The villagers do better when they decide collectively because they internalize the externality.

The Effects of Private Ownership

The example we have just considered is a version of a problem that is often referred to as the **tragedy of the commons**. When a resource is owned jointly, no one takes account of the negative externalities caused by overuse. We have seen in the previous section that taxes or other regulations can ameliorate the effects of externalities. But a simpler solution is to create property rights in the resource.

Suppose that in the previous example we allow for one of the villagers to purchase the lake. The owner can then decide how many boats to allow on the lake. We have seen that the most profitable choice is to allow a single boat on the lake, which generates an income of $30. So, if the lake is privately owned, resources will be allocated in the most efficient manner.

How much would one of the villagers be willing to pay to purchase the lake? Since the opportunity cost of investing in the boat is the $15 forgone interest, the owner of the lake would earn $15 profit if they could use the lake for free. The most one of the villagers would be willing to pay to purchase the lake is $100. At this price, the purchase of the lake yields the same return as buying a government bond. If the villagers invest the $100 paid by the purchaser in a government bond, then they can divide the additional income that it generates, thus raising all of their incomes.

Public and Private Goods

In the example we just considered, private ownership of the lake solves the allocation problem created by a common resource. But private ownership may not always be a feasible solution. Some resources like the oceans or the atmosphere are not easily privatized. Recent developments in economic theory have helped to clarify the characteristics of goods that can easily be privatized versus those that cannot. To understand this distinction, we need to differentiate goods along two dimensions.

The first of these dimensions is the extent of *rivalry in consumption*. Most goods have the characteristic that one person's consumption of them reduces the amount that is available for others. For example, if you consume a slice of pizza, then there is one less slice available for your friend. We say that pizza is a **rival good**. On the other hand, when you listen to a radio broadcast, your enjoyment of it does not diminish the ability of other listeners to enjoy it as well. The radio broadcast is a non-rival good. Note that rivalry is not always a black or white condition. On a lightly traveled highway, the presence of one driver may not interfere with the value of the road to other drivers. But as congestion increases, and traffic approaches the road's capacity, then additional drivers will begin to have a negative effect.

The second dimension is the degree of **excludability**. This describes the ability to control who consumes the good. National defense is a non-excludable good. If the military protects the country from invasion, all of its citizens benefit from this protection. Similarly, if your city puts on a fireworks display on the Fourth of July, it is difficult to prevent people from seeing it. In contrast, it is easy to exclude someone from consuming a slice of pizza by simply not giving it to them. Figure 29 summarizes this two-way categorization.

Private Goods

Conventional private goods are characterized by a high degree of rivalry in consumption and a high degree of excludability. This corresponds to the entry in the upper left corner of Figure 29. Examples of such goods are all around us—they include pizza, gasoline, and haircuts.

Common Resources

In the lower left-hand corner of the table in Figure 29 are goods that have a high degree of rivalry in consumption, but a low degree of excludability. These are common resources that suffer from the problem of the tragedy of the commons: because no one owns them, they will tend to be over-utilized. Fish in the ocean provide an illustration. Every fish that is caught by one person is not available to be caught by someone else. But, because it is difficult to limit access, it is difficult to make the fish a private good.

Goods that are rival in consumption but not owned are the source of externalities. As we discussed earlier, there are strong incentives for private actors to find ways to internalize these externalities. When these incentives are insufficient, however, public policy can seek to establish property rights or use taxes and other types of regulatory controls to address the inefficiencies created by a common resource.

Collective Goods

Goods that have a low degree of rivalry but a high degree of excludability (upper right corner) are termed collective goods. Such goods can easily be privatized, but they are often natural monopolies because non-rivalry in consumption means that the marginal cost of producing them is zero or close to zero. Examples include satellite radio and pay-per-view television.

A monopoly can profitably supply these goods, but it has an incentive to set the price too high and supply too little, thus leading to an inefficient outcome. This characteristic may lead to regulation or to government provision of collective goods.

Public Goods

The final category of goods combines non-rivalry in consumption with non-excludability. These are true *public goods*. Because it is difficult to exclude consumers, it is difficult for private actors to charge for these goods. And, because they are non-rival in consumption, the marginal cost of their provision is close to zero.

Many public goods are provided by the government. But, in some instances public goods, such as television and radio broadcasts, are supported in other ways—such as through advertising or private donations. It is likely, however, that when public goods are supplied this way that the quantity supplied will be too low. One illustration of this is the vastly greater number of channels available via cable and satellite TV than via over the air broadcast. Because subscribers to cable or satellite providers pay directly for programming, a much greater variety of content is available than can be

FIGURE 29

	EXTENT OF RIVALRY IN CONSUMPTION	
EXTENT OF EXCLUDABILITY	**HIGH**	**LOW**
HIGH	**Private Goods** • Pizza • Haircuts • Gasoline	**Collective Goods** • Satellite radio • Websites • Pay-per-view movies
LOW	**Common Resources** • Fish in the ocean • The environment • City streets	**Public Goods** • Radio broadcast • Tornado siren • National defense

Four Types of Goods

supported by advertising alone.

INSTITUTIONS, ORGANIZATIONS, AND GOVERNMENT

One of the central insights of economics is how markets help to convert the actions of self-interested individuals into socially desirable outcomes. As we have seen, however, this conclusion may not hold when producers have a degree of market power, or when market failures occur because of externalities or circumstances that make it difficult to define private property rights. In these cases, collective decision-making mechanisms may be necessary to overcome the effects of these departures from perfect competition.

Understanding how collective decision-making processes have emerged in modern economies is a complex topic, and we can only begin to touch on the most important insights of this branch of economics here. But the topic is, nevertheless, vitally important. Differences in standards of living around the world are vast today, and economists believe that in large part these differences are due to variations in the success with which different societies have dealt with the challenge of organizing collective decision-making.

Collective decision-making begins with **institutions**. Institutions are both formal and informal rules that structure human interaction. Most markets are, in this sense, institutions; so too are marriage and child-rearing practices and norms such as how much to tip a waiter in a restaurant. Like institutions, *organizations* help to organize human interaction, but do so through formal rules and structures. Commodity and stock exchanges are organizations as are corporations and organized religions.

An important limitation constraining institutions and organizations is the need for voluntary cooperation. For this reason, self-interested individuals will conform to social institutions or participate in voluntary organizations only so long as that cooperation makes them better off. Cooperation in some contexts can indeed improve social welfare, but, as we have seen in the case of cartels, there can be powerful incentives to cheat on voluntary agreements.

In comparison to private institutions and organizations, government possesses two distinctive powers. The first of these is the ability to tax its citizens. Private businesses can earn revenue only by selling their products. Consumers will only buy their products if they value them more than their prices. In contrast, government can compel the payment of taxes. Of course

this power is not absolute. In the United States, citizens are free to move between cities, counties, and states, and they can vote with their feet if they dislike the level of taxation in one area by moving someplace else. Similarly, citizens of any of the member countries of the European Union are free to move from one country to another. Other types of international mobility are more limited. The United States imposes significant restrictions, for example, on legal immigration into the country, as do most other countries.

The second distinctive power of government is the legal monopoly on the legitimate use of force. This power is used to restrain criminals, compel military service, and to protect national security. Clearly the government's ability to use force underlies its ability to collect taxes. The government's ability to compel citizens to act in ways that are not in their individual self-interest is also essential to supporting a system of private property on which the whole system of voluntary exchange rests.

Government also helps to support a broader range of voluntary cooperation than would otherwise be possible through activities such as the enforcement of contractual obligations. Contracts represent agreements entered into voluntarily because both parties anticipate that they will gain from the agreement. But, subsequent changes may cause one party to regret having entered into the agreement. Without the courts to enforce such agreements, individuals would be far more reluctant to enter into them in the first place.

The powers that governments possess are truly awesome. As we have suggested, they can be used to fix problems that prevent private economic actors from achieving efficient outcomes. But government can also be a source of inefficiency and corruption. We must remember that both elected officials and government employees are themselves self-interested economic agents, whose interests may diverge from those of the larger community. Economics can help us to identify and understand these conflicting forces more clearly.

Pork Barrel Politics

Pork barrel politics refers to the proclivity of elected officials to introduce projects that steer money to their communities. Such projects are often popular with the voters who matter for the particular legislator, but the combined effect of these projects is to increase the cost of government.

To understand this problem, it may help to think about an experience you may have had before. You have gone out with four of your friends to a restaurant and agreed that to simplify matters you will split the bill evenly. When the waiter asks if you want dessert, you look at the menu and see that you can purchase a hot fudge sundae for $4. You value the sundae at no more than $3, so if you were dining alone you would skip dessert. But you do the math and realize that if you order the sundae your share of the bill will only increase by $0.80 (= $4/5). As result, you order it. Not surprisingly, your friends make a similar calculation for themselves, and you wind up paying an additional $4 each.

A similar logic is at work in the legislative process. A member of the House of Representatives might, for example, be able to introduce an amendment to a bill that will bring $100 million in benefits to his/her district. The cost of the program to the federal government is $150 million (so clearly the costs outweigh the benefits). But the cost to the community is just a small fraction of this, since it will be supported by all taxpayers, not just those in the affected community. For the representative's constituency this is a terrific deal. They get $100 million in benefits for a small fraction of this amount in increased taxes.

Of course the legislation has to get the support of a majority of the House members to be passed into law. Why would a legislator representing another district support legislation that will increase the cost to his constituents without producing any benefits? The answer is that by supporting his/her colleague's pet project, the legislator can win support for his/her own pet project. This vote trading activity is commonly called **logrolling**, and much like the restaurant example above, it accounts for a certain amount of wasteful government spending.

Rent Seeking

A related source of inefficiency arises because the gains from many government programs are concentrated, while the costs are spread widely. An example of this problem is the current U.S. policy of price supports for domestic sugar producers. These supports combined with restrictions on the importation of cheaper sugar from outside the country keep U.S.

sugar prices at nearly twice world levels. A report in 2017 found that resulting losses to households are from $2.4–$4 billion in total. Spread across a population of over 300 million, the cost per person is relatively small. But the benefits to the small number of sugar producers are much larger. Sugar growers have a strong and compelling motivation to hire lobbyists and spend money to influence key legislators to continue price supports. Most voters, however, are unaware of this policy, and even those who are aware of it would be unlikely to find it worth the effort to oppose it.

Even when the overall benefits of projects exceed their costs, they may generate wasteful resource allocation. Competition to influence the location of expensive federally supported activities can lead to the expenditure of large amounts of money seeking to influence decision-makers.

In general, socially unproductive activities that seek simply to direct economic benefits to one set of actors rather than another are called **rent seeking**.

What Is the Proper Role for Government?

Determining what functions the government should play, how big it should be, and how much it should regulate are normative judgments that must be made on grounds that extend beyond purely economic considerations. Nonetheless, economics helps to illuminate the issues and frame these choices more clearly.

Government is not essential to the establishment of a market economy, but the enforcement of the rule of law helps to support a much broader range of transactions than would be possible without it. Most of us are willing to accept the small loss of individual autonomy for the protection of property and the individual security that this entails. But, unconstrained government can become an intrusive force that can substantially reduce individual freedoms.

Similarly, government can, as we noted earlier, correct market failures arising because of externalities and public goods; however, the ability to rectify these problems also gives rise to inefficiencies. People may genuinely differ in their evaluation of the relative costs and benefits of these trade-offs.

SECTION II SUMMARY

- The interaction of supply and demand in markets is the central topic of microeconomics.
- A market consists of all the buyers and sellers of a particular good or service.
- The model of a perfectly competitive market applies to situations in which the numbers of buyers and sellers is large, all the market participants are well informed about the market price, and the good or service being exchanged is highly standardized.
- The demand curve graphs the quantity of a good or service that buyers are willing and able to purchase at each price. According to the law of demand, the quantity demanded is negatively related to the price.
- The position of the demand curve depends on income, the prices of related goods, tastes, expectations, and the number of buyers.
- The supply curve graphs the quantity of a good or service that producers are willing and able to supply at each price. According to the law of supply, the quantity supplied is a positive function of the price.
- The position of the supply curve depends on the prices of inputs used in the production of the good or service being exchanged, the technology used to produce it, expectations, and the number of sellers.
- In a perfectly competitive market, equilibrium occurs when no market participant has any reason to alter their behavior. The only point that satisfies this requirement is the point where the supply and demand curves intersect.
- The competitive market equilibrium maximizes the combined benefits or total surplus of market participants.
- One important use of the competitive market model is to analyze how changes in economic conditions affect the equilibrium price and quantity as well as the surplus of market participants.
- Elasticity provides a measure of the responsiveness of supply and demand to price changes that is independent of the units used to measure price and quantity.

- Governments intervene in markets for a variety of reasons. They may set price ceilings or price floors. Governments may also impose taxes on certain types of transactions to raise revenues to pay for essential services.
- Trade makes people better off. International trade increases total surplus.
- Not everyone in an economy benefits from trade, however, which explains why there is often opposition to free trade.
- Firms are the economic actors who supply goods and services by combining labor, capital, raw materials, and other inputs to produce the products consumers want to purchase. Firms seek to maximize their economic profits.
- In a competitive market, the entry and exit of firms insures that the firms in the market earn zero economic profits.
- The model of perfect competition cannot be applied to all parts of the economy. There are many different types of imperfectly competitive markets. The most important cases are monopoly (a single supplier), oligopoly (a small number of suppliers), and monopolistic competition (many suppliers of similar but differentiated products).
- Imperfect competition arises because of barriers to entry into the market.
- Relative to perfectly competitive markets, imperfect competition results in a lower equilibrium quantity and a higher equilibrium price. This outcome causes total surplus to be lower than it would be in a competitive market.
- The economic profits that arise in imperfectly competitive markets are the incentive that motivates entrepreneurs to develop new goods and services, new markets, or new methods of production.
- Market failures occur when externalities or breakdowns in the system of private property cause market outcomes to deviate from the socially efficient outcome.
- Externalities occur when there are important economic interactions that do not take place through markets. One solution is to create a market for these interactions; another solution is government regulation.
- All goods and services can be classified along two dimensions: (1) the extent of rivalry in consumption and (2) the ease of excludability. This two-way classification allows us to identify four categories of goods and services: private goods, common resources, collective goods, and public goods.
- Institutions, organizations, and governments help to organize human interactions through formal and informal rules. Governments are distinguished from private organizations through their ability to compel citizens to pay taxes and their monopoly on the legitimate use of force.
- Government is an important factor in enhancing well-being through its support of private property and market transactions, but pork barrel politics and rent seeking are inefficient outcomes that arise because of how governments operate.

SECTION III
Macroeconomics

As is true in the physical sciences, the methods and approaches that are most effective in understanding economic phenomena depend on the type of questions we are asking. For example, biologists studying the operation of particular molecules use models and types of data that are different from those used by scientists who wish to understand larger ecosystems, even though the same fundamental principles apply.

In the same way, when economists wish to understand the performance of an entire economy—how much it produces or what causes national **unemployment rates** to fluctuate—the models and data they use are different from those that they use when they want to understand what happens in specific markets, such as the market for petroleum. The branch of economics that studies the performance of national economies is called macroeconomics.

This section of the resource guide provides an introduction to the major questions addressed in macroeconomics and describes the most important approaches to these questions. Broadly speaking, macroeconomics is concerned with two questions. The first concerns the factors that determine the long-run growth in the size of economies, the standard of living that they provide for their participants, and the price level. The second issue concerns the causes and consequences of short-run fluctuations in the level of economic activity, **unemployment**, and **inflation**.

We will begin this part of the resource guide by presenting some evidence about these issues that helps to motivate our subsequent analysis and by discussing the types of aggregate economic indicators that are used to describe the performance of the aggregate economy. These include measures such as **Gross Domestic Product (GDP)**, the cost of living, and the unemployment rate. These measures figure prominently in public discussion of the state of the economy and economic policy. Knowing how these concepts are defined and interpreted is important for everyone and is essential to understanding the behavior of the economy.

The remainder of this part of the resource guide will develop a theoretical framework for analyzing aggregate economic performance. We begin by describing factors that determine the size of an economy in the long run. We then will consider the role of the financial system and the uses of money. Finally, we will turn to the causes of short-run fluctuations in economic activity.

MACROECONOMIC ISSUES

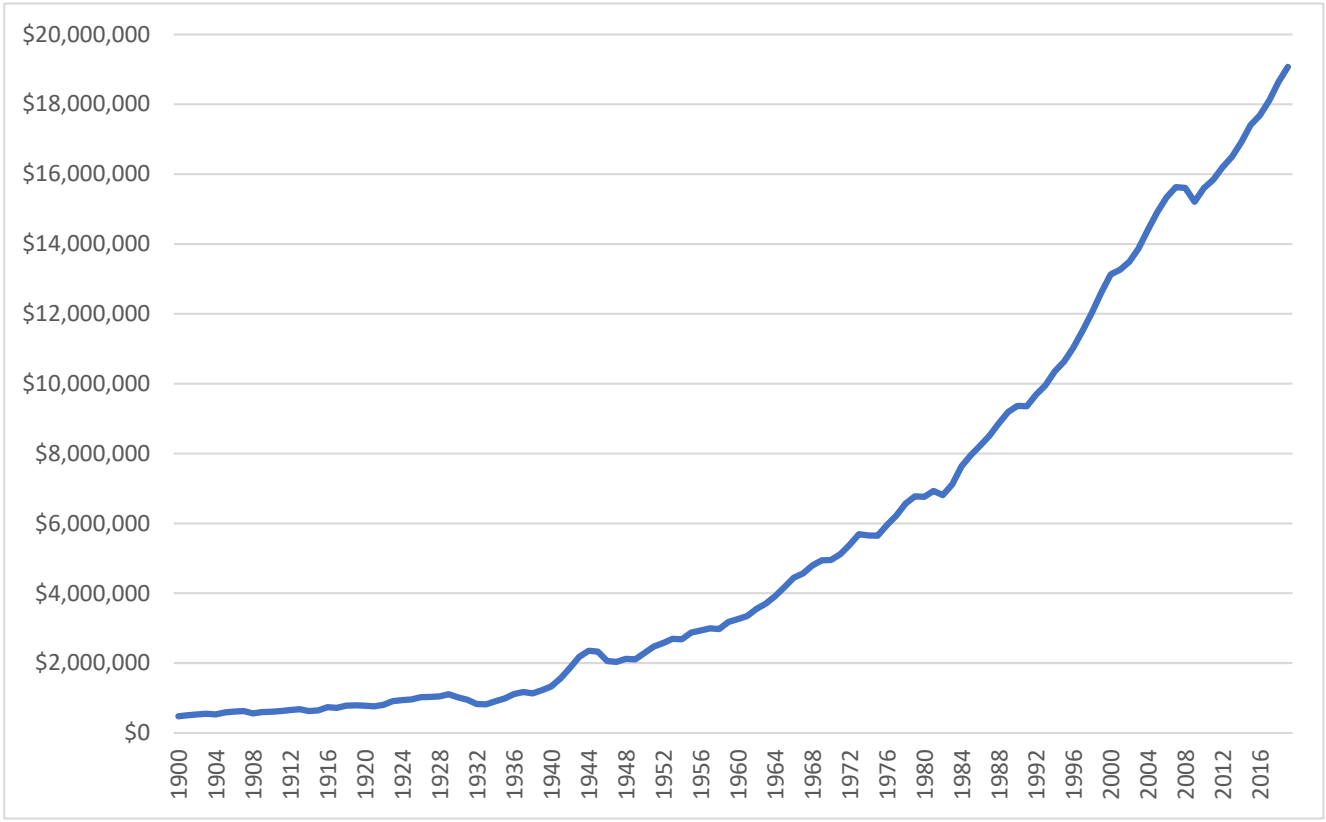
We have said that macroeconomics is concerned with the performance of national economies. To get a more concrete sense of why this is important and what it means, it will be helpful for us to look at a number of aspects of the U.S. economy.

Economic Growth and Living Standards

One of the most remarkable facts about the U.S. economy is its long-run history of growth. Figure 30 illustrates the growth of total output of the U.S. economy from 1900 to the present. The measure of output used in Figure 30 is real Gross Domestic Product (GDP). This is a measure of the total quantity of goods and services produced in the economy, adjusted to remove the effects of inflation. We will discuss in more detail how output is measured shortly, but for now, let's focus on what Figure 30 shows.

According to these data, since 1900, the total real output of the U.S. economy has increased by a factor of nearly forty.[9] There are some small ups and downs apparent in this chart—most notably the decline in output between 1929 and 1933 (the Great Depression) and the expansion of output from 1941 to 1945 (World War II). Viewed on this time scale, however, the

FIGURE 30



SOURCE: Louis Johnston and Samuel H. Williamson, "What Was the U.S. GDP Then?" MeasuringWorth, 2020
URL: http://www.measuringworth.org/usgdp/. All values expressed in 2012 prices.

Real Output of the U.S. Economy, 1900–2019

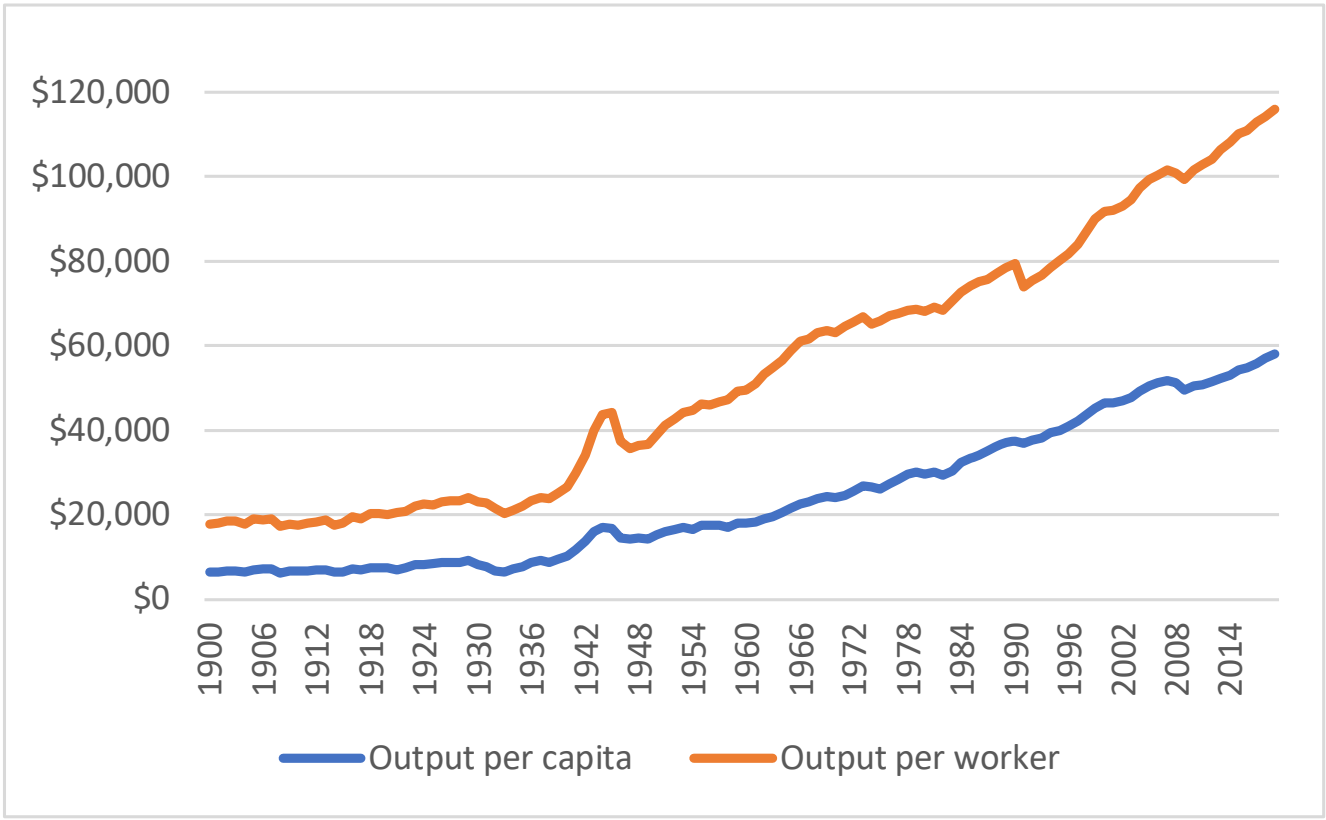
impact of these events is dwarfed by the expansion of the size of the overall economy.

At the level of the overall economy, what we can consume is limited by what we produce. One reason for the rising level of production historically has been the growth in population. More people can produce more output. But output has grown much faster than population. Since 1900, the U.S. population has increased by a factor of more than four. Combining this information with the data in Figure 30 implies the average output per person has increased by a factor of nine. Figure 31 illustrates the growth of output per person. Economists refer to this quantity as output (GDP) **per capita**. The term "per capita" is a Latin phrase literally meaning "per head," which is commonly used to denote averages calculated for an entire population.

While average output per capita provides an indication of what the typical person can consume, economists are also interested in changes in what the average person can produce. The economy's total output divided by the total number of workers employed is called **average labor productivity**. This is a measure of how much the typical worker can produce. The second (higher) line in Figure 31 shows the history of average labor productivity since 1900.

The average output per person in the U.S. economy in 2019 was over $65,000. To put this figure in perspective, Figure 32 compares total output and output per person in the United States to a selection of other countries around the world. The range of variation in production per person is remarkably large. Despite having a population nearly five times as large as the United States, China's total production is only

FIGURE 31



SOURCES: Louis D. Johnston and Samuel H. Williamson, "What Was the U.S. GDP Then?" MeasuringWorth, 2020. URL: http://www.measuringworth.org/usgdp/.
Carter, Susan B., "Labor force, employment, and unemployment: 1890-1990." Table Ba470-477 in *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*, edited by Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright. New York: Cambridge University Press, 2006. http://dx.doi.org/10.1017/ISBN-9780511132971.Ba340-651.
U.S. Department of Labor, Bureau of Labor Statistics, "Current Population Survey," ftp://ftp.bls.gov/pub/special.requests/lf/aat1.txt.
Output valued in prices of 2012.

Real Output per Capita and per Worker, 1900–2019

about two-thirds that of the United States while its per capita output is only about 15 percent as large as in the United States. The countries with the lowest levels of production per person in this list are in South Asia and Africa. While India, for example, is starting to catch up to wealthier countries, its output per capita is only 3.2 percent that of the United States.[10]

Even in the United States and other advanced economies, such as those of Japan and Western Europe, there are still many people living in poverty. But even the poorest citizens of these countries enjoy access to a bounty of material goods that far exceeds the consumption possibilities of the typical resident of countries at the bottom of the list in Figure 32.

Human happiness, of course, depends on more than just the material level of consumption that we are able to achieve. Living a long and healthy life, access to education, and a clean environment are also important. But, the reality is that the material resources created by higher levels of production make possible longer life, broader access to education, better healthcare, and a cleaner environment. These relationships are illustrated in Figure 33, which shows the relationship between output per person and several other indicators of quality of life.

FIGURE 32

	GDP		GDP per capita	
	Billions of $	Index (USA=100)	$	Index (USA=100)
United States	21,427.7	100.0	65,280.7	100.0
Germany	3,845.6	17.9	46,258.9	70.9
United Kingdom	2,827.1	13.2	42,300.3	64.8
France	2,715.5	12.7	40,493.9	62.0
Japan	5,081.8	23.7	40,246.9	61.7
South Korea	1,642.4	7.7	31,762.0	48.7
Russia	1,699.9	7.9	11,585.0	17.7
China	14,342.9	66.9	10,261.7	15.7
Mexico	1,258.3	5.9	9,863.1	15.1
Brazil	1,839.8	8.6	8,717.2	13.4
Egypt	303.2	1.4	3,020.0	4.6
Nigeria	448.1	2.1	2,229.9	3.4
Ghana	67.0	0.3	2,202.1	3.4
India	2,875.1	13.4	2,104.1	3.2
Pakistan	278.2	1.3	1,284.7	2.0

Source: https://data.worldbank.org.

Output and Output per Capita in 2019 in Different Countries

Recessions and Expansions

If you look closely at the line showing total output in Figure 30, you will see that the rate at which the U.S. economy's output has grown is not steady. There are periods of rapid growth and periods of slower growth, or even decline. The decline in real output during the Great Depression is particularly striking as is the sharp increase during the Second World War (1941–45).[11] We can also see in these data the recession that coincided with the 2008 financial crisis.

The variability of the growth of output is more obvious in Figure 34, which plots the percentage change in output between successive years. A period between a trough and a peak in economic activity is called an **expansion**; a period between a peak and a trough in economic activity is called a **recession**. When a recession is particularly severe, it is called a **depression**. The period from 1929 to 1933 is the most severe episode of economic decline observed to date and is called the Great Depression.

The alternation of periods of expansion and recession is referred to as the **business cycle**. These fluctuations are one of the fundamental features of the economy that macroeconomics seeks to explain. Because periods of recession are associated with declining employment opportunities and slower wage growth, a central focus of macroeconomic policy is to find ways to reduce the severity and duration of such periods.

Unemployment

The unemployment rate is the percentage of the labor force that would like to work but cannot find employment. The labor force is made up of all individuals who are employed or unemployed. When the unemployment rate is high, it is hard to find work, and people who do have jobs generally find it harder to earn promotions or increase their pay. Figure 35 shows the unemployment rate since 1900.

In general, the unemployment rate goes up during recessions and falls during expansions. You can see that the unemployment rate was especially high during

FIGURE 33

	GDP PER CAPITA IN U.S. DOLLARS (2019)	LIFE EXPECTANCY AT BIRTH, IN YEARS (2018)	ADULT LITERACY RATE (% IN 2018)	INTERNET USAGE (% IN 2018)
Brazil	8,717.2	76	93	70
China	10,261.7	77	97	54
Egypt	3,020.0	72	71	47
France	40,493.9	83	*	82
Germany	46,258.9	81	*	90
Ghana	2,202.1	64	79	39
India	2,104.1	69	74	34
Japan	40,246.9	84	*	91
Mexico	9,863.1	75	95	66
Nigeria	2,229.9	54	62	42
Pakistan	1,284.7	67	59	16
Russia	11,585.0	73	100	81
South Korea	31,762.0	83	*	96
United Kingdom	42,300.3	81	*	95
United States	65,280.7	79	*	87

SOURCE: https://data.worldbank.org.
*Data not reported.

Output per Capita and Other Development Indicators

the Great Depression. Figure 35 illustrates two other important points about the unemployment rate.

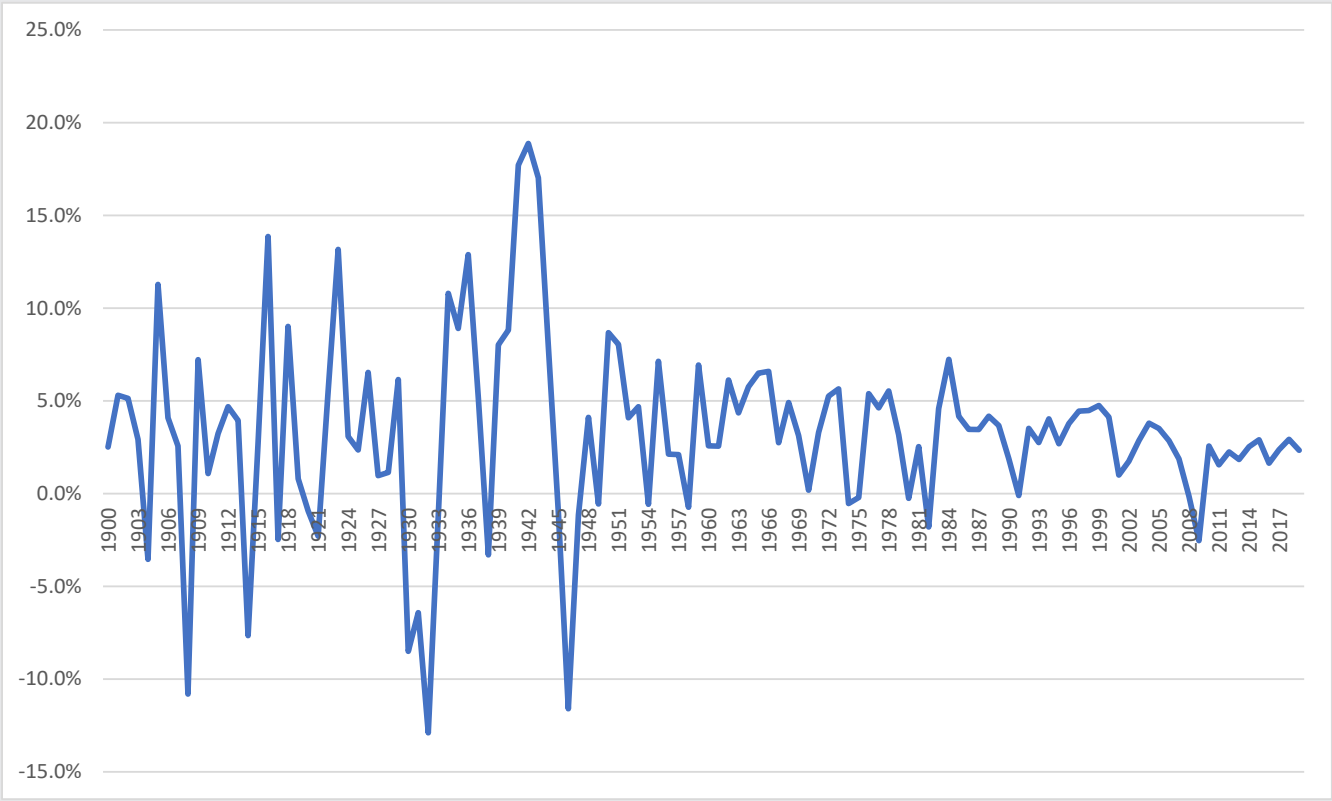
First, the unemployment rate is never zero. There are always some people searching for work. This reflects the continual entry of new job-seekers into the labor market as well as the shifting fortunes of different industries, regions, and businesses within the economy. Even in expansions, some companies are closing, while others are growing. Even during the Great Depression, when many employers were laying off workers, others were expanding their workforce. Second, despite the huge changes that have taken place in the economy since 1900, there is no indication that the unemployment rate is increasing in the long term.

Inflation

We have seen that the prices of individual goods and services play a central role in coordinating individual choices within markets. When the price of a particular good—say a gallon of gasoline—rises, this increase signals consumers to reduce their consumption and creates incentives for suppliers to increase production. When all prices rise together, economists call this inflation. Because inflation means that all the things people consume are becoming more expensive, inflation reduces purchasing power and makes people worse off. We will see that inflation imposes other economic costs as well. So, keeping inflation low is another important goal of macroeconomic policy.

Figure 36 shows the U.S. inflation rate since 1900. As this figure makes clear, the rate of inflation has varied considerably over time. Prices generally increase over time, but there have been some periods during which the price level fell. Most notably, this occurred during the Great Depression, but we also saw the price level fall during the 2008 financial crisis.

FIGURE 34



Annual Percentage Change in Real GDP, 1900–2019

International Trade

National economies are linked to one another through international trade. Because of its size, the United States is relatively less dependent on trade than many other, smaller countries. While international trade has generally increased since the 1950s, the level of exports and imports as a share of GDP has fallen over the past few years in the United States.

Figure 37 plots the volume of exports from the United States to other countries and the volume of imports to the United States since 1929 as a percentage of total output. When exports exceed imports, economists say that a country is running a *trade surplus*. When exports are less than imports, they say that a country is running a *trade deficit*.

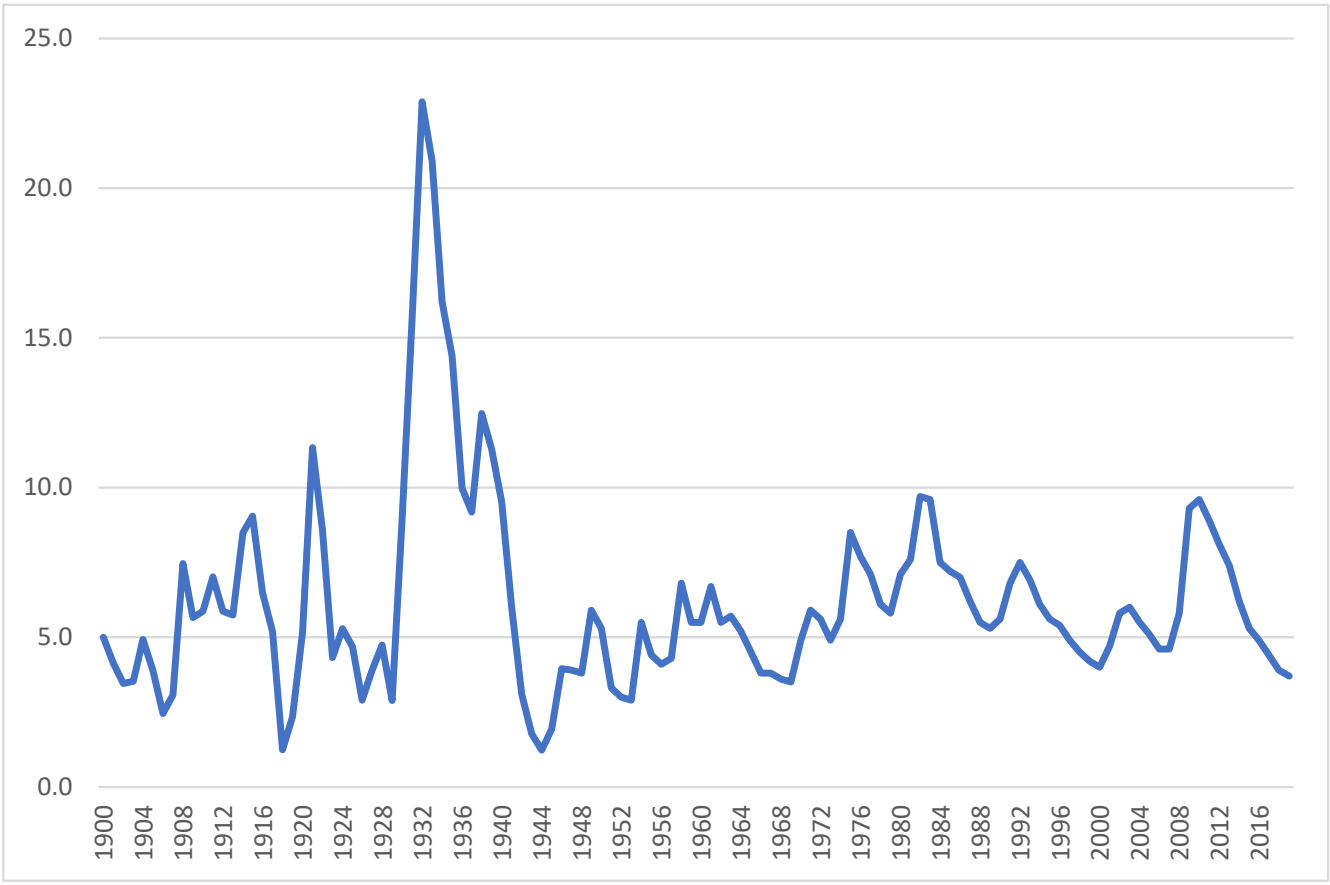
In the long run, the levels of imports and exports appear to move in similar ways. But there have been shifts in their relative levels. Up until the late 1950s, the United States generally exported more than it imported. Since the 1970s, the relationship has shifted, and imports are greater than exports.

MACROECONOMIC MEASUREMENT

In our description of the behavior of the U.S. economy in the previous section, we made use of concepts like the total national output, inflation, and unemployment. Constructing measures that capture the overall behavior of the national economy involves aggregation. Aggregation is the combination of many different things into a single economic variable. Well-constructed economic aggregates help us to see the big picture, but at the cost of obscuring important details.

Developing appropriate economic aggregates is an important branch of macroeconomics, and understanding the choices that go into the construction of these aggregates is important if we are to fully understand what their behavior tells us about the

FIGURE 35



SOURCES: Carter, Susan B., "Labor force, Employment, and Unemployment: 1900–2019." Table Ba470–477 in *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*, edited by Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright. New York: Cambridge University Press, 2006. http://dx.doi.org/10.1017/ISBN-9780511132971.Ba340–651.
United States, Bureau of Labor Statistics, http://www.bls.gov.

Unemployment as a Percentage of the Civilian Labor Force

economy. In this section, we will describe in more detail how the most important macroeconomic variables are defined, and we will discuss the significance of these definitions.

Measuring Total Output: Gross Domestic Product

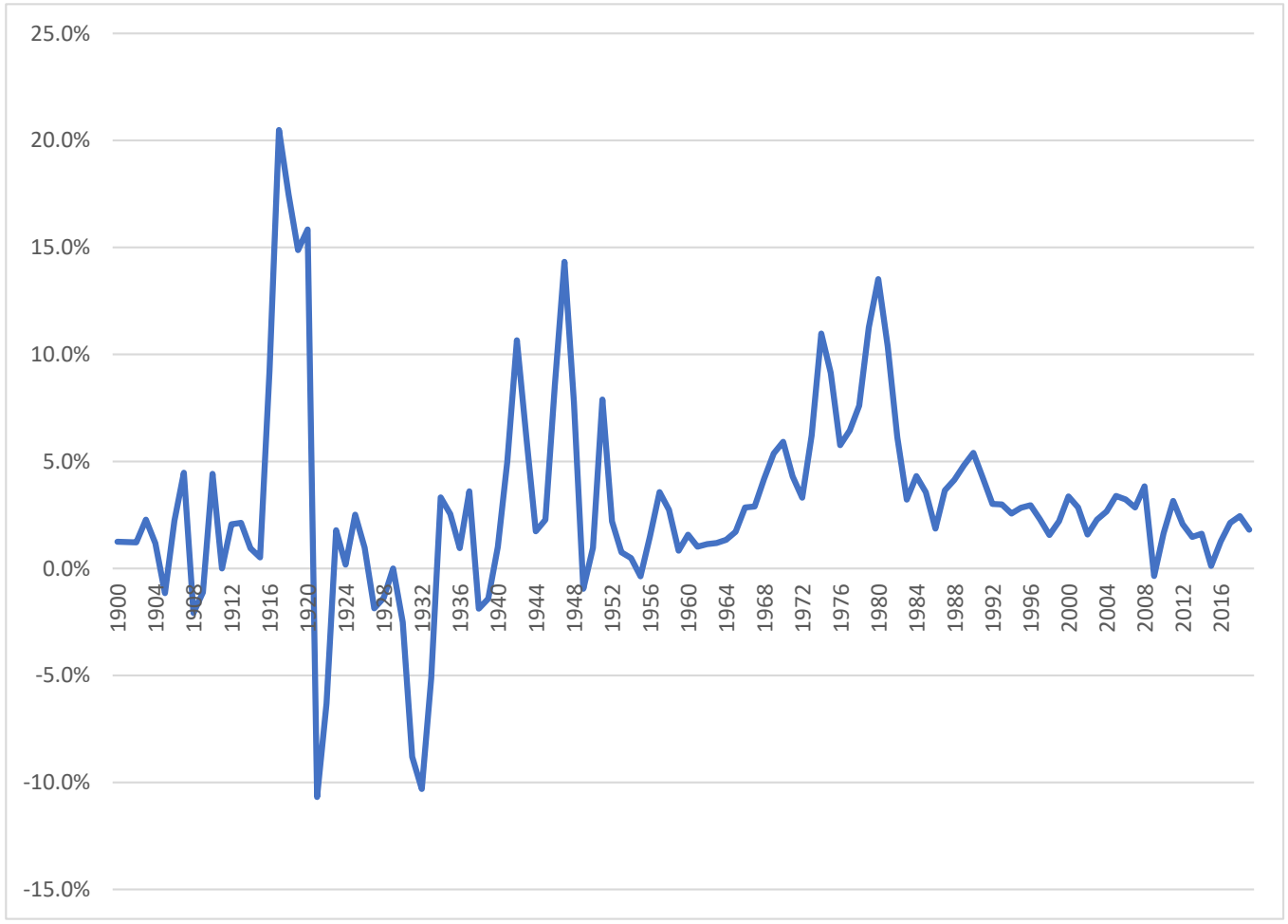
Earlier we presented data showing the growth of the total output of the U.S. economy. But, how can we measure the total output of an economy? How do we add up haircuts, personal computers, fast food hamburgers, financial advice, automobiles, and the myriad other goods and services produced by an economy?

The answer that economists have developed to this question is called Gross Domestic Product (or GDP). Formally, GDP is defined as: "the market value of all **final goods** and services produced within a country during a specified period of time." This definition is short, but there are several important points to note about it.

Market Value

To combine all the different types of things that a country produces, we use their dollar value to add them up. Suppose, for example, that an economy produced only two goods: t-shirts and shorts, and that t-shirts sell for $5 each, while shorts sell for $10.

FIGURE 36



NOTE: calculated as year-to-year change in the GDP Deflator
SOURCE: Louis D. Johnston and Samuel H. Williamson, "What Was the U.S. GDP Then?" MeasuringWorth, 2020.
URL: http://www.measuringworth.org/usgdp/.

Annual Rate of Inflation

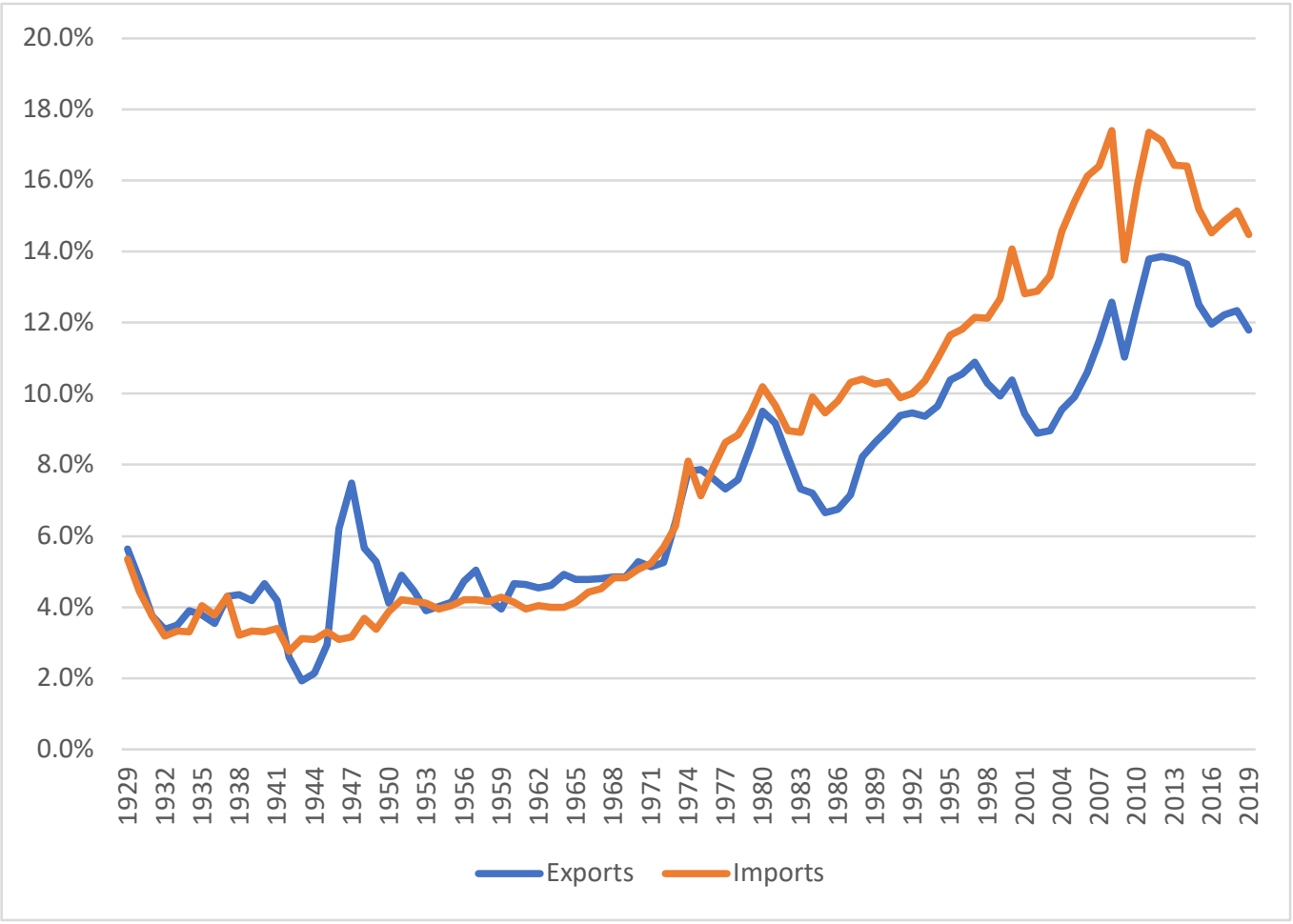
If the economy produced 100 t-shirts and 25 pairs of shorts, then its GDP would be 100 × $5 + 25 × $10 = $750. Because of the use of market prices, higher-priced goods contribute more to total GDP. Recall from our discussion of microeconomics that market prices reflect the value that the marginal consumer places on the good. So, goods that have higher prices have a higher value to consumers and therefore should contribute more to total output.

Final Goods and Services

Most of the products we consume are the result of a complex chain of production activities. For example, automakers purchase steel from refiners, who in turn purchase iron ore from a mining company. Because the automobile is the end product of this chain of purchases, we count only its value in GDP and exclude the purchase of inputs that are used up to produce the car. Goods that are used up in the production of a final good are called **intermediate goods**.

Excluding intermediate goods from GDP insures that our measure of GDP is not affected by the extent of vertical integration in the economy. This is important to avoid the possibility of double counting the value of some goods. To see this, consider the following alternative scenarios. First, suppose a steel producer sells $200,000 worth of steel to an auto manufacturer,

FIGURE 37



SOURCE: Irwin, Douglas A., "Exports and Imports of Goods and Services: 1929–2019." Table Ee376-384 in *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*, edited by Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright. New York: Cambridge University Press, 2006. http://dx.doi.org/10.1017/ISBN-9780511132971.Ee362-611.

Exports and Imports as a Percentage of GDP

and the auto manufacturer converts the steel into $1 million worth of automobiles. The steel is an intermediate good because it is used to produce the automobiles. Now, suppose the automaker produces its own steel and sells $1 million worth of automobiles. Notice that in both cases the value of the steel is included in the value of the automobiles. By excluding the transaction involving the intermediate good, we arrive at the same contribution to GDP regardless of the pattern of industry ownership.

Some goods can be either final goods or intermediate goods. In this case, we only count that portion of production that is sold to final users. As an example, suppose Sylvia raises tomatoes. In one year, she produces $200 worth of tomatoes. She sells $100 worth at a local farmers market and uses the other $100 worth to make tomato sauce, which she sells for $200. Sylvia's contribution to GDP is $300—the result of adding the $100 worth of tomatoes she sells to consumers and the $200 worth of tomato sauce. We do not count directly the $100 worth of tomatoes used to produce the sauce, but it is reflected in the value of the final product that it is used to produce.

Capital goods do not fit easily into either of the categories we have discussed so far. Capital goods are long-lived goods that are themselves produced and are

used to produce other goods and services but are not used up in production. Machinery and factory buildings are examples of capital goods. For the purpose of consistency, economists have adopted the convention that capital goods are included in GDP in the year they are produced. If we did not count them, then a country that invested in its future by building capital equipment would appear to have a lower GDP than one that used all its resources to produce consumer goods.

Within a Country

The word "domestic" in Gross Domestic Product indicates that we count only goods produced within the borders of the country that we are discussing. So, U.S. GDP includes the value of all automobiles produced in the United States, whether made by an American auto manufacturer or by a foreign-owned one.

During a Specified Period

Production takes a certain amount of time, but we only include items that are produced between the beginning and end of the period in question. Conventionally economists consider either annual or quarterly (three-month periods) GDP. An important implication is that the sale of goods produced in earlier periods is not included in GDP. For example if a twenty-year-old house is sold this year for $150,000, then this amount is not included in GDP. The house was not produced this year. It was included in GDP when it was produced, so we don't count it again when it changes hands. On the other hand, if the real estate agent who arranged the sale received a 6 percent commission, this $9,000 fee is included in GDP since the real estate services the agent provided were produced in the current year.

Understanding What GDP Measures

The conceptual basis for the measurement of GDP was developed in the 1930s. Interest in measuring economic output is longstanding, however. One of the earliest known efforts to measure national output was undertaken by Sir William Petty in the mid-1600s as part of the British government's effort to assess the ability of the Irish people to pay taxes to the crown.

Because the lack of comprehensive data on national economic activity was hampering efforts to respond to the Great Depression, in 1932 the U.S. Department of Commerce commissioned the economist Simon Kuznets to develop a system to measure national output. Kuznets presented his system in a report to the U.S. Senate in 1934. The U.S. entry into the Second World War provided an additional impetus for perfecting techniques of measuring output and establishing the necessary data collection tools to produce ongoing estimates of GDP. In 1971, Kuznets received the Nobel Prize in Economic Science in part for his contributions to the measurement of national production.[12]

The continued use of the concepts developed by Kuznets, and their subsequent refinement by other scholars, reflects the practical value of these concepts. But, it is important to recognize that despite the usefulness of these ideas, they have a number of important limitations. Three of these are described below.[13]

First, as we have already noted, it is not always easy to determine what constitutes final goods and services. One illustration of this point is the treatment of expenditures on national defense. Conventionally these are included in GDP, but Kuznets pointed out that they might equally well be viewed as an intermediate good that enables the citizens of a country to enjoy other final goods and services.

A second limitation of GDP arises from its exclusion of goods that are not bought and sold in markets. One very important example is unpaid household work. Housekeeping and childcare performed by family members are not counted in GDP, but if these services are purchased in the market, then they are. Over the past sixty years, as women have increasingly entered the paid **labor force**, the amount of commercially provided childcare and housecleaning has increased, causing GDP to rise. But, because some of this increase is simply a shift from non-market to market activity, it does not in fact reflect an increase in total production.

A third limitation of conventional GDP measurement is that it ignores activities that deplete a country's stock of natural resources or pollute the environment. Although economic theory provides some guidance about how natural resources and environmental quality should be valued, actually measuring their value has proved more difficult.

Other Ways to Measure GDP: Expenditures Equal Production

GDP is a measure of the quantity of goods and services produced in a country. But, since goods that are produced are also purchased, we can also think of GDP as a measure of the total value of expenditures within

FIGURE 38

	T-SHIRTS		SHORTS		GDP	
	QUANTITY	PRICE	QUANTITY	PRICE	NOMINAL	REAL
2015	100	$5.00	25	$10.00	$750.00	$750.00
2020	200	$7.50	50	$15.00	$2,250.00	$1,500.00
2020 relative to 2015	2	1.5	2	1.5	3	2

	T-SHIRTS		SHORTS		GDP	
	QUANTITY	PRICE	QUANTITY	PRICE	NOMINAL	REAL
2015	100	$5.00	25	$10.00	$750.00	$750.00
2020	200	$7.50	75	$15.00	$2,625.00	$1,750.00
2020 relative to 2015	2	1.5	3	1.5	3.5	2.3

Calculation of Real GDP

a country. Economists divide purchasers into four categories: households, firms, government, and the foreign sector (that is foreign purchasers of domestic products). Each of these categories corresponds to a category of spending.

Household purchases are called *consumption expenditures*, or consumption for short. These purchases are subdivided between consumer durables, nondurables, and services. *Consumer durables* are long-lived consumer goods such as automobiles, washing machines, and furniture. Note that expenditures on new houses are included in investment rather than in consumer durables. *Consumer nondurables* are goods that are used up more quickly than durable goods, such as food or clothing. *Services* are intangible goods such as education, legal services, insurance, and financial services.

Spending by firms on final goods and services, along with household purchases of new houses, comprise **investment**. Investment is subdivided into three categories. Business purchases of factories, offices, machinery, and equipment is called *business fixed investment*. The purchase of new homes and apartment buildings is called *residential fixed investment*. The final category of investment spending is *inventories*, which consists of additions of unsold goods to company inventories.

Notice that economists' use of the word "investment" is somewhat different from the word's use in ordinary conversation. In ordinary conversation, we often describe the purchase of financial assets, such as shares of stock or bonds, as making an investment. Such purchases transfer ownership of an existing financial or physical asset, but do not create new assets. In economics, the term "investment" is reserved for the purchase of new capital goods, such as buildings or equipment.

Government purchases include all of the goods and services purchased by federal, state, and local governments. These include wages paid to firefighters and teachers and purchases of fighter planes for the military. In addition to purchasing goods and services, governments make transfer payments, such as paying Social Security benefits. These transfer payments are not counted in government purchases of goods and services and neither is interest paid on government debt.

Net exports is the difference between the value of domestically produced goods sold to foreigners (exports) and the value of foreign-produced goods purchased by domestic buyers.

The relationship between GDP and the various categories of spending can be summarized by the equation **GDP = C + I + G + NX**, where C is consumption, I is investment, G is government

spending, and NX is net exports.

Yet Another Way to Measure GDP: Income Equals Production Equals Expenditures

We have seen that GDP can be measured either in terms of production or spending. In addition, GDP can be thought of as income. Whenever a good or service is sold, the revenue is distributed between the workers and the owners of the capital used to produce it. Except for some minor technical adjustments, the combined income of labor and capital equals expenditures, which equals production. As a result, we can state the following important identity: **GDP = Production = Expenditures = Income**.

For this reason, economists use these three different designations interchangeably when discussing the nation's GDP.

Real GDP

Recall that GDP is calculated by adding up the market value of all the goods and services produced (purchased) in a country during a specified period. As a result, the size of the resulting sum depends on both the quantity of goods and services produced and their respective prices. Because economists are often interested in comparing the level of economic activity over time or between different locations, it is important to have a way to separate the effects of changes in prices from changes in the quantity of goods and services produced.

The problem posed by changing prices is illustrated in the example shown in the top panel of Figure 38. This table reports prices and quantities for an economy producing just two goods in two years. Between 2015 and 2020, GDP tripled, rising from $750 to $2,250. But, if you look more closely at the quantity data, you can see that output of both t-shirts and shorts has doubled. Because prices increased by 50 percent, however, GDP tripled while the physical volume of production doubled.

In this case, it is simple to isolate the effects of changes in the physical quantity of production from the effects of changes in prices, but in most situations the quantities produced of some goods are increasing, while others are decreasing. Prices, too, will not change in a consistent way. To isolate the effects of changes in production from changes in prices, economists construct **real GDP** by using prices from a single year to value production in each year. This year is called the base year. For the example shown in Figure 38, if we use the prices in 2015 as the base year, then real GDP in 2020 would be calculated by taking the 2020 levels of production and multiplying by the 2015 prices for each good. For example, real GDP in 2020 = (200 x $5) + (50 x $10) = $1,500, twice the real GDP in 2015 and consistent with the doubling of production of each good.

The bottom panel of Figure 38 illustrates the calculation of real GDP in a more complicated situation where production does not grow at the same rate for the different goods. In this case, the quantity of t-shirts doubles, while the quantity of shorts produced triples. Using 2015 prices as the base year, GDP in 2020 is now $1,750 = 200 × $5 + 75 × $10. To clearly distinguish the current year GDP from real GDP, economists commonly call GDP calculated with current year prices **nominal GDP**. As Figure 38 shows, nominal GDP in 2020 is $2,625. The increase in real GDP is $1,750 / $750 = 2.33, which is somewhere between the quantitative increase of the two products of the economy.

Measuring Inflation

To measure inflation, the U.S. Bureau of Labor Statistics calculates the **Consumer Price Index** or CPI each month. The CPI measures the cost of purchasing a market basket of goods and services intended to be representative of the consumption of a typical consumer. To identify the components of the market basket, the Bureau of Labor Statistics (BLS) conducts periodic surveys of consumer expenditures in which a sample of households collects careful records of all of their expenditures. These responses are then aggregated to create a picture of the types and amount of goods and services purchased each month by representative households. Different market baskets are calculated for consumers at different income levels and for those living in different parts of the country to reflect differences in consumption patterns.

Each month BLS employees visit stores, check websites, and otherwise collect actual price information (including any temporary discounts offered by retailers) for all of the items in the market basket of goods determined by the Consumer Expenditure Survey. The BLS then combines these price data with the quantities in the market basket to calculate the cost of purchasing this bundle of goods and services. Finally, this cost is expressed as an index number relative to the cost of the

FIGURE 39

Household Consumption Bundle

	QUANTITY
Pants	2 pairs
T-Shirts	3
Shoes	1 pair

CPI Calculation

	PANTS		T-SHIRTS		SHOES		CONSUMPTION BUNDLE	
	PRICE	COST	PRICE	COST	PRICE	COST	COST	INDEX (2014=100)
2014	10	20	5	15	25	25	60	100.0
2015	10	20	7	21	30	30	71	118.3
2016	11	22	7	21	35	35	78	130.0
2017	12	24	8	24	50	50	98	163.3
2018	14	28	10	30	50	50	108	180.0
2019	13	26	10	30	40	40	96	160.0
2020	14	28	11	33	45	45	106	176.7

Calculation of the Consumer Price Index

bundle in the base year.

Figure 39 illustrates this calculation for an economy in which the consumption bundle consists of three items: pants, t-shirts, and shoes. We see that the quantity consumed each month is two pairs of pants, three t-shirts, and one pair of shoes. Using 2015 as the base year, we set the cost of the bundle in this year equal to 100, and calculate the CPI in the other years using the following formula: **CPI in year *t* = 100 × (cost of bundle in year *t*)/(cost of bundle in base year)**.

Notice that the quantities of each item in the bundle determine the impact of that item's price changes on the overall index. Because consumers purchase three t-shirts and only one pair of shoes, a change in the price of t-shirts will cause a larger change in the CPI than will an equivalent dollar increase in the price of shoes.

The CPI is of considerable practical importance in our economy. Each year, Social Security benefit payments are adjusted to reflect changes in the cost of living as reflected in the CPI. Similarly many union employment contracts include cost-of-living adjustment provisions that tie wage increases to the CPI. More informally, employers and employees take into account changes in the CPI when considering adjustments in wage rates.

The goal of the CPI is to measure how changes in prices affect the ability of households to maintain the level of well-being they enjoyed in the base year. What the CPI actually measures, however, is how changes in prices affect the cost of a fixed bundle of goods and services. This difference means that the CPI will typically overstate the true increase in the cost of living. This upward bias in the CPI arises for three reasons.

The first factor causing the CPI to overstate the effect of rising prices on the cost of living is *substitution bias*. As relative prices change, households will shift their consumption away from more expensive goods and services and toward less expensive ones. When the price of beef increases, for example, families will

consume more chicken; when airline ticket prices decline, consumers will choose to fly more and drive less. By adjusting their consumption toward less expensive goods, households can achieve the same level of well-being at a cost that is lower than the cost of buying a fixed basket of goods and services.

The second source of upward bias in the CPI is *unmeasured quality change*. Many goods and services get better over time due to technological change. In the past several decades, for example, personal computers have steadily become more powerful because of increased processor speeds, greater storage, and better software. Similarly, the addition of anti-lock brakes, airbags, satellite radio, and GPS systems has substantially improved the quality of the typical automobile. Such quality improvements would be expected to raise the price of these goods, so a simple comparison of prices between one year and the next will overstate the price increase or understate any decline in prices. Although BLS statisticians try to account for these quality changes, they are very difficult to remove completely from the CPI.

The third reason the CPI overstates the true rate of inflation is because of the introduction of *new goods and services*. A striking example of this is the cell phone. The first cell phones were introduced in the mid-1970s. Prior to this, mobile communication was simply unavailable at any price for most consumers. Because cell phones did not exist, they were not included in the market basket used by the BLS to calculate the CPI. During the early years of their development, prices for cell phones fell rapidly, and the quality of service vastly improved. But, because cell phones were not included in the CPI, none of these effects were reflected in measures of inflation. Only after cell phones had achieved a relatively large market penetration were they added to the CPI basket.

In 1996 the Boskin Commission, headed by economist Michael Boskin, carefully reviewed the methods used to calculate the CPI and concluded that the combined effects of substitution bias, quality improvement, and the introduction of new goods meant that the CPI overstated the rate of price inflation by 1.3 percent per year.[14]

The CPI is just one way that economists measure changes in the cost of living. The relationship between real and nominal GDP provides a slightly different perspective on inflation. This measure is called the *GDP deflator*, and it is defined by the following equation: **Nominal GDP = (GDP Deflator/100) × (Real GDP)**.

That is, we define the GDP deflator to be an index number, such that when we multiply real GDP by that index number we get the nominal GDP. Dividing both sides of the equation by Real GDP and multiplying both sides by 100, we can state this relationship as: **GDP Deflator = 100 × (Nominal GDP)/(Real GDP)**.

Figure 40 compares the rate of inflation as measured by the CPI and the GDP deflator since the early 1960s. As this comparison illustrates, they tell similar stories about the cost of living, but the GDP deflator is somewhat less volatile, rising less at peaks and decelerating less at low points. Over the entire period, the GDP deflator has risen somewhat less than the CPI.

There are several reasons for these differences. The first difference is that the GDP deflator reflects only the prices of domestically produced goods. To the extent that foreign produced goods have a larger role in the CPI market basket, differences in their behavior will show up in differences in the two indexes. One reason the CPI rose so much more than the GDP deflator at the beginning and end of the 1970s is that rising oil prices had a large effect on the CPI, but because this was mainly produced overseas, it did not affect the GDP deflator.

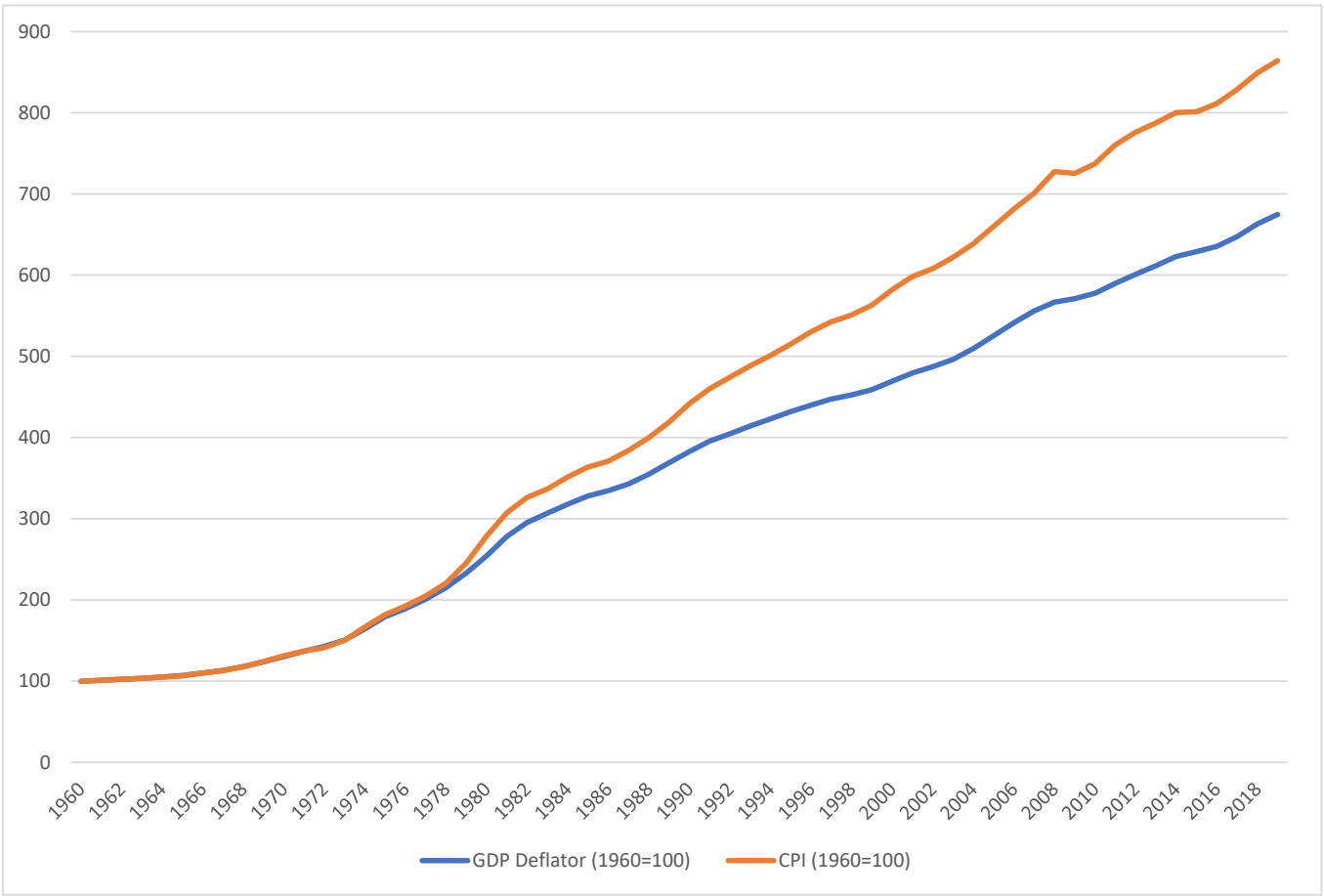
The second reason the GDP deflator and the CPI diverge has to do with the way in which they weight the prices of different goods and services. Whereas the CPI uses a fixed market basket to weight the prices of different goods, the GDP deflator weights prices by their current levels of production. As a result, the basket of goods used to weight prices in the GDP deflator adjusts to changing consumption patterns over time.

Unemployment

Macroeconomists use a variety of indicators to gauge the state of the economy. The unemployment rate is an especially sensitive indicator of how well the economy is performing at any moment. When the unemployment rate is low, workers feel secure in their jobs, and competition between employers helps to drive up wages. When unemployment is high, however, workers worry about losing their jobs.

The *unemployment rate* is defined as the percentage of the labor force that is unable to find a job. The labor force, in turn, consists of all working-age adults who

FIGURE 40



SOURCE: Federal Reserve Economic Data, Economic Research Division Federal Reserve Bank of St. Louis, https://fred.stlouisfed.org.

Comparison of CPI and GDP Deflator, 1960–2019 (1960=100).

are either employed or are actively seeking work. In the United States, the Bureau of Labor Statistics (BLS) is responsible for measuring the unemployment rate. To do this each month, the BLS surveys approximately 60,000 households. Based on a series of questions, interviewers classify every person age sixteen or older in the household into one of three categories:

- **Employed.** If that person worked for pay either full- or part-time during the previous week or is on vacation or sick leave from a regular job.
- **Unemployed.** If that person did not work during the previous week but made some effort to find paid employment during the past four weeks.
- **Out of the labor force.** If that person did not work during the past week and did not actively seek work during the previous four weeks.

Together these three categories comprise the working-age population. The sum of the employed and unemployed constitutes the labor force, and the unemployment rate is the quantity of people unemployed expressed as a percentage of the labor force.

Figure 41 shows data on the U.S. labor force collected by the BLS in July 2020. The table shows that there are approximately 260 million working-age persons in the United States. Of these, almost 160 million are in the labor force. The ratio of those in the labor force to the working-age population is called the **labor force participation rate**, which is about 61.4 percent. Of those in the labor force, about 143.5 million had jobs

FIGURE 41

Employment Situation Summary Table A. Household data, seasonally adjusted, July 2020

Civilian Population and Labor Force (in 1000s)	
Civilian noninstitutional population	260,373
Civilian labor force	159,870
Employed	143,532
Unemployed	16,338
Not in labor force	100,503
Unemployment Rates (percentage)	
Total, 16 years and over	10.2
Adult men (20 years and over)	9.4
Adult women (20 years and over)	10.5
Teenagers (16 to 19 years)	19.3
White	9.2
Black or African American	14.6
Asian	12.0
Hispanic or Latino ethnicity	12.9

SOURCE: United States, Department of Labor, Bureau of Labor Statistics, "Employment Situation Summary," July 2020. http://www.bls.gov/news.release/empsit.toc.htm.

Employment and Unemployment in the U.S., July 2020

while 16.3 million were unemployed, resulting in an unemployment rate of 10.2 percent. Just one year earlier, the unemployment rate was only 3.4 percent, so this significant increase in the fraction of the labor force that is unemployed reflects the significant economic damage caused by the COVID-19 pandemic. Note that the unemployment rate is highest among the teenage population, and it varies by race, ethnicity, and gender.

There are many reasons why some people are unemployed. Economists divide these reasons into three broad categories.

Frictional Unemployment

The U.S. economy is remarkably dynamic. Every month several million workers leave their jobs either voluntarily (i.e., they quit) or involuntarily (i.e., they get laid-off), and several million more are hired. Because job-searching takes time, many of these workers show up as unemployed for brief periods of time. An additional source of **frictional unemployment** comes from new workers entering the labor force for the first time. Frictional unemployment refers to the portion of the unemployed who are currently not working because of the normal process of matching employees and employers.

Structural Unemployment

Sometimes the jobs that are available require different skills or characteristics from those possessed by the workers who are seeking employment. The locations of job-seekers and vacancies may also be different, preventing those seeking employment from filling the available positions. That portion of total unemployment attributable to the mismatch between job openings and job-seekers is called **structural unemployment**. In the 1980s, for example, the U.S. steel industry was contracting while the computer industry was

expanding. Not only were laid-off steel workers located in the industrial northeast far from expanding Sunbelt industries, but many of them also lacked the skills to pursue such jobs. If self-driving vehicles begin to replace Uber drivers in the coming years, then those drivers would be structurally unemployed as a result of long-term changes in the structure of the economy (in this case due to automation).

Cyclical Unemployment

During recessions, unemployment rises as lay-offs increase, and new hires decline. In these circumstances, job-seekers find it harder to find employment, and many of them spend longer searching for work. The additional unemployment that occurs for this reason is called **cyclical unemployment**.

ECONOMIC GROWTH, PRODUCTIVITY, AND LIVING STANDARDS

Would you prefer to have an average income in the United States today or to have been the richest person living in 1900? Earlier we saw (Figure 31) that real GDP per capita grew by a factor of 9 between 1900 and 2019. In other words, the value of goods and services available to the average person today is nine times as large as what the average citizen could consume in 1900. But this comparison hardly captures the change that has taken place in our economy and consumption patterns over the past century.

In 1900, even the wealthiest American citizen could not go to the movies, could not travel from the United States to Europe in a single day, watch television, use a computer, or get antibiotics to treat an infection. How much income would it take to compensate you to live without computers, the internet, modern medicine, and all the conveniences we take for granted today that were not available a hundred years ago? Many people would conclude that no level of financial incentives would induce them to give up all of these modern conveniences.

The improvement in living standards that has taken place in the United States in the last century is a manifestation of a broader phenomenon that economists call *economic growth*. The phenomenon of sustained economic growth began a little more than two hundred years ago in the United States and Western Europe. During the nineteenth and twentieth centuries, it spread to Japan and parts of Latin America, and since the 1950s to a growing number of countries around the world. Yet, when we look around the world (Figure 33) there is still a strikingly large variation in material well-being and living standards.

In this section of the resource guide, we will look at what economists know about the factors that account for differences in the standard of living over time and between countries. That is, we will develop a theory that explains the size of a nation's economy in the long run.

The Circular Flow Model of the Economy

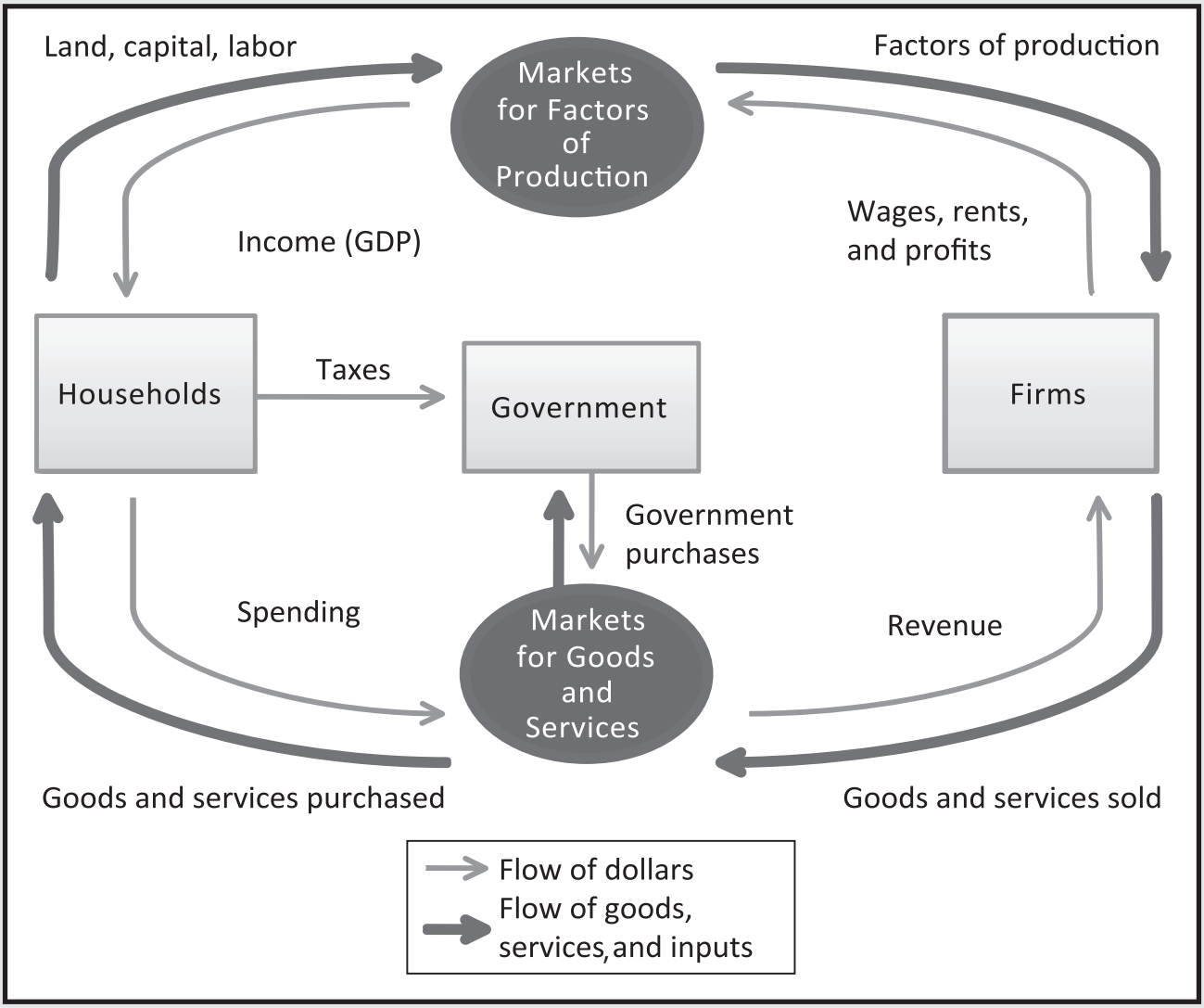
A useful starting point for our discussion is a conceptual model of the economy called the *circular flow* model, which is depicted in Figure 42. By tracing the flow of dollars through the economy, this diagram illustrates schematically the complex set of interactions between the major sets of economic actors in our economy: households, firms, and the government. In this diagram, the major actors are depicted by rectangles, while the markets through which they interact are depicted as ovals. Flows of money and flows of real things—goods and services, inputs to production—are drawn as arrows.

To understand the model, begin on the left-hand side with households. Households receive income by providing factors of production (labor, capital, land) to firms. This transaction is reflected in the arrow leading from the households' box to the factor market, and the parallel arrow labeled income in the other direction.

Even though firms purchase many of the capital goods in our economy, these capital goods are owned indirectly by households through their ownership of the firms, and it is appropriate to depict households as providing this capital to the firms in exchange for rental payments. Households use their income to purchase goods and services, to pay taxes, and to save through **financial markets**. These three uses of their income are illustrated by the three arrows leading out of the box labeled households.

Firms receive revenue from the sale of goods and services (the arrow leading from the markets for goods and services) and use this income to pay for the factors of production that they must hire to produce the goods

FIGURE 42



The Circular Flow Model of the Economy

and services that they sell.

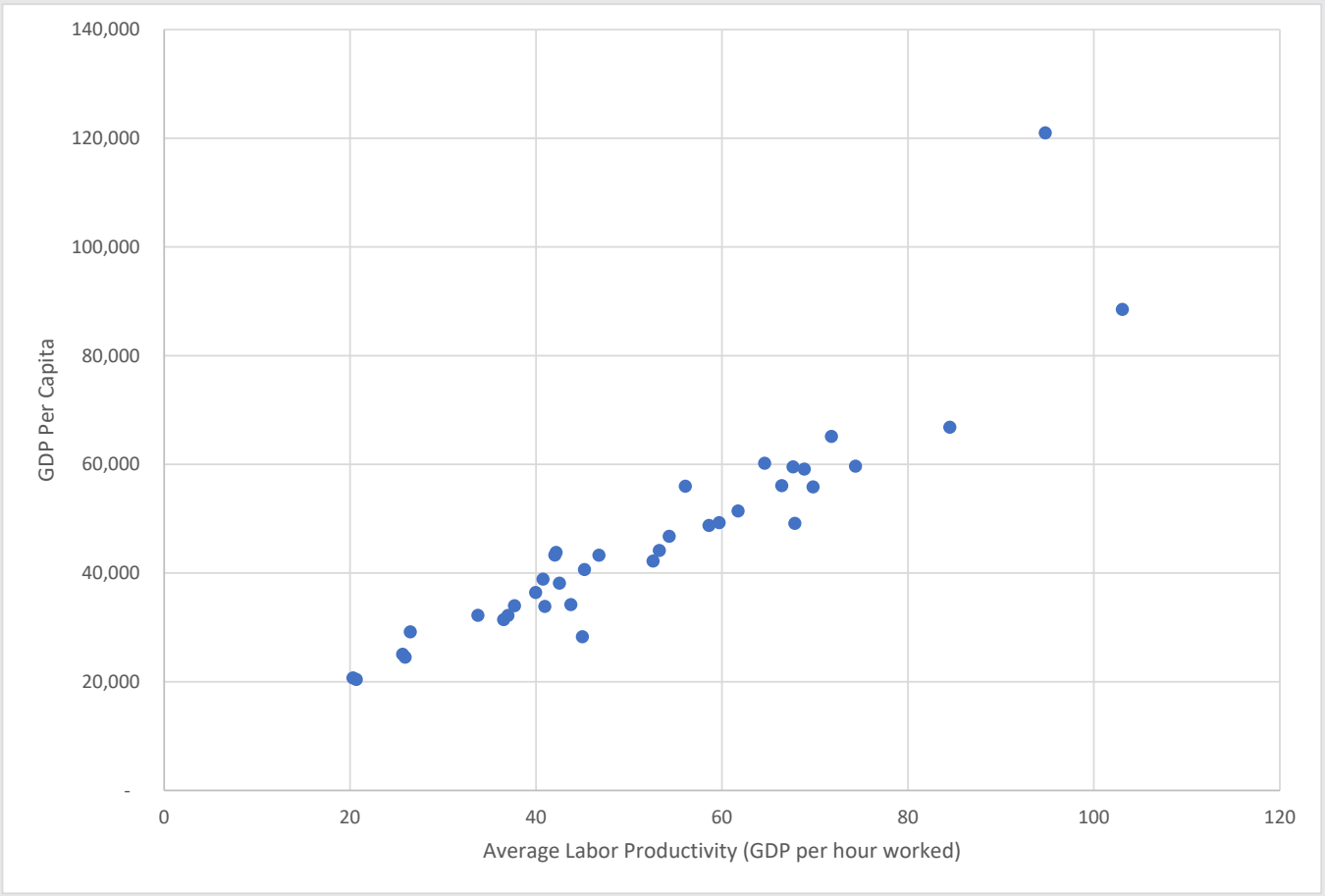
The government receives income from households in the form of taxes, and the government borrows from financial markets. It uses these sources of income to purchase goods and services.

The final flow of funds illustrated in this diagram is from financial markets to the market for goods and services. This flow represents borrowing by both households and firms, which is used to purchase consumer durable goods and capital equipment.

What Determines How Much an Economy Produces?

As the circular flow model emphasizes, an economy's output depends on the total quantity of goods and services that firms are able to produce. This in turn depends on the quantity of factor inputs that households are able to supply to the firms and the ability of the firms to transform these inputs into the outputs that households and the government choose to purchase. Larger economies will produce more (other things being equal) than smaller economies. But, this

FIGURE 43



Relationship Between GDP per Capita and Average Labor Productivity, 2019

source of variation cannot account for differences in **GDP per capita**.

To explain differences in GDP per capita, it is helpful to note that real GDP per capita is equal to real GDP per worker multiplied by the fraction of the population employed. Let POP stand for the country's population, and N stand for the labor force. Then, we can express this relationship in the following equation:

$$\frac{\text{GDP}}{\text{POP}} = \frac{\text{GDP}}{\text{N}} \times \frac{\text{N}}{\text{POP}}$$

The left-hand side of this equation is just real GDP per capita. By cancelling out N in the two fractions on the right-hand side, you can see that the right-hand side reduces to GDP per capita as well, so this relationship is always true. What this expression tells us is that the average quantity of goods and services available for each person to consume depends on the average amount that each worker can produce, or average labor productivity, and the proportion of the population that is engaged in production.

Most of the variation in GDP per capita occurs because of differences in average labor productivity. In the United States, labor force participation rates have increased modestly in the last century as more women have entered the labor force and as lower birth rates have reduced the share of children in the population and consequently increased the relative size of the working-age population. These trends have, however, been offset by earlier retirement and longer education. As a result, virtually all of the increase in output per person in the economy is explained by increased average labor productivity. Figure 43 shows that there is also a strong positive association between labor

productivity and real GDP per capita across countries.

Average labor productivity depends on a number of different factors. The most prominent of these are the following:

- **Physical capital.** Workers equipped with more and better tools, machinery, and up-to-date factories will be more productive. Modern manufacturing methods rely on the use of large quantities of capital per worker to achieve high levels of production. Recall that capital equipment is a produced factor of production; so it is an input that in the past was an output of the production process. As such, increasing the capital stock in the future requires giving up consumption in the present.
- **Human capital. Human capital** is the term that economists use to refer to the skills and experience that are acquired through education, training, and on-the-job experience. Unlike physical capital, human capital is not tangible, but like physical capital, creating it usually requires sacrificing current consumption. Students and trainees must reduce the amount of time they engage in productive activities while they are learning.
- **Natural resources.** Some countries or regions have natural resources like iron ore, petroleum, or natural gas reserves that contribute to the **wealth** of their citizens. The high standard of living of countries like Saudi Arabia and Kuwait are in large part due to the fact that they are located on top of large pools of oil. On the other hand, in an increasingly global world, natural resources are not essential to a high standard of living. Countries like Japan have been able to achieve high levels of per capita income by importing raw materials produced elsewhere.
- **Technological knowledge.** Economists refer to the knowledge about techniques by which inputs are transformed into the goods and services households desire as technological knowledge or simply technology. Advances in this know-how are the single most important factor in raising average labor productivity historically. These advances include the invention of entirely new products, like semiconductors, integrated circuits, lasers, and genetic engineering, as well as the development of better methods of organization, such as Henry Ford's introduction of the moving assembly line.
- **The political and legal environment.** Some kinds of technological knowledge are protected by patents, and others may be kept as trade secrets. But, most of the know-how behind the high levels of productivity in advanced countries like the United States is available to be learned and copied. The very rapid growth of living standards in Japan, South Korea, and China illustrates that countries can catch up quickly if they successfully borrow and adapt these techniques. Yet, the persistent poverty of other countries implies that there are obstacles to successful borrowing. The most persuasive explanation for this is that dysfunctional political and legal systems prevent many countries from fully exploiting the potential of modern manufacturing techniques. A stark illustration of this point is the divergent fortunes of North and South Korea. After World War II, both countries had similar resources, populations, and standards of living. Today, the South enjoys a standard of living comparable with the most developed countries while poverty is widespread in the North. This variance is almost entirely due to differences in governmental institutions.

The importance of the political and legal environment illustrates that creating the appropriate incentives is an essential prerequisite for achieving a high standard of living. But, what actions should policymakers seek to encourage?

Investment in both physical and human capital should be encouraged, but only up to a point. Recall that capital is created as part of the production process, so creating more capital to use in the future requires giving up current consumption. In the extreme, if all of our current output were directed to investment, there would be no goods and services available to consume, and we would all starve. Long before this, however, diminishing returns would make it undesirable to keep investing.

Similarly, investment in the creation of new technological knowledge through research and development (R&D) is desirable. Because new knowledge is a true public good—since the utility of a discovery is not diminished by other people knowing

it—private incentives to create new knowledge may lead to underinvestment. As a result, there is an important role for government to play in encouraging R&D either through tax credits, subsidies, direct expenditures, or legal protections such as the patent system that give inventors a temporary monopoly on the exploitation of their inventions in exchange for the disclosure of their discovery.

SAVINGS, INVESTMENT, AND THE FINANCIAL SYSTEM

As the preceding discussion makes clear, the quantity of resources that an economy directs toward the formation of capital—both physical and human—and toward the creation of new technological knowledge plays a central role in determining the rate of growth of productivity, and hence the standard of living. In essence, we face—both as individuals and collectively—trade-offs between how much we consume today and how much will be available to consume tomorrow. Devoting more resources to capital formation or to research and development means that there are fewer goods and services available to consume today. But, there will be more in the future.

Recall that economists use the terms "saving" and "investment" somewhat differently from how they are used in common conversation. To economists, saving is what happens when someone has more income than they wish to spend. Someone in this situation might put the money they don't want to spend now in a bank, or they might use it to buy shares of stock in a company. They might think of this as investing their money, but to an economist, the term "investment" is reserved to describe the purchase of new capital equipment. So, it is only when the bank lends the money to a business to construct a new factory, or the when the company uses the funds it receives from the sale of stock, that investment takes place.

A variety of different financial institutions help to coordinate the saving and investment decisions within our economy. It will be helpful to begin our discussion by examining several of these institutions in more detail.

Financial Markets

Financial markets are institutions through which individuals who have money they wish to save can supply these funds directly to persons or companies that wish to borrow money for investment.

The Bond Market

When a large corporation like Wal-Mart wants to borrow money to finance the construction of a new store, it can borrow directly from the public. It does this by selling bonds. A bond is a certificate of indebtedness that specifies the obligations of the borrower to the holder of the bond. In other words, it is a sort of IOU. The typical bond specifies when the loan will be repaid—called the date of maturity—and the rate of interest to be paid periodically until the loan is repaid.

The purchaser of the bond gives the company their money in exchange for the promise of repayment of the original amount, called the *principal*, and the periodic interest payments. The purchaser can hold the bond until maturity, or they can sell the bond to someone else. As market interest rates change, the price at which the bond can be sold will change to equate the promised payments of the bond with the new interest rate. This potential variation in the value of a bond is a risk that the buyer assumes. The longer the maturity of the bond is, the greater the risk of such changes in price, and the higher the interest rate that borrowers must pay to induce people to lend them money.

The buyer of a bond also assumes the risk that the borrower may fail to pay some or all of the principal or interest on the bond. The probability that the borrower will *default* on their obligation by declaring bankruptcy depends on the financial conditions of the borrower. The greater this risk is, the higher the rate of interest a borrower must pay to compensate lenders for this risk. Because the U.S. government is considered a safe credit risk, it can generally borrow at lower rates than private companies. By contrast, financially shaky corporations must pay high interest rates.

The Stock Market

Wal-Mart and other companies can also raise funds by issuing shares of stock and selling them to savers. Each share of stock represents ownership of a portion of a firm. If a company issues 10,000,000 shares of stock, then each share represents ownership of 1/10,000,000 of the business. The sale of shares of stock is called *equity finance*, whereas the sale of bonds is called *debt finance*.

Most companies use both equity and debt finance because these two methods of borrowing funds have very different characteristics. The purchaser of a share of Wal-Mart becomes a part owner of the company. If Wal-Mart is profitable, then the shareholders enjoy the

benefits of these profits either through the payment of *dividends* or through an increase in the value of their shares. The bondholders only receive their interest payments. If, however, Wal-Mart runs into financial difficulties, the bondholders are paid before stockholders receive any dividend payments. Purchasers of stock face greater risks than purchasers of bonds, but they also have a greater potential for high returns.

Someone who buys shares of stock in a corporation can sell those shares on an organized stock exchange, such as the New York Stock Exchange (NYSE) or NASDAQ (National Association of Securities Dealers Automated Quotation System). The price at which they can sell shares depends on the supply of and demand for shares in the company. These, in turn, respond to the current profits and future prospects of the company.

It is important to recognize that when shares of stock are traded on a stock exchange, the company does not receive any revenue from these transactions. Consequently, these transactions do not contribute to investment. Only new issues of stock contribute to a nation's investment. The ability of shareholders to easily buy and sell shares of stock on organized exchanges does, however, contribute to their willingness to hold these assets by making it easier for them to access the wealth that they represent.

Financial Intermediaries

An **intermediary** is a third party who acts as a link between two others. In developed economies, there are a great variety of intermediaries who help to link savers and borrowers. Two of the most important intermediaries are banks and mutual funds.

Banks

Many small businesses, such as local construction companies or retail stores, are too small to issue bonds. When these businesses need to borrow money to finance investments that they are undertaking, they are likely to turn to a bank. Banks get the funds that they lend by accepting deposits from people who have money they wish to save. Banks pay their depositors interest and charge borrowers more than they pay to depositors.

The difference between the interest rate banks charge and what they pay depositors covers the costs of accepting deposits and making loans, as well as the risk that some borrowers may be unable to repay their loans and provides profits for the bank owners.

Because most bank deposits are fully insured and can be withdrawn at any time, depositors correctly view them as having little or no risk. The value of the deposits does not fluctuate with the fortunes of the bank's borrowers, and all of the risks are borne by the bank owners.

In addition to their role as financial intermediaries, banks serve another important function in the economy—they facilitate purchases of goods and services by providing checking accounts. We will discuss this aspect of bank activities in greater detail in a later section of the guide when we turn our attention to monetary institutions.

Mutual Funds

Mutual funds provide a way for savers with small amounts of money to purchase bonds and stocks that would otherwise be difficult for them to purchase. Mutual funds purchase a portfolio of stocks and bonds and sell shares to savers. The value of the mutual fund's shares fluctuates with the value of the portfolio of assets that it owns. Mutual fund shareholders assume all of the risks of variation in the value of the shares.

Mutual funds are attractive to savers with small amounts of money for two reasons. First, mutual funds make it possible to achieve a higher degree of diversification than would be feasible through the direct purchases of stocks and bonds. Holding the stock or bonds of a single company is risky because the value of that financial asset depends on the fortunes of that one company. Diversification reduces the potential ups and downs because some companies will do well when others are suffering. For instance, discount retailers like Wal-Mart find that their sales may actually rise during recessions while department stores that cater to more upscale tastes see their sales fall. By diversifying, savers can avoid tying the value of their assets to the ups and downs of a single business.

The second advantage of saving through a mutual fund is that it provides access to the knowledge and insight of professional money managers. The skill and knowledge of these professionals mean that individuals do not have to closely follow market developments.

Saving and Investment in Aggregate

Saving occurs when individuals earn more than they wish to spend. Investment occurs when businesses or households purchase capital equipment or pay for the construction of new buildings. Before considering how

financial markets coordinate independent saving and investment decisions, we need to consider how saving and investment are measured at the aggregate level.

Recall that for an economy, production (GDP) is equal to income and to expenditures. We can express the equality of income and expenditures mathematically in the following expression: **Y = C + I + G + NX**. In this equation, Y stands for income, C is consumption expenditures, I is investment, G is government purchases, and NX is net exports. By virtue of the definitions of these quantities, this equality is an identity—it is always true.

To simplify, we will begin by assuming that the economy is closed; that is, it does not engage in any international trade. As a result, net exports are zero, and the identity between income and expenditures can be written as: **Y = C + I + G**.

Subtracting C + G from both sides of this expression, we obtain **Y – C – G = I**. The left-hand side of this expression (Y – C – G) is national savings, S, since it is the difference between income Y and expenditures by households, C, and government, G. In other words, the identity between income and expenditures implies a second important identity: savings equals investment. Written in symbols, this would be: **S = I**. Because this is an identity, by definition it is always true.

Further insight about this identity can be gleaned by some further rearrangement. In the expression above, we can add and subtract net taxes, T, from the left-hand side of the expression to obtain: **S = Y – C – G = (Y – C – T) + (T – G)**. The second and third expressions are equal because the two T terms in the last expression cancel each other out.

We can interpret this expression as saying that saving is equal to the sum of *private saving* (Y – C – T) and government saving (T – G). Private saving is the amount of money households have left over after they pay for their taxes and pay for their consumption. While taxes are an expense from the perspective of households, they are income for the government, and the difference between government income, T, and government purchases, G, is called government saving. If T – G is a positive number, then we say the government runs a *budget surplus*. If T – G is negative, then we say that it runs a *budget deficit*.

One important implication that emerges from breaking down saving into its components is that when the government runs a deficit, it reduces investment in the economy, which reduces the growth rate of living standards.

International Capital Flows in an Open Economy

In an open economy, domestic savings no longer have to equal domestic investment because of the possibility of international borrowing or lending. Nonetheless, there is an important parallel to the relationships we have just described and one that closely relates the level of international trade with domestic investment.

In an open economy, residents interact with citizens of other countries either in the world market for goods and services or in the world financial markets. In the same way that net exports measures the difference between the sale of domestically produced goods to foreigners and the purchase by domestic residents of foreign-produced goods, we can define a second concept—**net capital outflow**. The net capital outflow equals the purchase of foreign capital or financial assets by domestic residents minus the purchase of domestic assets by foreigners.

When Inbev, a Brazilian- and Belgian-owned brewing company, purchased the U.S. company Anheuser-Busch, it resulted in the purchase of domestic assets by foreign residents. This purchase added to the purchase of domestic (U.S.) assets by foreigners. Since we subtract such purchases, the net capital outflow decreased. When Intel builds a new factory in Taiwan, this results in the purchase of foreign assets by domestic residents, so it increases the first term in this expression, and increases the net capital outflow.

There are two types of international capital flows: **foreign direct investment** and **portfolio investment**. Foreign direct investment is used to describe situations in which a company or individual acquires assets in a foreign country that they will manage actively. An example of foreign direct investment in the United States is the purchase of Rockefeller Center in New York by the Japanese corporation Mitsubishi in 1989. Portfolio investment occurs when an individual or business purchases shares of stock or bonds issued by a foreign corporation. When the Chinese government purchases U.S. government bonds, it is making a portfolio investment.

In an open economy, net capital outflows (NCO) are precisely equal to net exports (NX). This equality always holds because, like the equality of saving and investment, it is an identity. To see why, it is helpful to consider an example. Suppose that Electronics Importers purchases a container full of video games from a Japanese manufacturer and pays them $100,000. This purchase is an import, so it reduces net exports by $100,000.

The Japanese video game producer could put the money in a safe. In this case, the owners of the company are using some of their income to invest in the U.S. economy by purchasing a domestic asset (U.S. currency). As a result, net capital outflows decrease by $100,000, thus balancing the change in net exports.

More realistically, the video game manufacturer might use the $100,000 to purchase U.S. government bonds. Or, they might take the money to a bank and exchange it for Yen. The company no longer has any dollars, but the situation has not really changed since now the bank faces the same choices as the company about what to do with the funds.

Another possible outcome is that the company uses the money to purchase U.S.-produced goods and services. For example, they might pay a U.S. advertising company to develop new advertisements. If they spend the entire amount of their revenue, then this causes U.S. exports of services to increase by $100,000, balancing the earlier imports. In this case, neither net exports nor net capital outflows change.

For the economy as a whole, the amount of net capital outflows must exactly equal net exports. Returning to the equality of income and expenditures for an open economy, we have: $\mathbf{Y = C + I + G + NX}$.

Rearranging the terms of this equation we obtain $\mathbf{Y - C - G = S = I + NX}$.

But, we have just shown that net exports equal net capital outflows, so we can replace NX with NCO to get $\mathbf{S = I + NCO}$. This states that domestic saving equals domestic investment plus net capital outflows.

In an open economy, savings can differ from investment, but only to the extent that the difference is offset by net capital outflows. If foreigners are willing to lend to domestic citizens (so NCO is negative), then investment can be larger than savings. Of course, foreigners make such loans with the expectation that they will be repaid at some point in the future. So, eventually the situation will likely be reversed, with saving exceeding investment to produce positive capital outflows.

How Financial Markets Coordinate Saving and Investment Decisions

We have seen that by definition saving must equal investment in a closed economy. And, even in an open economy saving and investment are closely linked with each other and with the net capital flows into or out of the economy. But, what determines the level of savings and investment that occurs in an economy?

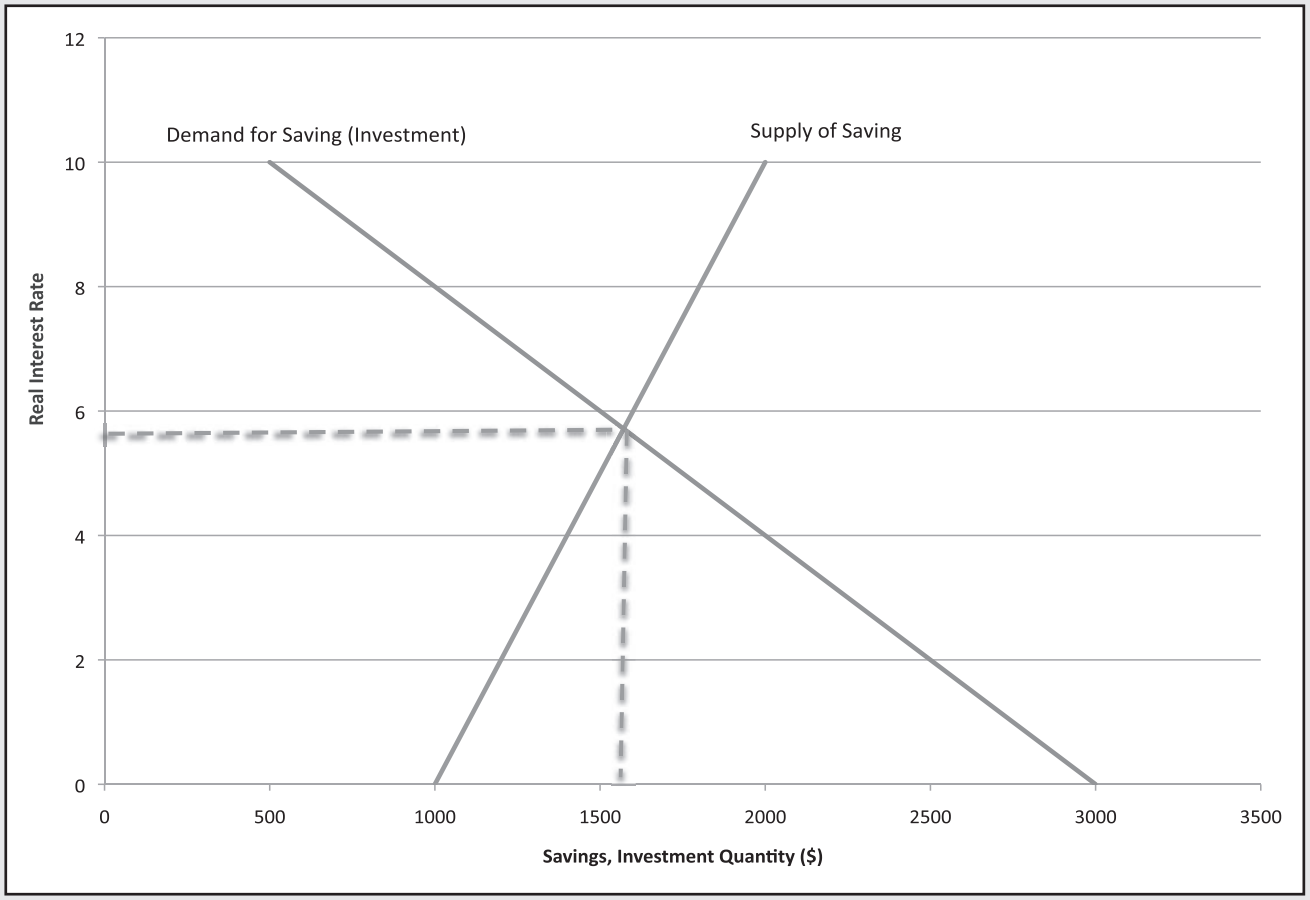
For simplicity's sake, we will again focus on a closed economy, but the situation would be quite similar in an open economy. In reality, there are a large number of financial markets, but they are all closely linked to one another because individuals with excess savings can easily move funds between markets to obtain the best return for their money, while borrowers can similarly choose between many different markets. As a result, it is convenient to collapse these many markets into a single financial market.

In the financial market, the supply of savings and the demand for savings (that is, the demand by firms for funds to purchase or construct new capital, or investment) are equalized through adjustments of the interest rate. This is illustrated in Figure 44. As before, we have graphed the quantity (in dollars) of supply (savings) and demand (investment) on the horizontal axis and the interest rate on the vertical axis.

In the financial market, the interest rate functions as the price of a loan. It is the amount that borrowers must pay for the loan, and it is the amount that savers receive for making the loan. For a lender, the decision to save a dollar today is, in effect, a decision to postpone consumption until some time in the future. Suppose the interest rate is 10 percent per year. A saver who lends $100 will receive $110 = $100 × (1 + 0.1) the following year. The possibility of consuming more in the future is one of the principal motivations for saving.

Of course, if prices are rising, the same bundle of goods becomes more expensive next year, so what matters is the real interest rate, which is the nominal rate minus the rate of inflation. If prices increase 10 percent per year, then it will take $110 next year to purchase a bundle of goods that costs $100 today. In

FIGURE 44



Equilibrium in the Financial Market

this case, the real interest rate will be zero, indicating that the saver receives no increase in purchasing power from postponing their consumption.

The higher the real interest rate is, the greater the rewards for being patient, and the greater the amount that people will choose to save. As a result, the supply of savings is drawn as an upward-sloping line in Figure 44.

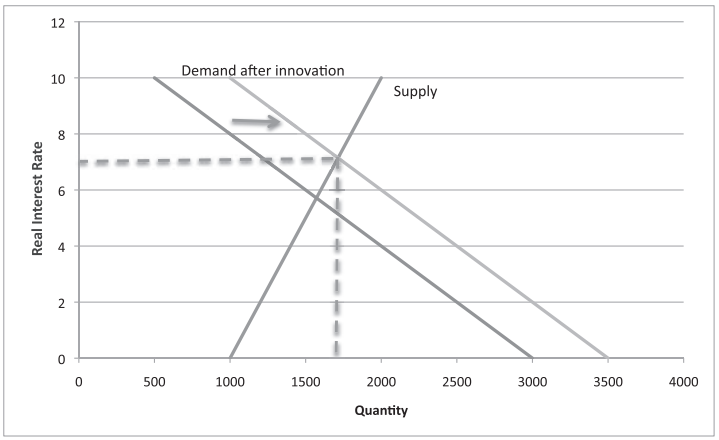
Businesses invest because they anticipate that the additional capital equipment they are acquiring will raise their revenues in the future. The price of making these investments is the real interest rate. So long as businesses expect that the additional revenues they will receive will exceed the cost of borrowing the funds, businesses will be willing to borrow. The lower the real interest rate is, the larger the number of investment projects that businesses will find profitable to pursue. As a result, the demand curve for savings is drawn as downward sloping.

In the same way that competitive forces move prices in other markets toward the market equilibrium level, there are strong pressures on the real interest rate that cause it to adjust to equilibrate the market. At an interest rate below the equilibrium level, borrowers would not be able to find enough savers willing to lend them funds, and competition to obtain the available funds would drive up the real interest rate. At an interest rate above the equilibrium, there would be an excess supply of funds, and competition between lenders to find borrowers willing to take their funds would cause the real interest rate to fall.

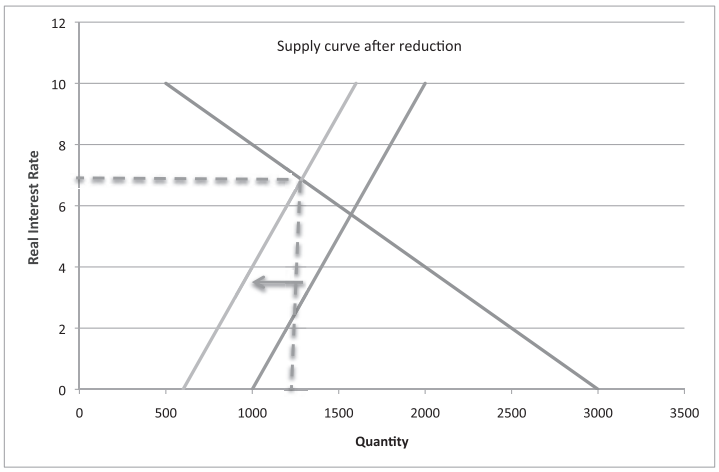
Now that we have seen how the financial market determines the real interest rate and the quantity of saving and investment, we are in a position to consider

FIGURE 45

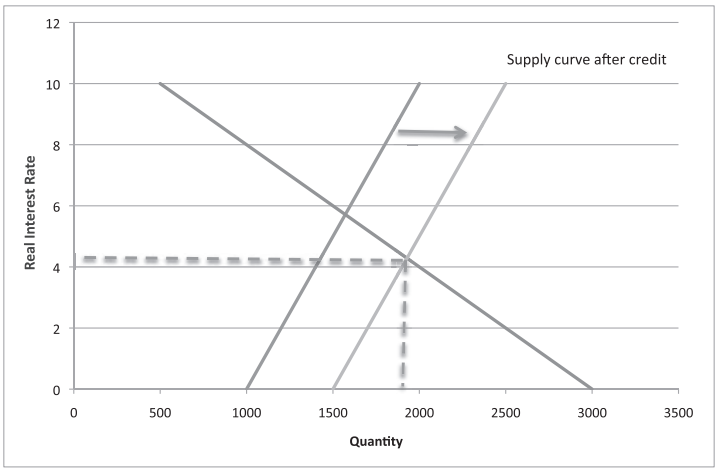
(a) TECHNOLOGICAL INNOVATION SHIFTS INVESTMENT (DEMAND FOR SAVINGS)



(b) REDUCTION IN GOVERNMENT SAVINGS SHIFTS THE SUPPLY OF SAVINGS



(c) TAX CREDIT FOR SAVING SHIFTS THE SUPPLY OF SAVING



The Effects of Supply and Demand Shifts on Financial Market Equilibrium

how various events affect this equilibrium. Figure 45 illustrates three possible changes in the market equilibrium. Panel (a) depicts the effects of a new technology that raises the productivity of capital. As a result, the demand for funds schedule shifts out to the right since businesses will want to borrow more money at every interest rate. Rising interest rates cause savings to rise, and the new equilibrium occurs at a higher interest rate and higher level of savings and investment.

In panel (b) we show the effect of an increase in the government deficit; or, equivalently, a reduction in government saving. With the government saving less or borrowing more, the supply of saving in the economy is reduced at every interest rate, which is shown as a leftward shift in the supply of savings curve. Now the equilibrium shifts up and to the left. As a result, interest rates are higher, and the total quantity of saving and investment in the economy is lower. This tendency of government deficits to reduce private investment is called **crowding out**.

The third example we will consider is the effect of a government tax credit to encourage savings. More concretely, suppose that the government reduces the tax rate on interest income earned on savings accounts. In this case, as is illustrated in panel (c), the supply of savings curve shifts out to the right. As a result, interest rates fall while saving and investment both increase.[15]

MONEY AND PRICES IN THE LONG RUN

Having grown up in an advanced market economy, it does not surprise us at all that we can walk into a store and hand over some small green pieces of paper and walk out with valuable merchandise. Nor does it surprise us that the store owner will allow us to simply swipe a credit card through a magnetic strip reader or write a check in payment for the merchandise. Money is a remarkable innovation that greatly facilitates exchange in our economy. Without it, we would be forced to barter, finding people who have the items we wish to acquire and who are willing to accept something that we are willing to give up in return.

Because money represents purchasing power, the quantity of money in circulation in an economy can have a powerful influence on the level of economic activity. Too much money can lead to inflation, and too little money can lead to deflation. This section begins by defining more precisely what we mean by money. Then, we will describe the forces that determine the quantity of money in the economy. Finally, we will consider how money affects prices and output.

What Is Money?

While we all have an intuitive sense of what money is, it is important in economic analysis to have a clearer and more precise definition. To economists, money is any asset that has three functions. It is a medium of exchange, a unit of account, and a store of value. These functions distinguish money from other assets, such as stocks and bonds, paintings, real estate, or barrels of oil. Let's consider each function of money.

- **Medium of Exchange.** A *medium of exchange* is an item that buyers can use to purchase goods and services. For money to function as a medium of exchange, sellers have to be confident that they can use the money they receive to pay for the things they wish to purchase. The usefulness of money as a medium of exchange explains why people are willing to hold onto it even though it earns no interest. The ability to quickly and easily complete a transaction compensates us for the interest payments we give up.
- **Unit of Account.** A *unit of account* is a yardstick used to establish the value of different goods and services. Expressing the prices of goods and services in a common unit of account greatly facilitates comparisons of economic value. The use of money as a medium of exchange is closely linked to its use as a unit of account. Because money is used to buy and sell things, it makes sense to express prices in money terms.
- **Store of Value.** A *store of value* is an item that people can use to transfer purchasing power from the present into the future. When a seller accepts dollar bills today in exchange for a good or service, that seller can hold onto those bills for weeks or months before becoming a buyer. Paper currency is only one of many stores of value, but—unlike stocks or bonds—it pays no interest and offers no opportunities for appreciation in value.

Economists use the term "wealth" to describe all of the different stores of value in an economy. An

FIGURE 46

M1	**$5,209.90**
Currency	$1,855.90
Demand (Checking) Deposits	$2,211.70
Other Checkable Deposits	$1,142.30
M2	**$18,166.80**
M1	$5,209.90
Savings Deposits	$11,391.80
Small Denomination Time Deposits	$428.40
Retail Money Funds	$1,136.70

Note: Outstanding amount of U.S. dollar-denominated traveler's checks of nonbank issuers. Publication of new data for this item was discontinued in January 2019. Traveler's checks issued by depository institutions are included in demand deposits.
SOURCE: https://www.federalreserve.gov/releases/h6/current/default.htm.

Components of the Money Stock, June 2020 (in Billions).

important characteristic that distinguishes different assets that make up wealth is their **liquidity**. Liquidity is a measure of the ease with which an asset can be converted into the economy's medium of exchange. Currency is clearly the most liquid asset, but deposits held in checking accounts, most stocks and bonds, and shares of mutual funds can be easily used to complete transactions and are thus also highly liquid. In contrast, real estate and collectable antiques require more effort to sell and are consequently less liquid.

Throughout history many things have functioned as money. These can be divided into two categories: commodity money and *fiat money.* When an item with some intrinsic value is used as money it is called *commodity money.* The use of precious metals such as gold or silver is an example of commodity money. Similarly, during World War II prisoners of war used cigarettes as money to trade goods and services with one another. When an item with no intrinsic value is used as money it is called fiat money. A fiat is simply an order or decree. The value of dollar bills as legal tender is established by government decree.

Measuring Money

To be able to analyze the effects of money on the economy, a first step is simply to be able to measure the amount of it. In the United States, the stock of money is made up of several components. The most obvious of these is **currency**, which includes the paper bills and coins in the hands of the public. But currency is not the only asset that functions as money. The wealth represented by your checking account is nearly as good as (if not better than) currency. By writing a check or swiping your debit card, you can use this wealth to make purchases in the same way you can use currency. Many other types of accounts, such as savings accounts or mutual fund accounts, are essentially equivalent to checking accounts.

It is not easy to draw a line between assets that are "money" and those that are not. Dollar bills in your wallet are money, whereas your house is not; but there are many assets somewhere between these two extremes. For this reason, monetary economists have developed several different measures of the stock of money in the economy. The most widely used are called M1 and M2. The table in Figure 46 lists the components of each. M2 includes all of the items in M1 plus a broad array of other assets.

Notice that neither M1 nor M2 includes credit cards as part of the stock of money, even though credit cards are often used to make purchases. The reason is that a credit card is not so much a way of making a payment as it is a way of putting off a payment. When you pay

for your groceries with a credit card, the bank that issued the card pays the supermarket, and then at a later date you pay the bank. Although credit cards are not part of the money stock, people who use them are able to pay many of their bills at one time, and they are therefore likely to hold less currency than they otherwise would. To this extent, credit cards help to reduce the economy's need for money.

The Federal Reserve System, Banks, and the Supply of Money

The amount of money in the U.S. economy is determined by the interaction between the public, commercial banks, and the Federal Reserve System. The Federal Reserve System, often called "the Fed," is the central bank of the United States. A central bank is an institution created to oversee the banking system and regulate the supply of money.

The Federal Reserve System was created in 1913 and consists of twelve regional banks owned by the commercial banks in their region, and the Federal Reserve Board in Washington, D.C. The Fed is run by a board of governors that consists of seven members who are appointed by the President and confirmed by the Senate. Governors' terms are fourteen years, which helps to insure that the actions of the Federal Reserve system are insulated from political pressures.[16]

The twelve regional banks are largely responsible for overseeing commercial banks in their respective regions and for facilitating transactions by clearing checks. They also act as a sort of bankers' bank, making loans to banks when they wish to borrow funds. When a member bank is unable to obtain funds from other sources, the Federal Reserve banks act as a *lender of last resort* to maintain the stability of the overall banking system.

The task of controlling the quantity of money in the economy, called the **money supply**, is the responsibility of the Federal Open Market Committee (FOMC). The FOMC is composed of the seven governors of the Fed plus five regional bank presidents. The president of the New York Fed is always a member, but the other four places on the FOMC rotate among the remaining banks. The FOMC meets about every six weeks in Washington, D.C., to assess the state of the economy and determine if any changes in **monetary policy** are necessary.

When the FOMC decides that the money supply should be adjusted, the Fed achieves this goal primarily through **open market operations**. If the Fed wishes to increase the money supply, then it purchases U.S. government bonds from banks or the public. As a result, the amount of currency and deposits in the hands of the public increases. If the Fed wishes to reduce the money supply, then the Fed reverses the process, selling bonds to the public and removing money from circulation.

Open market operations are a powerful tool, but by themselves they do not determine the stock of money in the economy. The money supply also depends on the behavior of banks and of the public.

Let's begin by considering how banks affect the money supply. To begin, let's suppose that there are no banks and that the money supply consists of $100 of currency. Now suppose that someone establishes a bank offering depositors a safe place to store their currency. The bank accepts currency and stores it in its vault; when a depositor wants to make a purchase, the depositor goes to the bank, withdraws the necessary funds, and uses them to make a purchase. After the transaction is completed, the seller takes the funds and deposits them in their account with the bank.

We can summarize the bank's financial position as shown in Figure 47. In this table, there are two columns: on the left we list the bank's assets, while on the right we list the bank's liabilities. The bank's assets consist of the $100 in cash that it holds in its vault; and, its liabilities are the $100 in deposits that its depositors can withdraw at any time. The bank's assets and liabilities are in balance. Whether people hold currency or place it in bank accounts, the money supply in this economy is $100.

The situation depicted in Figure 47 is simple, but it doesn't offer the bank's owners much opportunity to earn a profit, and they will have noticed that most of the money on deposit remains unused. Instead of holding all $100 in deposits, they could lend some of this out to people who wish to borrow funds to purchase a house, pay for college, or make some other major purchase. The bank needs to keep some **reserves** to be able to pay its depositors, but this is likely only a small fraction of total deposits.

Suppose that the bank owners determine that they need

FIGURE 47

ASSETS		LIABILITIES	
Reserves	$100	Deposits	$100

Bank Balance Sheet with 100 Percent Reserves

to hold reserves equal to just twenty percent of their liabilities. Then, they can lend out $80 to borrowers and receive interest income on this. Figure 48(A) illustrates the bank's situation now. On the right-hand side the bank still has $100 in liabilities, but now its assets consist of $20 in reserves and $80 in loans. Once again, assets and liabilities exactly balance.

Notice, however, what has happened to the money supply. The bank's depositors have $100 in deposits, and its borrowers have $80 in currency. The money supply has grown to $180. By holding only a fraction of deposits as reserves, the bank is able, in effect, to create money. This may seem to be too good to be true. But, it is important to understand that while the bank has created more money, it has not created any more wealth. Its borrowers have an additional $80 in assets (the money they have borrowed), but they also have an additional $80 in liabilities (the debt that they have to repay). Because of fractional reserves, the bank makes the economy more liquid, but it doesn't increase the total amount of wealth in the economy.

The process of money creation does not stop with the initial loans made by the bank. Its borrowers may deposit the loan in another account until they make a purchase with the funds. Or, once they have made a purchase, the seller will deposit the funds that they receive in their bank account. Figure 48(B) shows that now the bank's liabilities have increased to $180 and its assets have grown to $180 as well—$100 in reserves and $80 in loans.

With $180 in liabilities, the twenty percent reserve ratio suggests that the bank should hold reserves equal to $36, which means it can lend an additional $64. Figure 48(C) shows the situation once it has made these loans. Its liabilities remain the same, but now it has $144 in loans and $36 in reserves. At this point, the money supply has increased by $64, reflecting the additional loans the bank has made.

In due course the additional funds that the bank has loaned will find their way back to the bank as additional deposits. And, the cycle of loans and money creation will continue until the total deposits equal $500, and the bank has $100 in reserves and $400 in loans. At this point, the bank cannot make any additional loans without falling below its twenty percent reserve ratio.

The amount of money the banking sector creates from each dollar of reserves is called the **money multiplier**. The money multiplier is the reciprocal of the reserve ratio. If R is the reserve ratio, then each dollar of reserves will support $1/R of money supply. When banks change the reserve ratio they hold, they can alter the stock of money in the economy.

To keep matters simple, we have thus far assumed that the public holds all of its money as deposits. In reality, the public's behavior also affects the money supply through decisions about how much money to hold as bank deposits and how much to hold as currency. As we have seen, the Federal Reserve can adjust the amount of currency in circulation through open market operations.

Suppose that the Fed has provided M dollars of currency. If the public chooses to hold C dollars as currency, then the banking sector must be holding M–C in reserves. The amount of currency plus reserves is often referred to as the **monetary base** or *high-powered money*. If banks hold a fraction, R, of each dollar of deposits as reserves, then there will be (1/R) × (M–C) dollars of deposits, and C dollars of currency, so the money supply (which is deposits plus currency) will equal.

$$C + \frac{M - C}{R} = \frac{R \times C + M - C}{R} = \frac{M + (R - 1) \times C}{R}$$

FIGURE 48

Panel (a)

ASSETS		LIABILITIES	
Reserves	$20	Deposits	$100
Loans	$80		

Panel (b)

ASSETS		LIABILITIES	
Reserves	$100	Deposits	$180
Loans	$80		

Panel (c)

ASSETS		LIABILITIES	
Reserves	$36	Deposits	$180
Loans	$144		

Bank Balance Sheet with Fractional Reserves

If you experiment with this equation, you will find that the smaller that C is, or the smaller that R is, the larger the money supply will become.

In addition to open market operations, the Federal Reserve has several other tools it can use to influence the supply of money in the economy. The Fed has the power to set **reserve requirements** for commercial banks. Banks can, of course, choose to hold reserves beyond this requirement, but manipulation of required reserves is nonetheless a powerful lever. Because it is disruptive to the business of banking, however, the Fed only rarely makes changes in reserve requirements.

The third tool available to the Fed is the **discount rate**, which is the interest rate that the Federal Reserve charges on loans that it makes to banks. Although banks rarely borrow directly from the Federal Reserve because such borrowing suggests they may be in financial difficulty, the discount rate is closely linked to the **federal funds rate**, which is the rate charged by banks when they lend reserves to other banks. A higher discount rate discourages banks from borrowing reserves. Thus, raising the discount rate helps to reduce the quantity of borrowed reserves and therefore reduces the supply of money.

Bank Runs

One problem that can arise in a system based on fractional reserves occurs when the public suddenly decides that it wants to hold substantially more currency than it has been holding. Since banks have reserves equal to only a fraction of their liabilities, they will not be able to pay all their depositors. If depositors begin to fear that they may not be able to withdraw their deposits, they will hurry to the bank to get their deposits ahead of other depositors.

Such a rush of withdrawals is called a **bank run**. Even if a bank is solvent, meaning that its assets exceed its liabilities, it will not have enough cash on hand to meet all of the demand, and it will be forced to shut its doors until loans are repaid or it can borrow additional funds or sell assets. When a solvent bank experiences a spike in demand, it is the Fed's responsibility to act as lender of last resort to prevent disruptions to the banking system.

Today bank runs are very infrequent, but in the past they were a significant source of financial disruption.

Money and Inflation in the Long Run

Earlier we discussed how economists measure inflation. Figure 40 showed how the cost of living has changed since 1960. The increase in the CPI shown in Figure 40 implies that over the last half-century the cost of a fixed basket of consumption goods has increased by a factor of about 8.6—the CPI in 2019 was about 8.6 times larger than it was in 1960. While this is a significant increase in the overall price level over the long-term, there have sometimes been periods of large declines in the price level. The most significant such decline occurred during the Great Depression when the price level fell by about 25 percent from 1929 to 1933. We also saw a small decline in the CPI between 2008 and 2009 as a result of the 2008 financial crisis.

What causes the price level to rise or fall over time? To begin with, suppose that the price of a can of soda increases from $1 to $2 over some period of time. What does it mean when people are willing to give up twice as much money in exchange for a can of soda? It could be that they have come to enjoy soda more. But this is probably not the case. It is more likely that their enjoyment of a can of soda has remained the same, but, over time, the money they use to buy soda has become less valuable. In other words, inflation is more about changes in the value of money than about the value of goods.

When the economy's overall price level rises, it takes more money to purchase a fixed basket of goods. Or, looking at the matter differently, we can say that the value of money relative to goods and services has declined. It may be helpful to state this observation more formally. Suppose P is the price level—measured by the CPI or GDP deflator—then P measures the cost in dollars of a basket of goods. The quantity of goods and services that can be bought with $1 is 1/P. If P is the price of goods and services measured in money, then 1/P is the value of money measured in terms of goods and services.

In the long run, the value of money is determined in the same way as the value of any other item in an economy: by the interaction of supply and demand. We have just seen how the supply of money depends on the Federal Reserve and the banking system. When the Federal Reserve uses open market operations to sell bonds, the supply of money contracts; when the Federal Reserve uses open market operations to buy bonds, the supply of money expands. Because of fractional reserves, the effects of these actions are magnified. But, the key point is that through its policy actions the Federal Reserve can choose the supply of money.

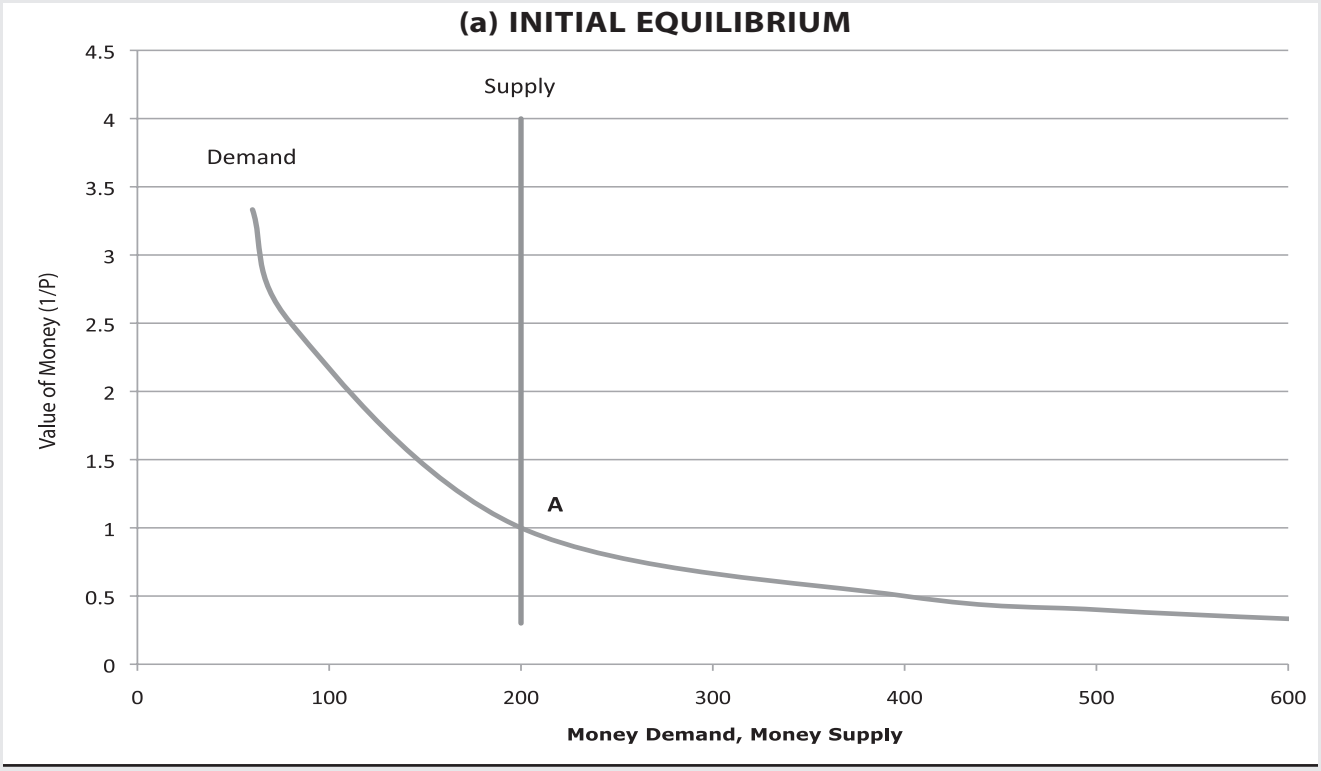
The demand for money depends on how much of their wealth people wish to hold as money, instead of in the form of other less liquid assets. The chief reason that people choose to hold money rather than other assets is because of the usefulness of money as a medium of exchange. The greater use of digital payments (for example, using Apple Pay on an iPhone) will reduce the need to use money. A desire for contactless payment, whether for convenience or to avoid possible disease transmission, as was a concern early in the COVID-19 pandemic, may speed up this transition in the coming years. But, the most important determinants of how much money people demand are the volume of transactions they engage in and the prices at which these transactions take place. Holding constant the real level of activity in the economy, we would expect that a doubling of all prices would cause the demand for money to double.

How does the economy balance people's demand for money with the level of money that the Fed chooses to supply? The answer depends on the time horizon that we are considering. For the moment, we will focus on the long run, by which we mean a time period over which the price level adjusts to equate the demand for money with the available supply.

Figure 49(a) illustrates this equilibrium. In this figure, the horizontal axis measures the quantity of money. On the vertical axis we have plotted the value of money (= 1/P). In Figure 49, the money supply is drawn as a vertical line, indicating that the Federal Reserve has fixed the supply. The demand for money is drawn as a downward-sloping line, reflecting the fact that as the value of money rises (the price level falls), people need less money to purchase a given quantity of goods and services. The equilibrium occurs at the point labeled "A" in the diagram, where the demand curve crosses the supply.

In Figure 49(b) we illustrate the effect of a doubling of the money supply. As the Fed adds to the money supply by purchasing government bonds, people find that they have more money than they want to have. They may attempt to reduce their cash holdings by

FIGURE 49



Equilibrium in the Market for Money

purchasing additional goods and services, or they may lend the additional money to someone else by depositing it in a bank or using it to buy stocks or bonds. The extra supply of savings will cause interest rates to fall and will encourage businesses and consumers to increase their spending.

The injection of more money into the economy thus causes an increase in the demand for goods and services. But, the economy's supply of goods and services has not changed. We have seen that the ability of an economy to produce goods and services depends on the available technology and on the quantities of labor, capital, and natural resources available. None of these has been changed by the additional money, so the supply of goods and services should not change.

The combination of higher demand with a fixed supply will cause the price of goods and services to rise. And, this increase in prices will continue until prices have risen enough to cause the demand for money to once again equal the supply. Once the economy has adjusted, the new equilibrium occurs at the point labeled "B." At this point, the value of money has fallen by half (or equivalently the price level has doubled). In the long run, assuming nothing else changes, the increase in prices will be exactly proportional to the change in the supply of money.

This result—that in the long run, an increase in the supply of money leads to a proportional increase in the price level—reflects the long-run **neutrality of money**. The neutrality of money means that changes in the quantity of money have no effect on real quantities in the economy. Monetary changes only affect nominal quantities. *Real* quantities are things that are measured in physical units; for example, a bushel of wheat and a ton of steel are real quantities. Nominal quantities are things that are measured in monetary units; examples would include the price of a bushel of steel or GDP in current prices.

Notice that the relative prices of different goods and services are real quantities. For example, if a bushel of wheat costs $6, and a ton of steel costs $600, then the cost of steel relative to wheat is

$$\frac{\$600/\text{ton}}{\$6/\text{bushel}} = 100\ \frac{\text{bushels}}{\text{ton}}.$$

Since dollars appear in both the top and bottom terms of this ratio, they cancel out of the equation, and we are left with a ratio of physical quantities. Similarly, if the wage rate is $10/hour and the price of an iPad is $500, then taking the ratio of the price of an iPad to the hourly wage, we can express the price of an iPad as 50 hours of work.

The neutrality of money gives rise to a very useful tool called the quantity equation. As a starting point, let us define the **velocity of money** as the average number of times a typical dollar bill is used during a year. If Y stands for real GDP and P is the price level, then the nominal GDP = P × Y measures the value of goods and services (and hence dollars) that change hands. To find the velocity of money, V, we divide P × Y by the number of dollars in circulation, M. That is: **V = (P × Y)/M**.

To see why this makes sense, let's consider a very simple economy that produces only t-shirts. If this economy produces 500 t-shirts and each sells for $5, then nominal GDP is $2,500. Suppose the supply of money is $250, then velocity in this economy is $2,500/250 = 10. For $2,500 in spending to occur using only $250 in cash, each dollar must change hands an average of ten times during the year.

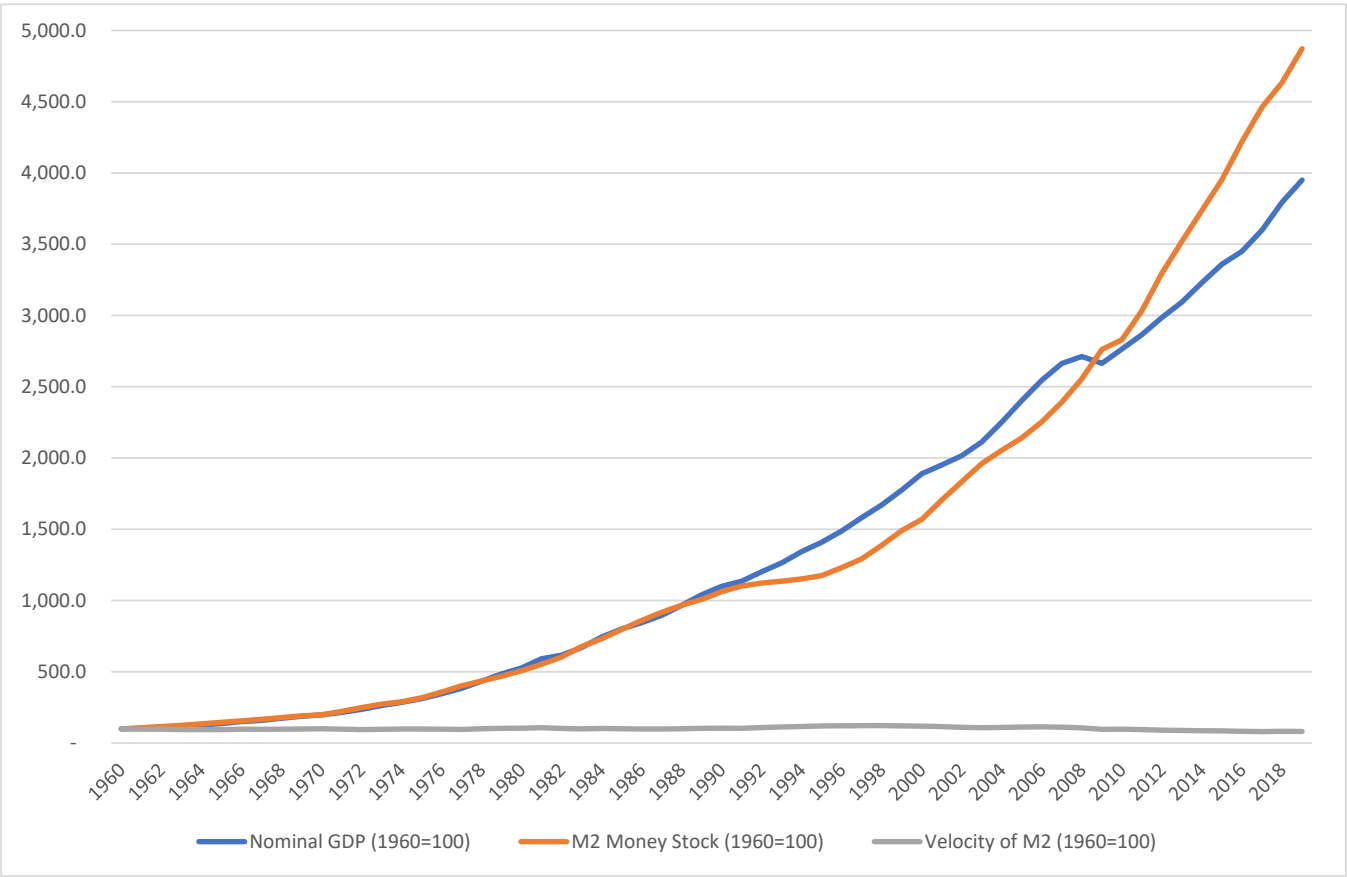
Figure 50 graphs nominal GDP, M2, and the velocity of money that they imply. As you can see in this figure, between 1960 and 2019, the growth trajectories of nominal GDP and the stock of money have followed very similar paths, with the velocity of money remaining relatively stable. Using this stability of the velocity of money, we can rearrange the quantity equation to obtain the following expression: **V × M = P × Y**.

This equation states that the velocity of money times the quantity of money will be equal to nominal GDP. So, any increase in the supply of money will be reflected in one of three ways: 1) as a fall in the velocity of money, 2) an increase in real GDP, or 3) an increase in the price level.

Why Worry about Inflation?

Inflation is unpopular. During the 1970s when inflation rates reached double digits, many consumers viewed inflation as the number one economic problem of the country. But, the neutrality of money suggests that changes in the aggregate price level should not matter because they do not affect real quantities. Despite the

FIGURE 50



SOURCES: GDP: Louis Johnston and Samuel H. Williamson, "What Was the U.S. GDP Then?" MeasuringWorth, 2020, URL: http://www.measuringworth.org/usgdp/.
M2 1960–99: Anderson, Richard G., "Federal Reserve Board monetary aggregates and major components: 1959–1999." Table Cj84-99 in *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*, edited by Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright. New York: Cambridge University Press, 2006.
M2 2000–2019: Federal Reserve Economic Data / Link: https://fred.stlouisfed.org.
Velocity: Federal Reserve Bank of St. Louis, Velocity of M2 Money Stock [M2V], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/M2V, August 12, 2020.

Nominal GDP, Money Stock, and Velocity, 1960–2019

neutrality of money, inflation does impose real costs on the economy.

First, although inflation does not alter relative prices, it does reduce the value of money. In effect, inflation is a tax on people who hold money. As prices rise, the value of the currency people have in their wallets declines relative to the goods and services they want to purchase. As a result, people will reduce the amount of money they hold. This means they have to go to the bank or ATM more frequently, which imposes an inconvenience. Inflation also imposes a cost on firms because firms have to adjust the prices of their products more frequently, and this can be a costly process.

Second, inflation introduces distortions into pricing. Because firms will not all adjust their prices at the same time, relative prices will not always accurately reflect the relative costs of production. Recall that these prices play an important role in coordinating economic decisions in market economies. Because of these distortions, the information conveyed by market prices becomes less valuable.

Third, inflation introduces confusion about the true value of goods and services in the future. Remember

that when someone with savings lends it, they are compensated by an interest payment for postponing their use of that money until a future date. But, if they cannot accurately forecast the rate of inflation, they cannot calculate how much purchasing power they will have in the future. Uncertainty about the rate of inflation adds to the risks that both borrowers and lenders face in credit markets, and this increased risk reduces both the supply of savings and the demand for investment. Because investment is crucial to economic growth, inflation reduces economic growth.

SHORT-RUN ECONOMIC FLUCTUATIONS

We noted earlier that macroeconomics is concerned with two issues: the long-term growth of the aggregate economy and short-term fluctuations. In the preceding sections, we have developed a framework for understanding the forces that determine the long-run performance of national economies. This theory provides a useful description of how the economy evolves over long periods of time of several decades or more. But, it does not provide much guidance for understanding the shorter-run deviations of economic growth from these long-run trends.

In Figure 30, we graphed the growth of real GDP in the United States between 1900 and 2019. If you look closely at that figure, you can see that superimposed on the upward trend in total output are some significant fluctuations. In particular, the drop in output during the Great Depression stands out, as does the rapid growth of production during World War II. The downturn during the 2008 financial crisis is also clearly evident. According to the National Bureau of Economic Research (NBER), a *recession* is a period between a peak and a trough in economic activity, and an *expansion* is a period between a trough and a peak in economic activity. During a recession, a significant decline in economic activity spreads across the economy and can last from a few months to more than a year. Similarly, during an expansion, economic activity rises substantially, spreads across the economy, and usually lasts for several years. In both recessions and expansions, brief reversals in economic activity may occur—a recession may include a short period of expansion followed by further decline; an expansion may include a short period of contraction followed by further growth. A *depression* is a particularly severe or protracted recession.

The recurrent alternation of expansions and recessions is commonly referred to as the *business cycle*. Business cycles have been a characteristic of industrial societies since at least the late eighteenth century. The table in Figure 51 shows the dates and duration of U.S. business cycles. A commonly used rule of thumb is that periods when real GDP declines for two consecutive quarters are recessions. The determination of the dates on which recessions and expansions begin and end is performed by the NBER, a non-profit organization of economists that has been a major source of research on short-term fluctuations in the economy. The NBER considers a broad array of different economic indicators in fixing the dates listed in Figure 51.

Looking at the data in Figure 51, the longest and deepest period of recession is the 43-month decline that began in August 1929, which has come to be known as the Great Depression. During this episode, the nation's real GDP fell by more than one-quarter. Since the Second World War, periods of recession have tended to be relatively short, with only three stretching longer than twelve months, and relatively mild in terms of the decline in real GDP. Expansions have tended to be much longer than the recessions, with most lasting more than two years—a fact that is reflected in the sustained upward trend of real GDP.

We will begin our examination of short-run fluctuations by describing their characteristics in greater detail. We will then develop a model that can account for recessions and expansions and will use this model to consider the role that government economic policy can play in mitigating the negative effects of business cycle fluctuations.

Characteristics of Short-Run Fluctuations

Expansions and recessions have effects that are visible throughout the economy and are characterized by systematic patterns of change in a wide array of different macroeconomic variables. Two of the most important correlates of fluctuations in the economy's aggregate growth are unemployment and inflation.

Figure 52 shows the unemployment rate from 1960 through 2019. Recessions are generally characterized by rising unemployment. Typically businesses are slow to increase hiring in the early phases of an expansion,

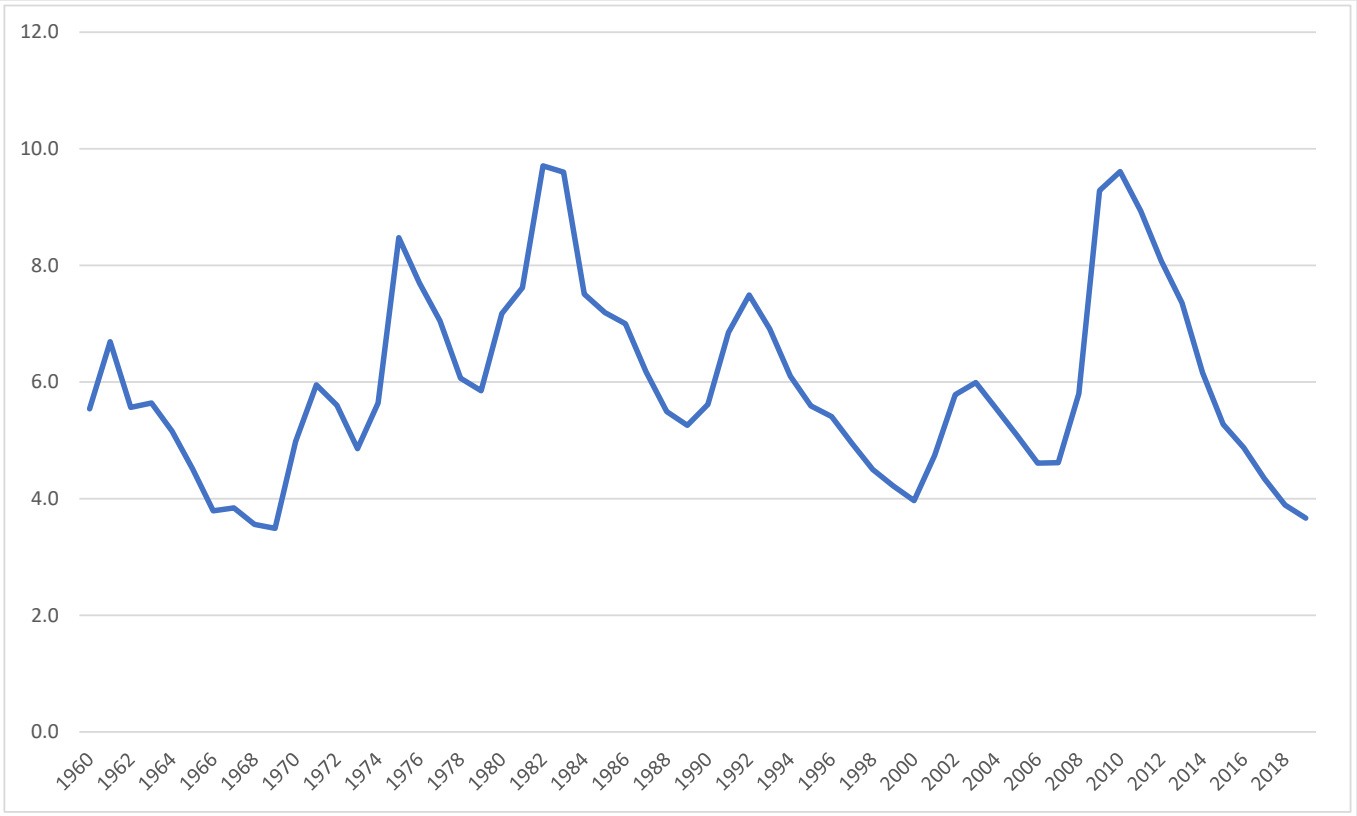
FIGURE 51

Peak	Trough	Contraction	Expansion Cycle		
Quarterly dates are in parentheses		*Peak to trough (months)*	*Previous trough to this peak (months)*	*Trough from previous peak (months)*	*Peak from previous peak (months)*
June 1857(II)	December 1858 (IV)	18	30	48	--
October 1860(III)	June 1861 (III)	8	22	30	40
April 1865(I)	December 1867 (I)	32	46	78	54
June 1869(II)	December 1870 (IV)	18	18	36	50
October 1873(III)	March 1879 (I)	65	34	99	52
March 1882(I)	May 1885 (II)	38	36	74	101
March 1887(II)	April 1888 (I)	13	22	35	60
July 1890(III)	May 1891 (II)	10	27	37	40
January 1893(I)	June 1894 (II)	17	20	37	30
December 1895(IV)	June 1897 (II)	18	18	36	35
June 1899(III)	December 1900 (IV)	18	24	42	42
September 1902(IV)	August 1904 (III)	23	21	44	39
May 1907(II)	June 1908 (II)	13	33	46	56
January 1910(I)	January 1912 (IV)	24	19	43	32
January 1913(I)	December 1914 (IV)	23	12	35	36
August 1918(III)	March 1919 (I)	7	44	51	67
January 1920(I)	July 1921 (III)	18	10	28	17
May 1923(II)	July 1924 (III)	14	22	36	40
October 1926(III)	November 1927 (IV)	13	27	40	41
August 1929(III)	March 1933 (I)	43	21	64	34
May 1937(II)	June 1938 (II)	13	50	63	93
February 1945(I)	October 1945 (IV)	8	80	88	93
November 1948(IV)	October 1949 (IV)	11	37	48	45
July 1953(II)	May 1954 (II)	10	45	55	56
August 1957(III)	April 1958 (II)	8	39	47	49
April 1960(II)	February 1961 (I)	10	24	34	32
December 1969(IV)	November 1970 (IV)	11	106	117	116
November 1973(IV)	March 1975 (I)	16	36	52	47
January 1980(I)	July 1980 (III)	6	58	64	74
July 1981(III)	November 1982 (IV)	16	12	28	18
July 1990(III)	March 1991(I)	8	92	100	108
March 2001(I)	November 2001 (IV)	8	120	128	128
December 2007 (IV)	June 2009 (II)	18	73	91	81
February 2020 (2019 IV)			128		146

SOURCE: https://www.nber.org/cycles.html.

Business Cycle Peaks, Turning Points, 1857–2020.

FIGURE 52



SOURCE: U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, August 13, 2020.

U.S. Unemployment Rate, 1960–2019

so declines in unemployment typically lag somewhat behind the onset of the next phase of economic growth.

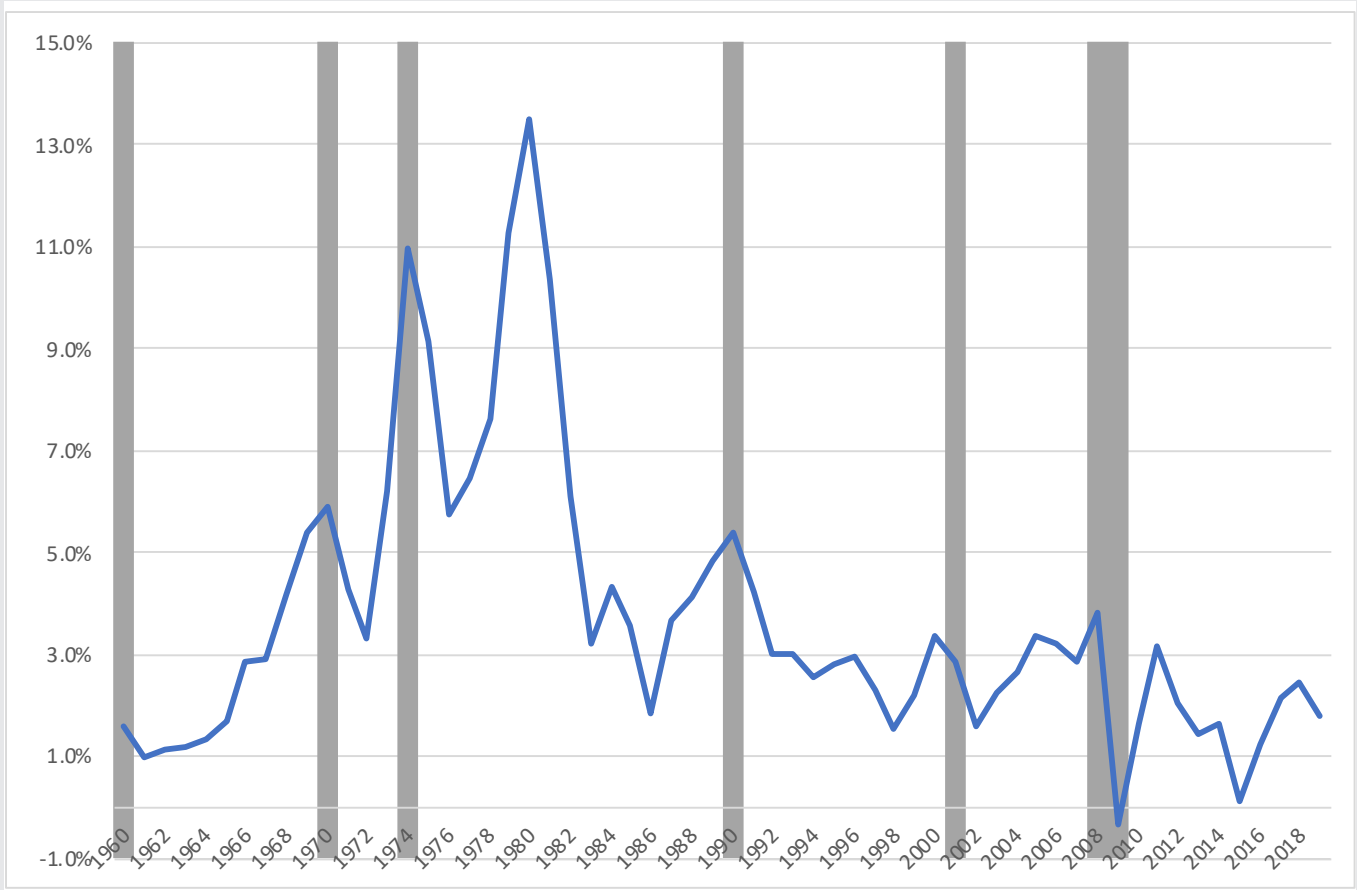
Like unemployment, the rate of inflation is also tied to the business cycle. Periods of expansion are often characterized by accelerating inflation, and recessions typically are linked to a slowing in the rate of inflation. Figure 53 graphs the rate of inflation since 1960. Between 1960 and 1979, there was a generally upward trend in the rate of inflation, which makes the business cycle effect somewhat difficult to see. But, if you look closely, you can see that the rate of inflation was declining during recessions.

Potential Output, the Output Gap, and the Natural Rate of Unemployment

In thinking about the short-run performance of the economy, it is useful to think of the actual level of GDP at any time as consisting of two parts: the **potential output** of the economy and an **output gap**. Potential output is the quantity of goods and services that the economy can produce when using its resources (such as capital and labor) at normal rates. The level of potential output is not fixed, of course, but increases over time as technology improves, and the economy accumulates additional resources.

In the subsequent discussion, we will use the variable Y* to denote potential output. The output gap consists of the difference between actual output, which we'll denote by Y, and potential output. In other words, the output gap = Y – Y*. Figure 54 plots the growth of actual output in the postwar period along with the trend growth of output between successive business cycle peaks, which approximates the growth of potential output. Relative to the trend growth of output, deviations appear small in this figure, but they nonetheless result in significant economic hardships.

FIGURE 53



SOURCE: Lawrence H. Officer and Samuel H. Williamson, 'The Annual Consumer Price Index for the United States, 1774–Present,' MeasuringWorth, 2020.

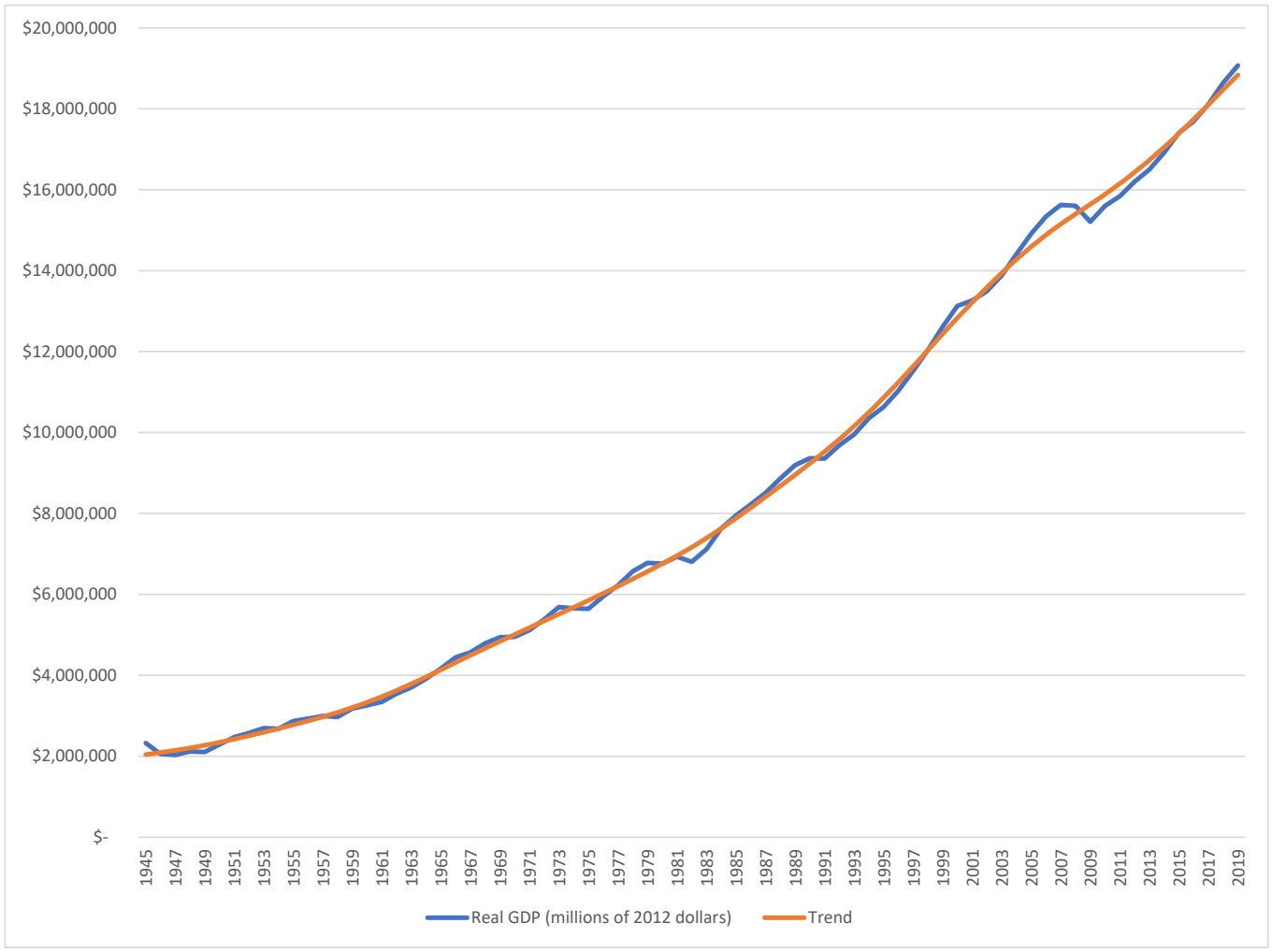
U.S. Price Inflation, 1960–2019. Average 1982–84 = 100.

When output is below potential output, the economy's productive resources are not being completely utilized. In particular, unemployment rises when the economy is below its potential output. Recall that unemployment is conventionally divided into frictional, structural, and cyclical components. The cyclical component is the part that rises when the economy is in a recession. Economists call the level of unemployment due to frictional and structural causes the **natural rate of unemployment**. It is the level of unemployment that would exist when the actual output is equal to potential output.

The natural rate of unemployment varies over time due to changes in the labor market. During the 1970s and 1980s, the entry of many more women into the paid labor force helped to raise the natural rate of unemployment, as did the decline of traditional manufacturing industries and the growth of the service sector. More recently, the natural rate of unemployment has fallen.

In the early 1960s, Arthur Okun, who was one of President Kennedy's chief economic advisors at the time, noted that there was a relationship between the output gap and the level of cyclical unemployment. Specifically, he observed that every one percent that the unemployment rate differed from the natural rate was associated with a two percent deviation in the output gap. In other words, if cyclical unemployment increased from 1 percent to 2 percent, then the output gap would rise from 2 percent to 4 percent. This relationship is called **Okun's Law**.

FIGURE 54



SOURCE: Louis D. Johnston and Samuel H. Williamson, "What Was the U.S. GDP Then?" MeasuringWorth, 2020.
URL: http://www.measuringworth.org/usgdp/.

Actual and Trend Real Output for the U.S. Economy, 1945–2019

Explaining Short-Run Fluctuations in Output

What explains the recurrent alternation between periods of expansion and recession in the aggregate economy? Logically, variations in the rate of growth of output over time could be caused either by changes in the growth rate of potential output or could occur because actual output falls above or below potential. The rate of growth of potential output depends on the growth rate of the population, the rate at which the capital stock increases, and changes in the pace of technological advances. Over long periods, shifts in these underlying forces do produce important modulations in the pace of economic growth. But, most of the short-run variation in the level of economic activity appears to be due to the divergence between actual and potential output.

In a world in which prices would adjust immediately to balance supply and demand in all markets, the economy's resources would always be fully employed, and actual output would not deviate from potential output. Accounting for deviations of actual output from potential output requires that we modify the basic microeconomic model of markets to account for the fact

that in many markets prices do not adjust immediately.

The most common approach to modifying our model of the economy rests on the observation that in many parts of the economy, firms do not constantly adjust prices in response to fluctuations in market demand. Instead, firms tend to set prices and sell as much or as little as is demanded. It is only after a sustained period of imbalance between demand and desired supply that firms adjust prices.

Because in the short run firms respond to variations in demand by adjusting production rather than prices, output in the economy is determined by the level of aggregate demand rather than by potential output. Aggregate demand is the total desired spending on final goods and services by everyone in the economy. Recall that when we discussed the equality of GDP and expenditures, we saw that total expenditures had four components:

- *Consumption (C)* is spending by households on final goods and services.
- *Investment (I)* is spending by firms on new capital goods, such as machinery and structures, as well as spending on the construction of new houses and apartment buildings. In addition, increases in inventories are also included in investment.
- *Government purchases (G)* is spending by governments—federal, state, and local—on goods and services. Transfer payments, such as Social Security benefits and unemployment insurance, as well as interest payments on government debt are not included in this category.
- *Net Exports (NX)* is the difference between the value of goods and services produced domestically and sold to foreigners and the value of goods and services produced abroad and purchased by domestic residents.

Although firms initially respond to variations in demand by adjusting quantities, in the longer run firms will adjust their prices to move back toward their normal level of production. When demand is above their desired level, firms will raise prices, causing inflation to accelerate; when demand is below their normal level of production, firms will lower prices, causing inflation to slow.

Over the long run, price changes eliminate the gap between actual and potential output and ensure that the economy's resources are once again fully employed. Because these adjustments can take a significant amount of time, however, there may be the potential for government policies to help eliminate output gaps more quickly.

This explanation for short-run fluctuations in the level of economic activity was developed by the British economist John Maynard Keynes (1883–1946) in his 1936 book *The General Theory of Employment, Interest, and Money.* The theory that Keynes developed in this book was a response to what he perceived as the inadequacy of prevailing microeconomic models to account for the events of the Great Depression. In recognition of Keynes's contribution, the resulting model of the economy is often called the **Keynesian model**.

According to Keynesian theory, the causes of short-run fluctuations in the level of economic activity can be summarized in terms of the interaction between an aggregate demand (AD) curve and a short-run aggregate supply (AS^{SR}) curve, as is illustrated in Figure 55. In addition to the short-run **aggregate supply curve**, the diagram also includes a long-run aggregate supply curve, which is drawn as a vertical line at the point Y=Y*; that is where output equals potential output.

In this diagram, the horizontal axis measures real GDP, and the vertical axis measures the aggregate price level. This diagram looks quite similar to the demand and supply diagrams we have used before to analyze individual markets, but it is important to understand that the reasons for the shapes of the AD and AS^{SR} curves are entirely different from the demand and supply curves we have considered up to now.

The Aggregate Demand Curve

We will begin by considering the derivation of the **aggregate demand curve**. Recall that in the conventional analysis of a single market, the quantity demanded increased as the price fell primarily because a lower price meant that the good in question had become less expensive relative to other goods. For example, as the price of bagels falls, they become a better bargain relative to muffins, and people will buy fewer muffins and more bagels. At the level of the aggregate economy, however, such an explanation no longer makes sense since a decline in the aggregate

price level means that the prices of all goods and services have declined.

If shifts in relative prices don't account for the downward-sloping aggregate demand curve, then what does? There are three reasons for the negative relationship between aggregate demand and the aggregate price level.

Wealth Effects

With a fixed supply of money, when the aggregate price level declines, the money that people have in their wallets and bank accounts will allow them to purchase a greater quantity of goods and services. Lower prices in effect increase their wealth and encourage a higher level of spending. Notice that this conclusion follows directly from the quantity equation that we introduced earlier ($M \times V = P \times Y$). If the velocity of money (V) doesn't change and the quantity of money (M) in the economy is constant, then a lower price level (P) must lead to a higher level of real GDP (Y).

Interest Rate Effects

At a lower price level, people will find that they are holding more money than they want to have. As we saw when we analyzed the demand for and supply of money, when people are holding more money than they view as optimal, they will attempt to reduce their monetary assets by using their money to acquire less liquid assets, including bank certificates of deposit, stocks, and bonds. All of these actions increase the supply of savings. Increased saving causes interest rates to fall and encourages households and firms to borrow more funds and increase their spending.

Foreign Exchange Effects

The third channel through which a lower aggregate price level affects aggregate demand is exports. At a lower domestic price level, domestically produced goods and services are less expensive relative to foreign-produced goods. As a result, domestic consumers will buy fewer imported goods and services, and foreign consumers will purchase more domestically produced goods and services, causing net exports to increase.

Now that we have explained the downward slope of the aggregate demand curve in Figure 55, we can consider the factors that influence its position. Anything that influences the consumption decisions of households or foreign residents, or leads firms to increase investment, will cause the aggregate demand curve to shift. The introduction of a promising new technology, such as high-speed satellite internet, will cause some firms to increase investment spending. For instance, Elon Musk's SpaceX company is manufacturing over a hundred internet satellites per month on its way to building a network with 12,000 satellites that will be able to provide internet from orbit to anywhere in the world. Such an increase in investment spending increases aggregate demand at every price level and causes the AD curve to shift to the right.

Similarly, changes in consumer sentiment will affect consumption spending and will shift the AD curve. For example, a drop in stock prices, such as occurred in the wake of the 2008 financial crisis, reduces wealth and causes consumers to reduce their level of spending at every price level. Such a change would be reflected as a leftward shift of the AD curve.

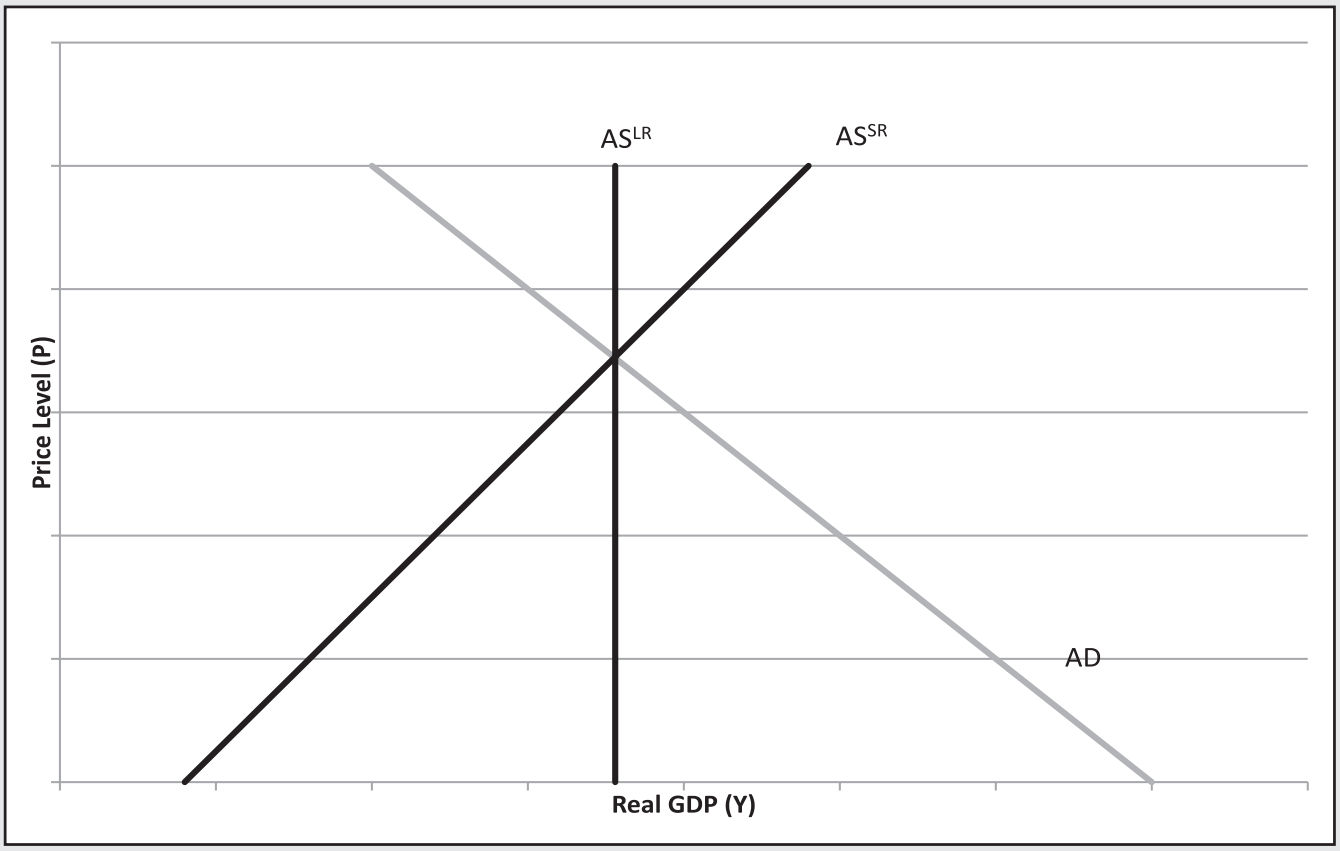
The AD curve can also be shifted by changes in government spending or taxes. An increase in spending by the federal government will, other things being equal, increase spending at every price level, a change that can be illustrated as a rightward shift in the AD curve. When state governments reduce spending by furloughing employees as some did during the recession of 2007–2009, this reduces spending and shifts the AD curve to the left. When the government reduces taxes, it increases households' disposable income, which should increase consumption spending, which will shift the AD curve to the right.

The Aggregate Supply Curve

In Figure 55, the aggregate supply curve is drawn as upward sloping, indicating that the quantity of goods and services supplied is an increasing function of the aggregate price level. Once again, while the shape of the aggregate supply curve is similar to the supply curves we encountered earlier in our microeconomic analysis of markets for particular goods, it is upward sloping for different reasons. In the microeconomic analysis of markets, the supply curve was upward sloping because higher prices are necessary to attract resources from producing other products. For example, in the market for bagels, as the price of bagels increases, bakers who were previously producing muffins will be induced to shift over to bagel production. At the aggregate level, however, resources cannot be shifted from other less profitable activities.

As we noted at the beginning of our discussion of

FIGURE 55



Aggregate Demand and Aggregate Supply in the Keynesian Model

short-run economic fluctuations, many firms do not immediately adjust prices in response to variations in demand. Instead, they fix prices for some period of time and sell as much or as little as consumers choose to purchase. Over time, firms adjust their prices in response to the gap between actual and anticipated sales. The aggregate supply curve slopes upward to reflect the relationship between this price adjustment process and the size of unanticipated sales.

The position of the aggregate supply curve depends on the economy's long-run potential output and on what people expect the aggregate price level to be. The short-run aggregate supply curve will pass through the vertical line at Y* at a price level equal to the prevailing expectation about aggregate prices. Resources will be fully employed, and aggregate supply will equal its long-run potential output, Y* in our earlier discussion, when the aggregate price level is equal to the level that firms and consumers anticipated.

Thus, there are two reasons for the short-run aggregate supply curve to shift. The first and most common cause of shifts in the position of the aggregate supply curve is changes in the expected price level. Since AS^{SR} is equal to Y* at the expected aggregate price level, an increase in the expected price level will cause the aggregate supply curve to shift upward. A decrease in the expected price level will cause the aggregate supply curve to shift downward.

The second cause of shifts in the aggregate supply curve is aggregate supply shocks. For example, weather and climate conditions that affect agricultural production may shift the aggregate supply curve. An especially good harvest means that more agricultural commodities are available at every price, an event that would cause the aggregate supply curve to shift rightward. An important example of a shock to aggregate supply in recent history is the OPEC-initiated oil embargo that began in 1973. Because of the importance of fossil

fuels as a source of power throughout the economy, the shortage of imported oil had a widespread effect on the U.S. economy, causing a reduction in quantities supplied at every price, and a leftward/upward shift of the short-run aggregate supply curve.

In addition to these movements of the short-run aggregate supply curve, technological progress will cause the economy's potential output to shift out to the right over time. It is this increase in potential output that accounts for the long-run growth of real GDP we noted at the beginning of this section of the resource guide.

The Keynesian Model of Short-Run Fluctuations

At any point, the Keynesian model implies that the economy's aggregate production and price level are determined by the intersection of AD and AS^{SR}. Figure 55 is drawn so that this intersection occurs at the point where actual output is equal to its potential, so the output gap is zero. At this point, there is no cyclical unemployment, and resources are being fully employed. But, if the aggregate supply or aggregate demand curve were to shift for some reason, the intersection of the two curves would be either above or below potential output.

We will begin by illustrating the effects of a negative aggregate demand shock. In February 2020, the United States economy went into recession as a result of the nationwide shutdowns that followed the spread of COVID-19 in the country. As consumers stayed at home, and many business closed in an effort to slow the spread of the virus, consumption and business investment spending both fell sharply. Government spending increased to try to fill the gap, but the end result was a recession.

In panel (a) of Figure 56, we show the combined effect of these events as a leftward shift in the AD curve. As a result, the economy's short-run equilibrium now occurs at Y1, which is less than Y*. The economy is now in a recession because of the shift in consumer sentiment.

As the equilibrium point in Figure 56 shifts down and to the left along the AS^{SR} curve, some businesses lower their prices, and the aggregate price level begins to decline, a response that moderates the impact of the shift in consumer sentiment.

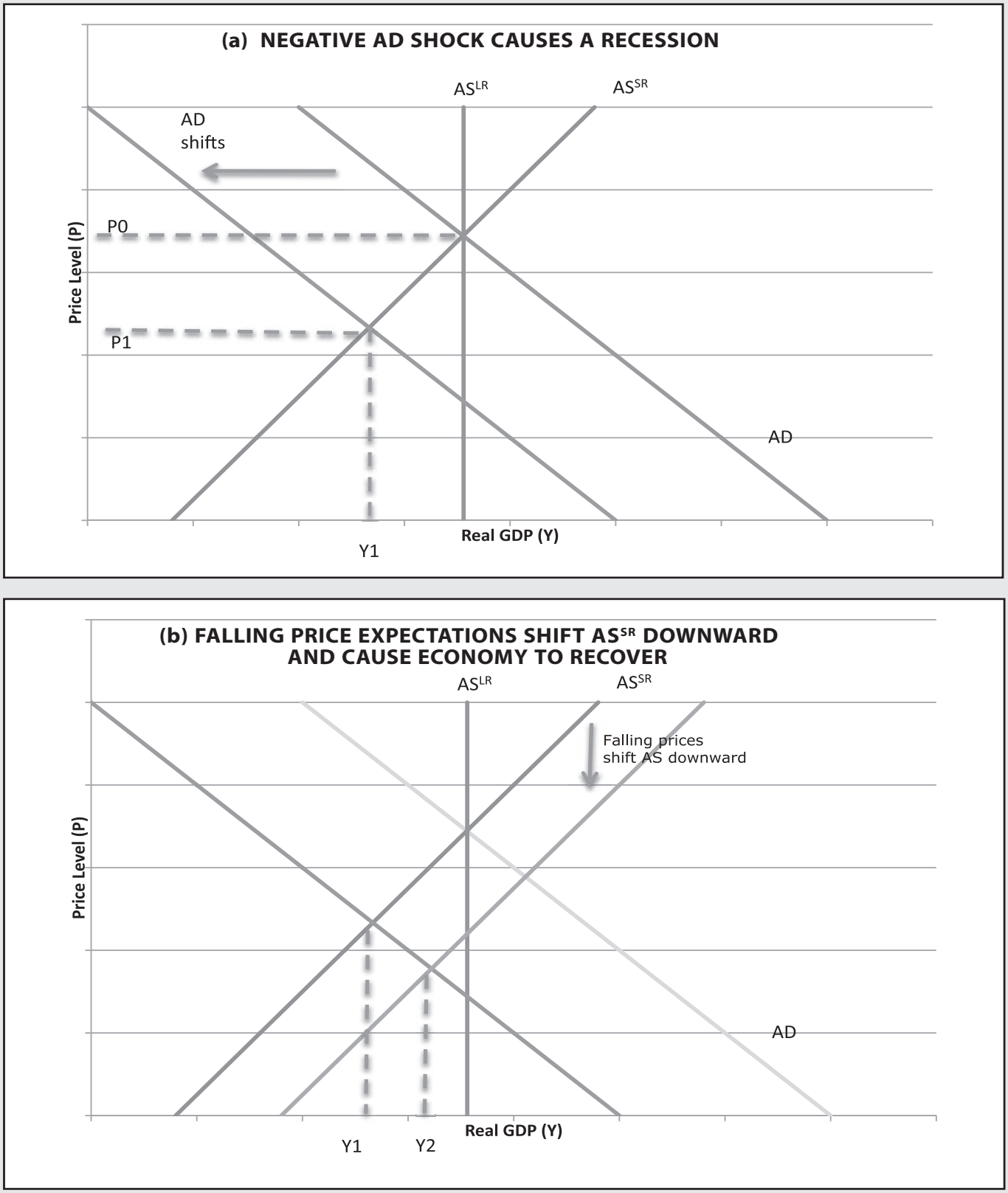
Panel (b) of Figure 56 illustrates the adjustment of the economy as it begins to recover from the recession. The position of the aggregate supply curve in Figure 56(a) indicates that businesses and households were expecting that prices would be at P0, since this is where the curve passes through the point where Y=Y*. However, the actual price level, P1, is now below P0. As time passes and firms find that they are selling less, they will begin to adjust their expectations about the aggregate price level downward. As a result of the decline in prices, output has increased to Y2, but it is still below potential output. Prices will continue to fall until the AD and AS^{SR} curves once again intersect at the point where output equals potential (Y*).

In Figure 57 we illustrate a recession caused by an aggregate supply shock such as the OPEC oil embargo of 1973. The shortage of petroleum and higher prices of gasoline and other products is shown in Figure 57 (a) as a leftward shift of the AS^{SR} curve. Beginning at the point Y=Y* and price level P0, output falls to Y1, and the aggregate price level rises to P1. With the economy in recession, firms find that actual sales are falling short of expectations. Eventually they will begin to cut prices, causing the AS^{SR} curve to shift out to the right. This movement is illustrated in Figure 57(b). Prices will fall until output once again equals potential. At this point, the price level will have returned to its original level at P0.

Figures 56 and 57 illustrate the basic explanation of recessions and expansions in the Keynesian model. Recessions and expansions occur because of the sequence of unpredictable shocks to aggregate demand or aggregate supply that strike the economy and cause the equilibrium level of production to move away from its potential. The reason these shocks are translated into recessions or expansions is because of the short-run inflexibility of prices. If prices everywhere in the economy adjusted instantly to the effects of shocks, then output would never differ from potential. It is the short-run inflexibility of prices that is the basic explanation for recessions and expansions.

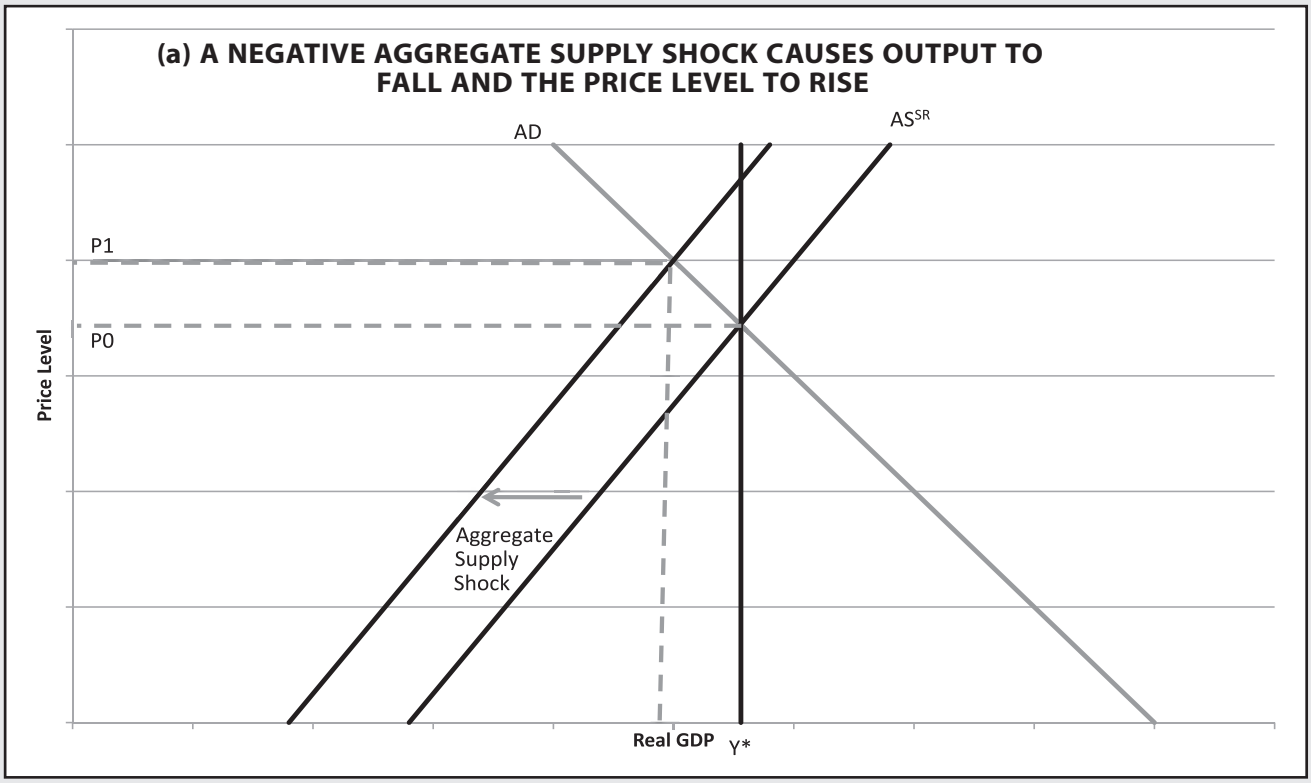
You will notice that up until now, we have been somewhat vague about the period of time that is represented by the "short run." That is because the definition of the short run is effectively the period of time in which the performance of the economy deviates from the predictions of the long-run model. Judging from the length of typical economic cycles, this is usually from one to three years.

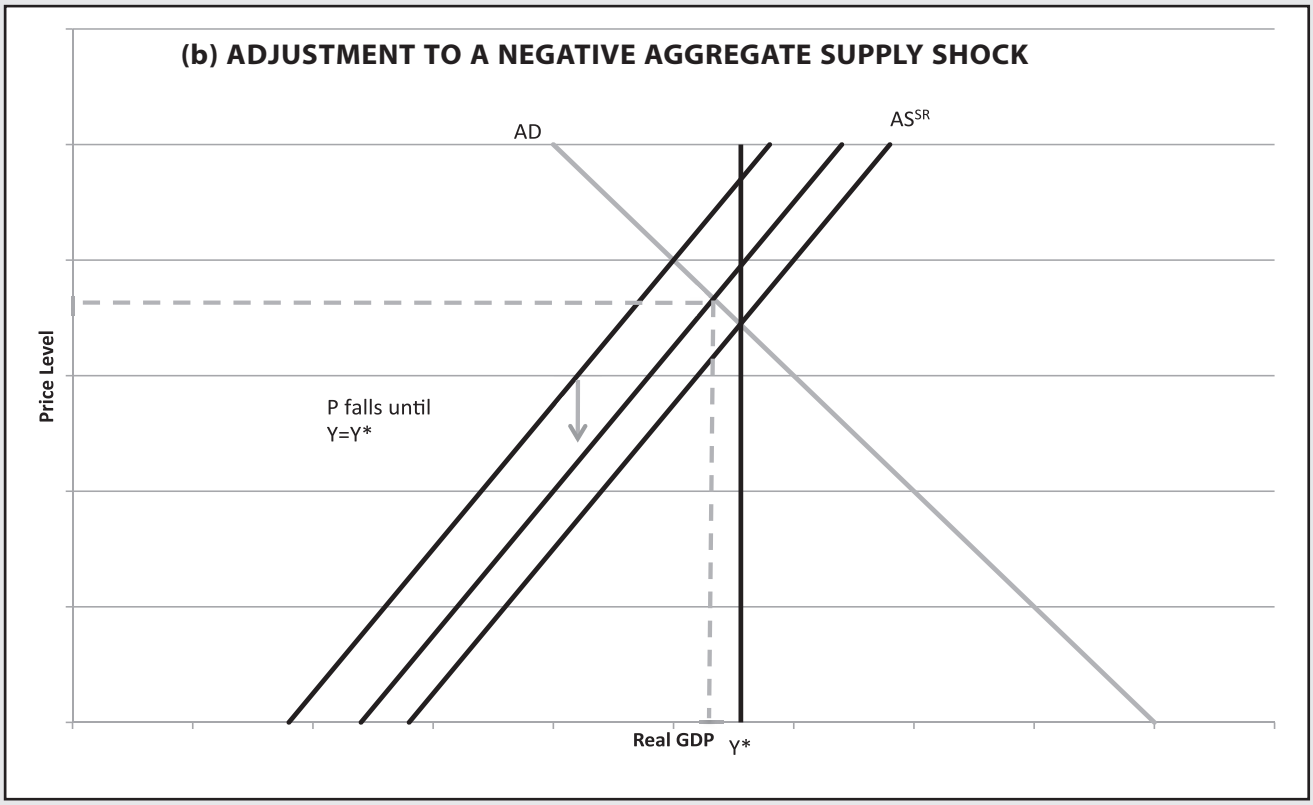
FIGURE 56



Aggregate Demand and Aggregate Supply in the Keynesian Model

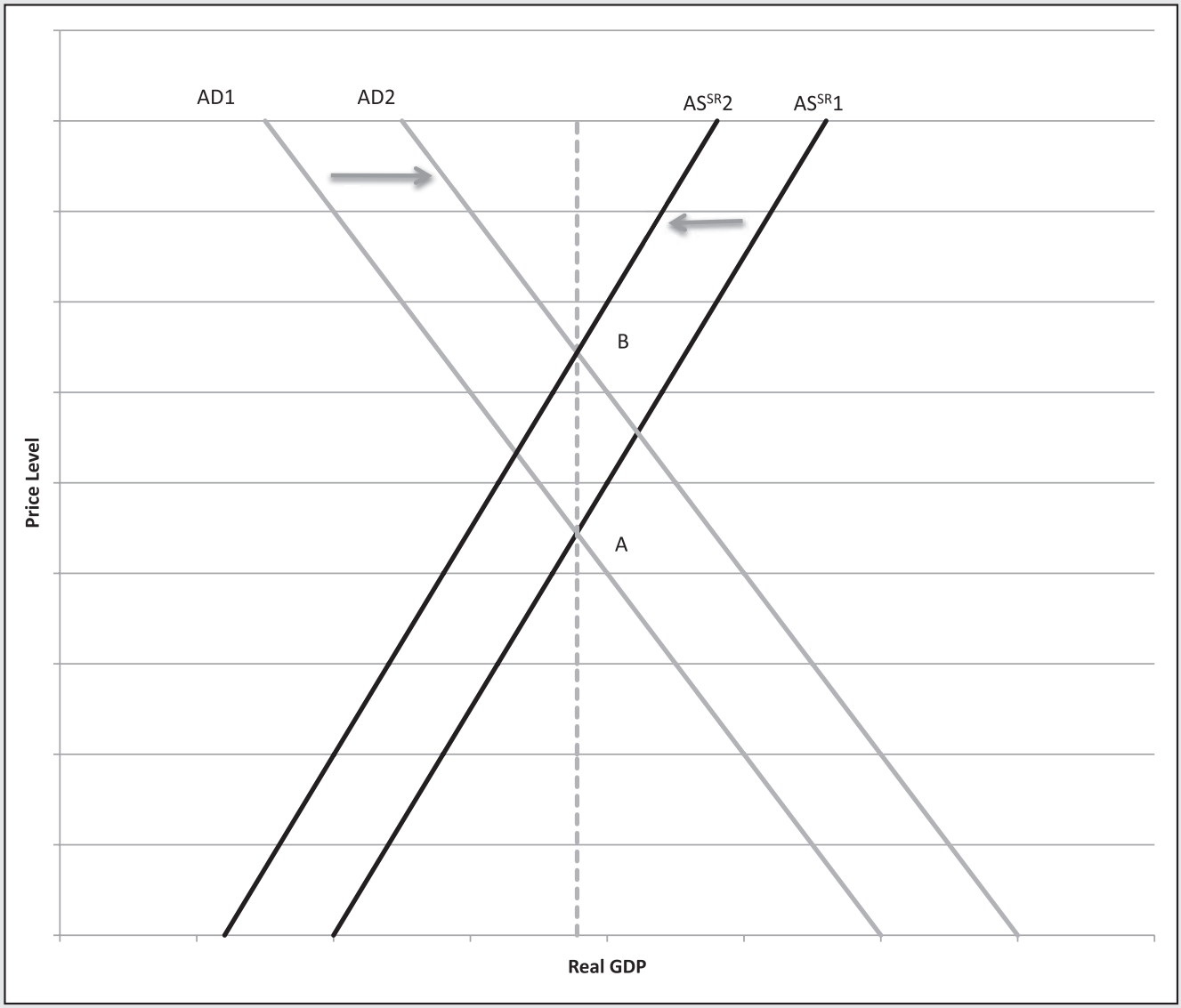
FIGURE 57





Effects of an Aggregate Supply Shock

FIGURE 58



Effects of a Fully Anticipated AD Shock

Inflation in the Keynesian Model

The model of recessions and expansions we have sketched so far has assumed that the level of inflation in the economy is zero. That is, we have drawn the AD and AS^{SR} curves on the assumption that everyone believes the aggregate price level is stable. We have seen, however, that since the Second World War the aggregate price level has followed a generally upward trend. Moreover, as the previous section demonstrated, the process of adjustment by which the economy returns to full employment after a shock involves changes in the price level.

To begin with, we need to ask why an economy would experience persistent inflation. The quantity equation implies that in the long run, the aggregate price level can rise only if the money supply is growing faster than the economy's potential output. Suppose, for example, that because of technological change, potential output increases 2 percent per year, and the stock of money increases at 5 percent per year. The quantity equation can be rearranged to show the price level must equal $(M \times V)/Y^*$. So long as velocity is constant, prices will rise at 3 percent (= 5% – 2%) each year.

In the aggregate demand–aggregate supply model, an increase in the money supply causes the AD curve to shift to the right. But, if people have become accustomed to an increasing money supply and rising prices, then they will expect the price level to rise each year, and the AS^{SR} curve will shift upward so that AD and AS^{SR} continue to intersect at the economy's potential output. This is illustrated in Figure 58.

Thus, full employment equilibrium is consistent with any anticipated level of inflation. In this context, unexpected shocks that move the economy away from full employment cause actual inflation to deviate from the anticipated level. For example, suppose that the federal government decides to begin a military buildup, but chooses to finance it through borrowing because it is afraid that increased taxes will be unpopular. This is roughly what happened in the 1960s under President Lyndon Johnson.

The increased government spending causes output to increase precisely because it is not anticipated. In the long run, however, there is no policy that will maintain output at a level different from the economy's potential output. Thus, when we use the aggregate demand–aggregate supply model, we need to remember that the changes in aggregate prices that it indicates are unanticipated deviations from the prevailing (and expected) rate of inflation.

Using Fiscal and Monetary Policy to Stabilize the Economy

The adjustment of the aggregate price level when output differs from potential suggests that the economy has a natural tendency to return to a situation in which resources are fully employed. Such an adjustment can, however, take a year or more to significantly affect the economy, a fact that leads many economists to argue that fiscal or monetary policy measures should be used to help speed up the adjustment process. Nonetheless, the use of activist policies remains controversial, and a significant number of economists believe that such interventions are generally counterproductive. We will begin here by describing how government policy affects the short-run equilibrium in the economy, and then we will discuss arguments for and against intervention.

As we have seen, total expenditures in the economy are described by the equation $Y = C + I + G + NX$. The term G stands for government purchases of goods and services. If shifts in household and business spending and net exports reduce expenditures, increased government spending can be used to make up the shortfall in aggregate demand. Increased government spending, or expansionary **fiscal policy**, is one form of intervention that should offset a recession and restore full employment.

Fiscal policy can also be used to indirectly increase spending through a tax cut. Lower taxes (with a constant level of government spending) mean that consumers have a higher level of disposable income. Higher income should encourage increased consumer spending and cause the AD curve to shift to the right, thus mitigating the effects of a recession.

In addition to fiscal policy, the Federal Reserve can use monetary policy instruments to offset short-run economic fluctuations. By varying the amount of money it supplies to the economy, the Federal Reserve can control the interest rate. And, as we have seen, changes in the interest rate can affect the level of both investment and consumption spending. If the economy is producing above potential output, a situation that would cause inflationary pressures, the Federal Reserve can help to reduce consumption and investment spending by decreasing the money supply and causing interest rates to rise. Conversely, if the economy is in recession, increasing the money supply will lower interest rates and stimulate additional consumption and investment spending.

The main argument in favor of using monetary or fiscal policy to stabilize the economy is that deviations of actual output from potential output are costly. In recessions, when some resources are not fully employed, the economy forever loses the output that these resources could have produced. Moreover, unemployment imposes significant hardships on those who lose their jobs or see their incomes reduced. When output is above potential, inflation will accelerate. We have seen that inflation is costly for a variety of reasons.

Controversy about the desirability of fiscal and monetary policy interventions arises for two reasons. The first is the difficulty of identifying precisely what the economy's potential output is and thus the difficulty in determining when interventions are needed. The second and more significant concern centers on the practicality of carrying out such fiscal and monetary policy effectively.

One of the biggest challenges that economic policymakers face is that information about the aggregate economy takes time to collect. It takes about three months to calculate the first estimates of GDP, and these estimates are subject to substantial revision over the next few months as additional data becomes available. Other data are available more quickly, but almost all economic information has some lags, meaning that policymakers must act on partial and incomplete evidence about the state of the economy.

Moreover, the effects of their actions take time to be felt. When interest rates are reduced, for example, it can take many months for businesses to undertake new investment projects since they often require considerable planning. Efforts to increase government spending operate with even longer lags. It can easily take six months or a year from the time Congress authorizes additional spending until projects are actually undertaken. So, even if Congress acts quickly, which is not usually the case, the additional spending may not begin to take effect until the economy has already begun to recover.

If the effects of increased government spending begin to be felt only after the economy has begun to recover on its own, they may cause the economy to overshoot full employment and contribute to inflationary pressures rather than mitigating the effects of the recession. For this reason, many economists believe that activist policies are as likely to be counterproductive as to be helpful.

SECTION III SUMMARY

- Macroeconomics is concerned with two questions: (1) What determines the long-run growth in the size of economies? (2) What are the causes and consequences of short-run fluctuations in the level of economic activity, employment, and inflation?
- Economists measure the total output of the economy using Gross Domestic Product (GDP). GDP is the market value of all final goods and services produced within a country during a specified period of time.
- In the United States, output has grown much faster than population. Since 1900, the U.S. population has increased by a factor of about four, while GDP has grown by a factor of almost forty.
- The rate of growth of output is quite variable. A period between a trough and a peak in economic activity is called an expansion; a period between a peak and a trough in economic activity is called a recession.
- The alternation of periods of expansion and recession is referred to as the business cycle.
- The labor force is the total of all individuals who are either working or are available for work but are not currently working. The unemployment rate is the percentage of the labor force who would like to work but cannot find jobs.
- Economists often break down unemployment into frictional unemployment, structural unemployment, and cyclical unemployment.
- Inflation occurs when prices in the economy are all increasing. The Consumer Price Index and the Gross Domestic Product Deflator provide two different measures of inflation.
- Gross Domestic Product is defined as a measure of production; but at the level of the economy, production equals expenditures equals income.
- Economists divide expenditures into four categories: Consumption, Investment, Government Purchases of Goods and Services, and Net Exports.
- The quantity of GDP per capita that an economy produces is closely related to the level of average labor productivity. Labor productivity depends on many things, the most important of which are the quantities of physical and human capital an economy has accumulated, its natural resource supplies, the level of technological knowledge, and the political and legal environment.
- Economists use the term “savings” to describe income that is not spent on the consumption of goods and services in the current period. “Investment” is the term used to describe the purchase of new capital equipment.
- Financial markets are the institutions through which individuals who have money they wish to save can supply these funds to persons or companies who wish to borrow money to invest.
- Because of the way they are defined, savings

must equal investment in a closed economy. In an open economy, savings equals investment plus net capital outflows.

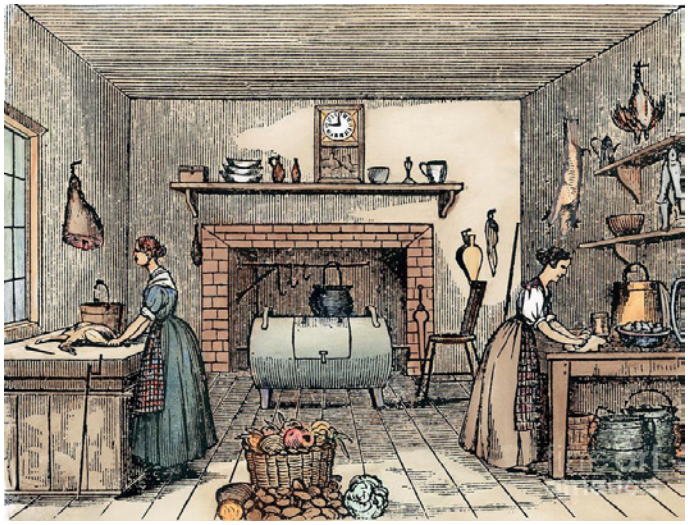
- In the financial markets, the interest rate adjusts to equate the supply of saving to the demand for saving (investment).
- Money is any asset that serves the functions of: (1) a medium of exchange, (2) a unit of account, and (3) a store of value. Because it is not easy to draw an absolute distinction between assets that are and are not money, economists use several different measures of money. The most common are M1 and M2.
- The Federal Reserve System is the central bank of the United States. It was established in 1913 and consists of twelve district banks located in major cities across the country and the Federal Reserve Board, which is located in Washington, D.C.
- The Federal Reserve controls the supply of money in the economy and acts as lender of last resort for the banking system.

- In the long run, increases in the supply of money do not affect the real economy, but affect only prices. But, in the short run, changes in the supply of money alter credit conditions and influence the level of economic activity.
- To analyze short-run variations in the level of economic activity, economists divide actual output into two parts: potential output and the output gap. Potential output is the quantity of goods and services that would be produced if all resources were fully employed. The output gap is the difference between actual output and potential output.
- In the long run, an economy's output is determined by its potential output. But, in the short run, many firms set prices and sell as much or as little as is demanded. As a result, output is determined by the level of aggregate demand, which may be more or less than potential output.
- Deviations of actual output from potential output eventually cause the aggregate price level to change so that the economy returns to potential output.
- When actual output deviates from potential output, monetary and fiscal policy tools can be used to help speed up the adjustment process. In practice, however, changes in government spending or the money supply affect the economy with long and variable lags. Consequently, attempts to stabilize the economy may actually magnify economic fluctuations.

SECTION IV
The Economics of Technology and Innovation

LIFE IN THE PREINDUSTRIAL WORLD

You may not think of yourself as rich. Living in the modern world, it is easy to take its comforts for granted: fresh food in the refrigerator, a closet full of clothes, central heating, running water, toilets. But most people in preindustrial societies lived in what today is considered extreme poverty: in 1820, almost 90 percent of the world had an income of less than $1.90/day.[17] Such a number is difficult to comprehend. To give it meaning, we have to understand what daily life used to be like. We can summarize the standard of living in three themes: material hardship; isolation; and disease and disaster.



Keeping a home in preindustrial times was a full-time, labor-intensive job.

Material Hardship

Work

In preindustrial societies, most jobs were manual labor. In European countries in 1500, 50–75 percent of the workforce were farmers.[18] In the United States, even as late as 1870, 46 percent of the workforce were farmers and farm laborers, 34 percent worked blue-collar jobs, and 8 percent worked in domestic service. Only 8 percent of workers worked in jobs that required problem-solving or creativity, such as managers or professionals.[19] Retirement was unknown; people worked until they were physically unable to continue, or until they died.

Home

Houses had no electricity or gas, and most had no plumbing for running water or sewage disposal. Water had to be hauled in from the well or pump; there were no toilets, faucets, or bathrooms. Heating and cooking were done on a stove or simply over a fireplace—by burning wood, coal, peat, turf, or even animal dung. These fuels also had to be carried into the home and would create smoke that could damage lungs and eyes. Light came from candles that often were made of animal fat; they were expensive, so evenings were dim. There were no washing machines, dishwashers, or vacuum cleaners, making it a full-time, labor-intensive job to keep house even at a meager level.

Food

In the preindustrial agricultural era, diets were not varied, and the food was generally not fresh. There was no way to refrigerate or freeze food; instead, it was preserved through drying, salting, smoking, pickling, or turning into preserves—all of which changed its flavor. Fruits that are common today, such as oranges, were rarely seen far from the few places where they grow.

Possessions

Most people only owned a few sets of clothes, some kitchen implements and other tools, and not many other possessions. For example, in the United States in 1800, a clock cost $50—well over $1,000 in today's dollars—and only about 2 percent of white adults could afford one, mostly shopkeepers and other professionals.[20] (Look around your home and count how many clocks your family owns—and don't forget the ones in the car,

The Great Fire of London of 1666, depicted by an unknown painter. Fire was a constant danger in cities largely built of wood.

on the stove and microwave, on your phone, etc.)

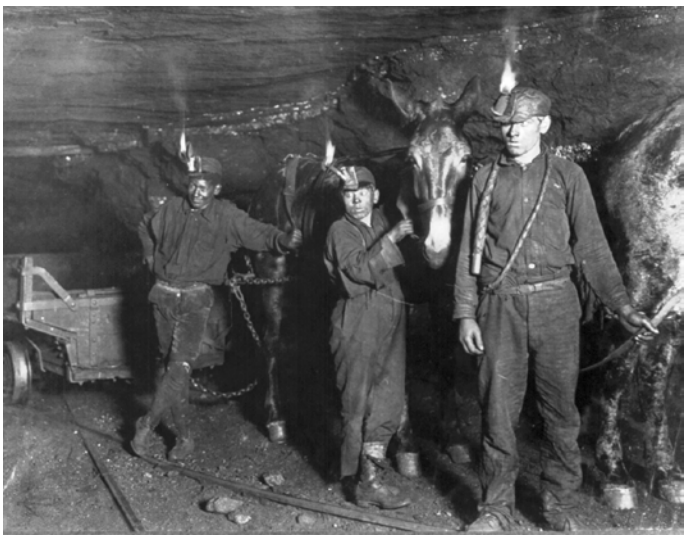
Isolation

Due to limited transportation and communication, people were isolated from each other and also from the vast store of human knowledge and culture. Travel was slow and uncomfortable. News and other information could only travel at the speed of horses and sailing ships. This limited business opportunities. A craftsman or even a manufacturing company would usually sell to their local area and would source materials locally as well. Job mobility was also limited. Socially, the lack of transportation led to limited social circles. Many people never traveled far outside their village. Marriage options were limited, and many marriages were between first or second cousins.

Limited transportation and communication also meant limited knowledge. Most people had only a few years of schooling. In 1800, only 12 percent of the world population was literate.[21] They also had little exposure to other peoples and cultures; only the wealthy could afford to travel abroad. And there were limited entertainment options: no recorded music, no movies, no TV or radio, and definitely no Netflix.

Disease and Disaster

Infectious disease was rampant and caused at least half of all deaths: plague, smallpox, malaria, tuberculosis, cholera, polio, etc. Epidemics have occurred throughout recorded history, such as the Black Plague of the Middle Ages that probably killed around half the population of Europe.[22] Disease struck rich and poor, young and old, even royalty. Children were particularly susceptible, with a quarter dying in their first year, and about half



Children working in a coal mine, 1908. In the United States and other industrialized nations, children now go to school instead of having to work.

dying before reaching adulthood.[23] Thus, most parents knew the loss of a young child, and most children felt the loss of a sister or brother.

Famine was also common throughout the ancient and medieval world, caused by drought, floods, frost, blight, or pests. During famines, people went hungry; ate roots, weeds, and bark; often starved to death; and even sold their children or themselves into slavery to escape.

Fire was a constant danger in cities largely built of wood, especially when open flames were used for lighting, cooking, and crafts. Virtually every major city has burned multiple times, including ancient Rome, medieval London, colonial Boston, and modern San Francisco and Tokyo.

Even a personal injury was more dangerous when there was no 911 to call (indeed, no phone to call with), and no ambulance to rush to the scene.

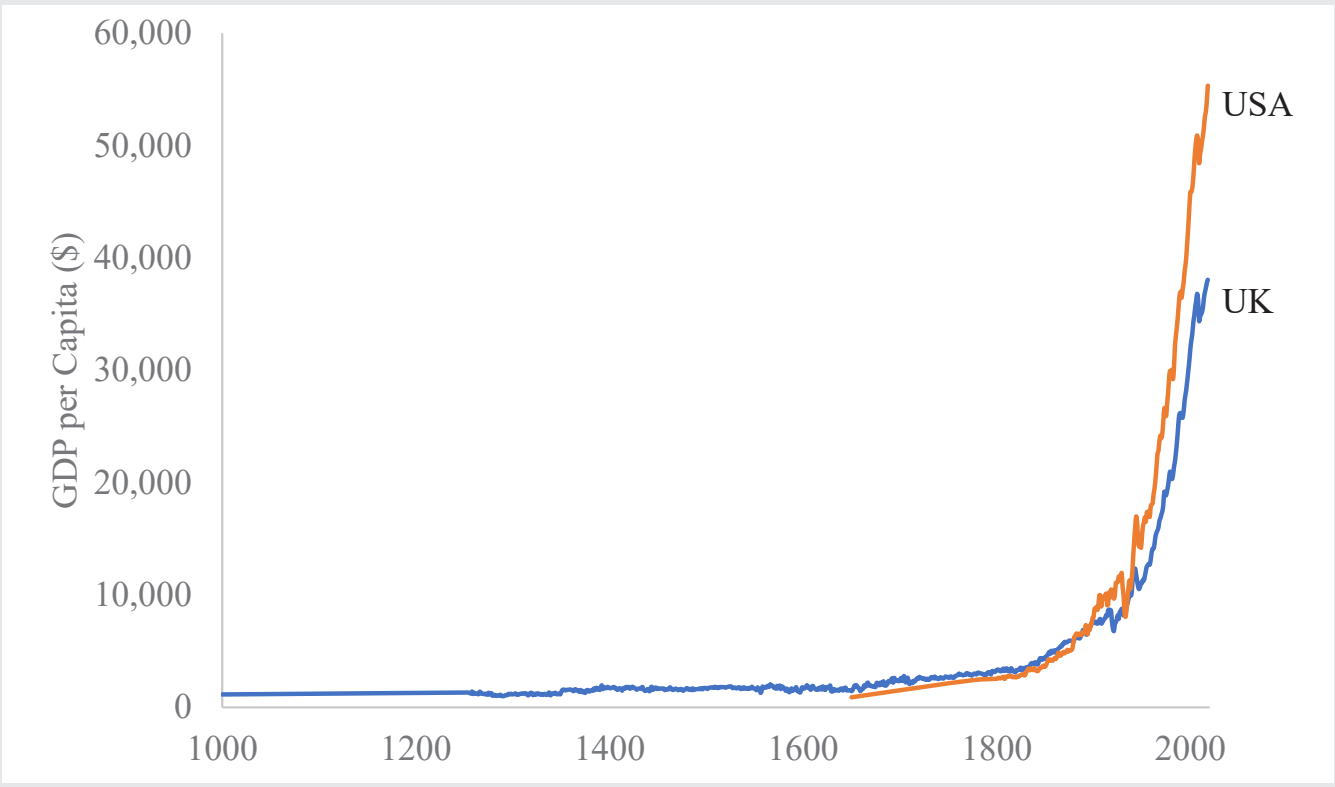
DRAMATIC IMPROVEMENTS IN THE LAST TWO HUNDRED YEARS

The improvements in living standards over the last couple centuries are so dramatic that economic historian Deirdre McCloskey has termed this transformation "the Great Enrichment."

Wealth

GDP per capita has increased some twenty to thirty times in industrialized nations,[24] representing all the comforts many of us can now afford: bigger homes

FIGURE 59



GDP per Capita, 1000 CE–2018 CE

Source: Our World in Data

with central heating and air conditioning, running water for toilets and bathrooms, gas and electricity; cars and trucks; closets full of clothing; and electronics such as smartphones, laptops, and enormous flat-screen televisions.

We work fewer hours—from an average of over sixty hours per week in the United States and many European countries in 1870 to thirty-five to forty hours per week today[25]—and often at more comfortable and enjoyable jobs. Children no longer have to work—they go to school instead. The elderly can retire. Many jobs offer paid vacations.

Connectedness

In industrialized nations, we travel in comfort and safety and can get almost anywhere in the world in about twenty-four hours. We fly home for the holidays and take international vacations. We can even have long-distance romantic relationships. And we can communicate with almost anyone, anywhere on Earth, anytime, instantly—for business, commerce, politics, or socializing.

Health and Safety

Mortality rates for infectious disease have declined by more than 90 percent.[26] Infant and child mortality are now quite rare; the death of a child is a rare tragedy rather than a common occurrence. Overall, world average life expectancy at birth has more than doubled, from under thirty years in 1800 to over seventy years today.[27]

Costs and Risks of Progress

Progress comes with costs and risks. Burning coal and gasoline creates air pollution, which created terrible air quality in nineteenth-century Pittsburgh and 1970s Los Angeles as well as present-day Beijing. And the CO_2 generated is now affecting our climate—potentially a multi-trillion-dollar problem. Car crashes kill more than one million people per year.[28] Chemicals and radiation can cause health risks. Agricultural plenty

Members of the Highland Park Optimist Club wear smog-gas masks at a banquet, Los Angeles, c.1954. The burning of coal and gasoline has had a significant detrimental impact on air quality.

eliminated famine in much of the world, but it has enabled obesity. Technology can be used for evil, such as war or oppression.

But despite all of this, the overall quality of life has vastly improved for the majority of people on the planet.

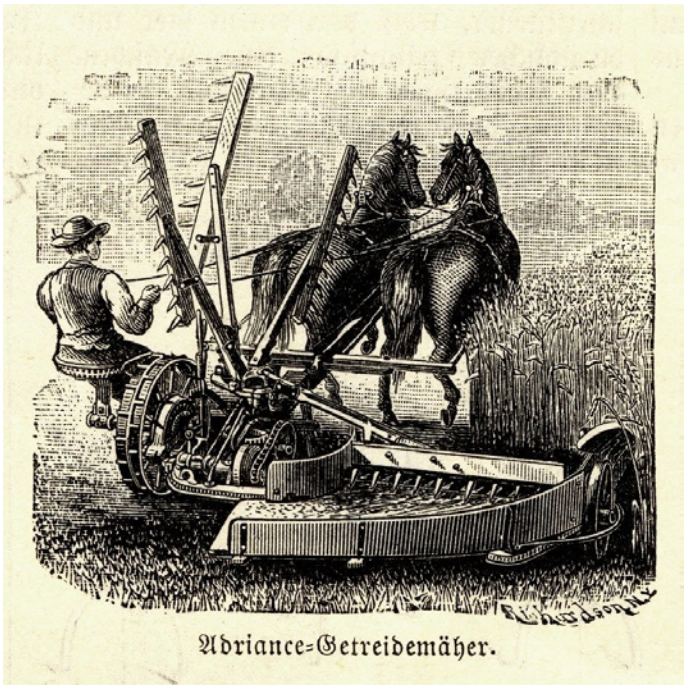
HOW DID THE "GREAT ENRICHMENT" HAPPEN?

Such a transformation is nothing short of awe-inspiring. Anyone who cares about human well-being ought to be interested in understanding how the Great Enrichment came about.

Following is a capsule summary of some of the key developments that created industrial civilization and modern living standards.

Mechanization

We built machines to automate human labor, especially in factories and on farms.[29] Often these machines allowed one person to do the work of ten or more. In one early example, a set of machines was created for the British navy to manufacture pulley blocks—a crucial component of sailing ships that holds the pulleys that the ships' many lines go over and through. The new pulley block factory, staffed by only ten workers, could produce more than one hundred workers could make using only hand tools.



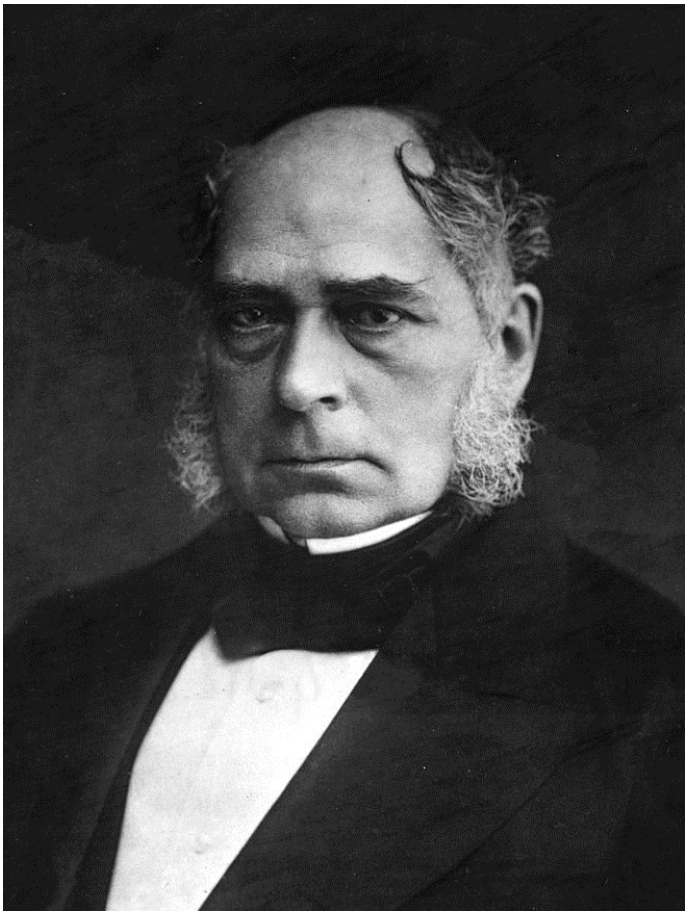
A nineteenth-century reaping machine. In the mid-1800s, a reaping machine was developed that allowed one person (and two horses) to reap as fast as ten men with scythes.

For another example, consider the task of reaping wheat: cutting down ripe stalks of wheat at harvest time, which had traditionally been done by hand-held blades such as scythes. In the mid-1800s, a reaping machine was developed that allowed one person (and two horses) to reap as fast as ten men with scythes. By the mid-1900s, gasoline-powered harvester machines could reap as fast as *one hundred* men. In addition, these machines would further process the wheat to extract the grain, another labor-intensive task.

Mechanization lowered the cost of producing goods, but it could also improve their quality since machines could perform even better and more consistently than humans. For instance, some of the first thread-spinning machines not only saved human spinning labor, but were also able to spin very fine thread, which was used for the highest quality fabrics.

Materials

We also developed better materials and improved processes to make them. Consider steel: the best grade of iron, which has great strength without being brittle. It is an alloy of iron with just the right proportion of carbon. Through the early 1800s, the only processes to make steel were labor-intensive, taking weeks of heating in special ovens or crucibles. In the



English inventor Henry Bessemer discovered a far more efficient way to make steel by blasting the molten metal with air.

mid-1800s, the English inventor Henry Bessemer (1813−98) discovered a way to make steel by blasting the molten metal with air—reducing the process to about twenty minutes. Over the coming decades, through trial and error and applied chemistry, both the cost of steel and the quality and reliability of the product were improved. Today, thanks to the use of modern chemistry and computer-controlled processes, the composition of metal alloys is accurate to the hundredths of a percent.

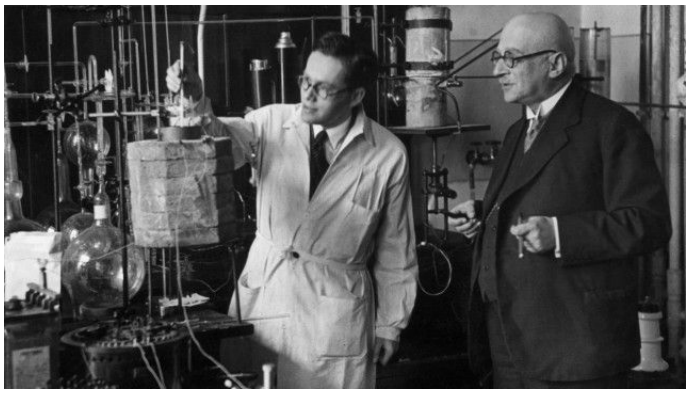
Another crucial material is cement, which is used for the construction of everything from homes, offices, and factories to roads, bridges, and airport runways to dams, levees, and harbors. Its main ingredient is crushed limestone that has been kilned at high temperatures to transform it from calcium carbonate ($CaCO_3$) to calcium oxide (CaO). But the cement can be improved by mixing in silicates, such as clay or volcanic ash, which can make the material harder or allow it to set underwater. During the **Industrial Revolution**, builders began to more systematically investigate and experiment with different formulas for cement, and by the mid-1800s, a new formula (known as Portland cement) was discovered that was stronger and quicker-setting than anything previously known.

We have also created entirely new, synthetic materials, such as plastic. Before plastic, when people needed a lightweight, waterproof, insulating material, they often used animal parts, such as elephant ivory, tortoise shells, or whale bones. These were used for everything from combs to knife handles to billiard balls to electrical insulation. But by the 1800s, animal sources were running out, and some were on the brink of extinction. Plastic came to the rescue: a synthetic material made from more abundant mineral sources. Plastic was lightweight, waterproof, and insulating. It could be made into any shape, rigid or flexible, as well as any color or even transparent. And it was extremely cheap. Plastic quickly replaced expensive leather and silk for clothing, brittle glass jars for food, and heavy wood and metal for children's toys. It has also greatly improved hygiene in food packaging and medical products, a boon to health.

Agriculture

Agriculture was improved not only through mechanization, but through better chemistry and biology. One of the central challenges of agriculture is soil fertility: nutrients in the soil are depleted with each harvest, and natural processes do not replenish them fast enough. In ancient and medieval times, farmers left fields "fallow" (without crops) part of the time, planted special crops, such as legumes, which replenish nutrients, and used manure as fertilizer. By the 1800s, higher yields were achieved from natural fertilizers sourced from around the world, such as rock salts and guano (seabird droppings) from South America, but these sources were rapidly running out. Chemist Fritz Haber and industrialist Carl Bosch came up with a process to combine nitrogen gas with hydrogen gas to create ammonia (NH_3), a precursor of fertilizer. It has been estimated that fertilizer from the Haber-Bosch process is now responsible for about half of all food produced on Earth.

Agriculture was also improved by creating better varieties of crops. Each crop variety flourishes under certain growing conditions, including temperature, sunlight, rain, and length of the growing season.



Fritz Haber and Carl Bosch developed a process to create ammonia (NH_3), a precursor of fertilizer. It has been estimated that fertilizer from the Haber-Bosch process accounts for about half of all food produced on Earth.

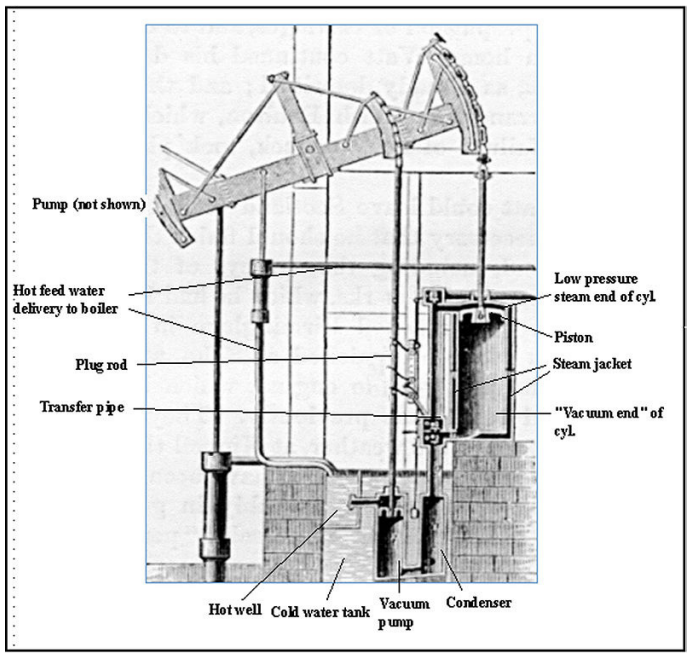
Breeding new varieties allows cropland to be extended into new regions, as for example wheat was extended throughout the Midwest United States. Breeding can also create varieties that are easier to harvest; resistant to drought, pests, or disease; or higher-yielding. One of the greatest successes of increasing yields through breeding was by crossing lines of corn to create hybrids. Hybrid corn helped yields in the United States to soar from under two tons per hectare before 1940 to over ten tons per hectare today.[30]

Energy

Energy was, and is, fundamental to the economy. Engines and motors power factories, agricultural machinery, construction equipment, and home appliances. Heat and electricity are required for industrial processes, and lighting brightens our homes, offices, factories, and streets. Before the Industrial Revolution, mechanical energy came mainly from water, wind, and muscles (both animal and human). Wind and water mills powered the grinding of grain, saws that cut timber, and bellows and hammers for ironmaking. Muscles were used for everything else: for instance, horses were harnessed to turn large wheels to thresh wheat or pump water. The problem with these sources was the difficulty of storing and transporting the energy and of concentrating large amounts of it in one place. Water wheels had to be built on the river; the wind did not always blow; and horses were limited in power and expensive to feed.

The Steam Engine

One of the most important breakthroughs of the Industrial Revolution was the *steam engine*. This was a way to create mechanical power from a different



A schematic of James Watts' steam engine from an 1878 book. The steam engine was one of the most important breakthroughs of the Industrial Revolution.

source of energy: heat from fire, fueled by wood or coal. Fuel could be stored, so it could be collected now and used later. It could be transported, so it could be taken from wherever it happened to be found and brought to where it was needed, such as a mine. The steam engine uses fire to heat water into steam in a boiler. The first steam engines, such as Thomas Newcomen's engine of 1712, condensed the steam in the cylinder to create a vacuum, allowing external atmospheric pressure to drive the piston; later engines achieved higher power by creating high-pressure steam and using that to drive the piston directly. Steam engines were first applied to pump water out of mines, but by the end of the 1700s, they were driving all manner of factory machinery, and by the early 1800s, they were powering locomotives.

The Oil Industry

The next major breakthrough was the development of the oil industry, beginning in the United States in the 1860s. Whereas wood and coal are used basically in the form in which they are found in nature, crude oil was refined into purified products that were better for lighting and for powering engines. Kerosene, the first major refined oil product, was excellent for lighting lamps. Kerosene, and later gasoline, also provided the fuel for the internal combustion engine. Compared to the steam engine, this type of engine can be made smaller and lighter for a given power output, making it

ideal for powering all kinds of vehicles.

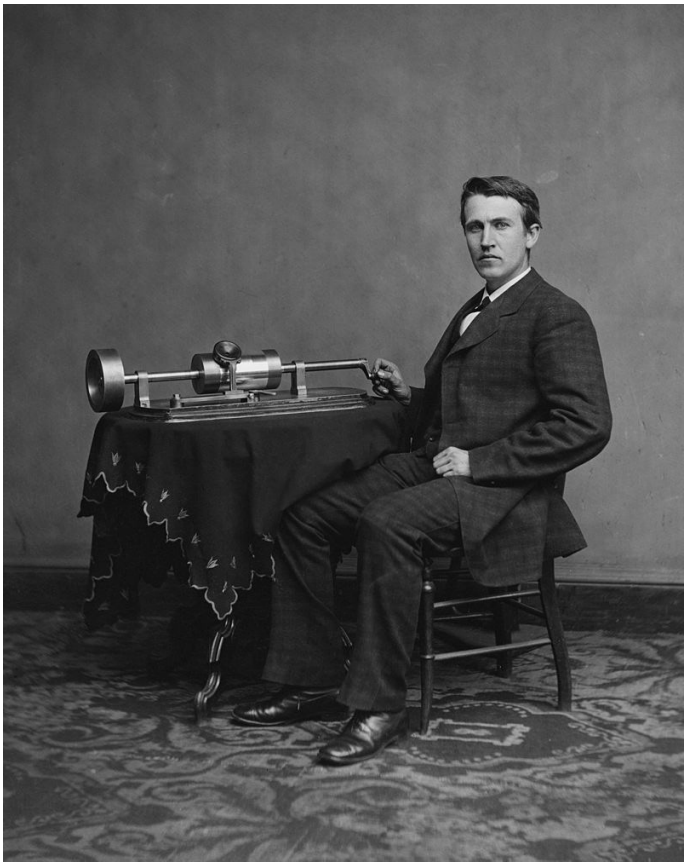
Electricity

The third major breakthrough was electricity. The electric generator allowed this new form of energy to be generated from engines or from traditional sources like wind and water. The American inventor Thomas Edison first demonstrated his incandescent light bulb in 1879. Edison's electric light bulb provided bright light without flickering, soot, smell, or risk of fire. Electric motors provided a clean, quiet source of power to run home appliances. Electricity also revolutionized factories: electric machines allowed the factory floor to be reorganized for maximum efficiency, and electric lighting increased comfort and safety. Further, the electric power grid enabled instantaneous long-distance *transmission* of energy. A large waterfall like Niagara could now power not only local mills, but also lighting and streetcars in cities miles away.

Transportation

Engines also revolutionized transportation. High-pressure steam engines enabled the invention of the locomotive. Starting in the 1830s, a network of railroads was built out across the United Kingdom, then the United States, and soon the world. Railroads cut down the time of a land voyage tremendously. In the mid-1800s, it took six months to travel from New York to San Francisco, and travelers faced deadly perils. If they went by land, they were hundreds of miles from civilization and needed to provide all their own food, water, and medical supplies; if by sea, they faced storms, shipwrecks, and all the other perils of the ocean. After the completion of the transcontinental railroad in 1869, the journey took closer to six *days*, in comfort and safety. Railroads also greatly reduced the cost of freight: to move a ton of goods a thousand miles by wagon in 1830 cost $173.82; to move them the same distance by rail in 1910 cost only $22.43 (both figures are given in 2008 dollars).[31]

In the 1900s, the internal combustion engine enabled the automobile and the airplane. The automobile liberated farm families from their rural isolation, allowing them to go into town to find better opportunities for shopping and socializing. It let city families visit the countryside on weekends; the Sunday drive became an institution. And it helped to spread a new, in-between form of living: the suburb; families could now have space for lawns and play places for children while staying close to city jobs and shopping.



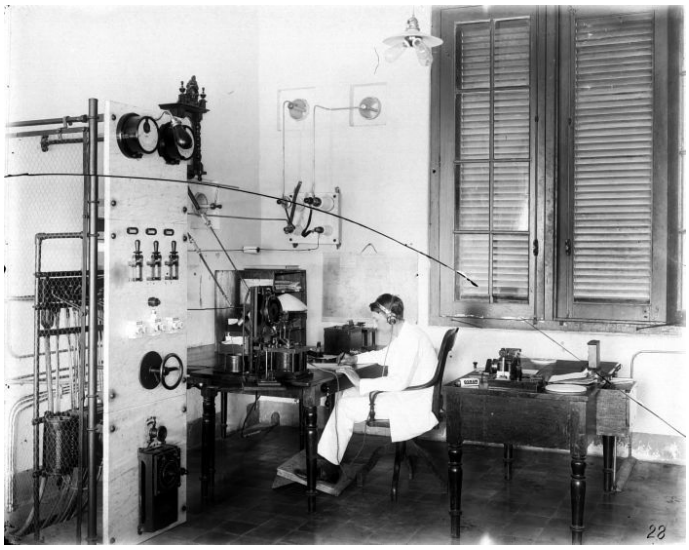
The inventor Thomas Edison, photographed with his phonograph in 1878. Edison first demonstrated his incandescent light bulb a year later, in 1879.

The airplane similarly broadened people's horizons. From the first wood-and-canvas contraption flown by the Wright brothers in 1903, the airplane rapidly developed into the passenger plane and, by the 1960s, the jet. Air travel enabled a variety of activities from foreign vacations to international business deals to national sports leagues to world tours by famous musicians.

Powered transportation has also created a worldwide shipping network. In the mid-1900s, the standardized metal shipping container created huge efficiencies in loading and unloading; today these containers are carried on trucks, by the hundreds on freight trains, and by the tens of thousands on enormous cargo ships. This has created truly global markets and supply chains.

Communications

The science of electromagnetism not only gave us electric lights and motors, but also a revolution in communications. Before the 1800s, there was no way



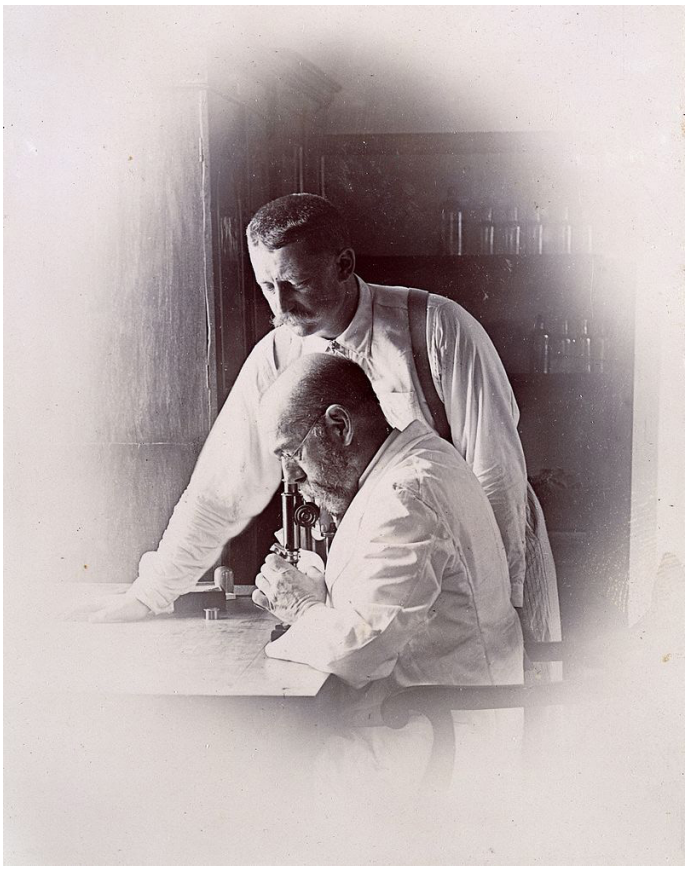
A telegraph operator at work. The later development of the telephone not only enabled voice conversations, but also placed that ability in the hands of ordinary people and didn't require skilled operators as the telegraph had.

By Tropenmuseum, part of the National Museum of World Cultures, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=8615724

to send information faster than writing or printing it on paper and sending it by horse or sailing ship. Even crucial messages could take months to cross a continent or an ocean. Electronic communications traveled at lightning speed. The telegraph, which was developed in the 1830s and 1840s by American inventor Samuel Morse (1791–1872) and others, came first: it could only send brief text messages, but this was enough to revolutionize commerce, finance, and news. The telegraph was followed decades later by the telephone, which not only enabled voice conversations, but also placed that ability in the hands of ordinary people, without requiring skilled operators. The first telephone lines were constructed in the 1870s, and use of the new communications technology grew rapidly.

The next generation of communications was wireless broadcasting over the airwaves: radio (developed in the 1920s) and television (1940s), carrying news, sports, music, and other entertainment. Along with the automobile, these technologies alleviated the isolation felt by rural families: they were now connected to the happenings of the world. They also created a new phenomenon: mass culture, a world unified as never before by witnessing the same events and entertainment at the same time.

The third and current generation of communications is the Internet. Begun in the 1960s under the military



Robert Koch (on the microscope) and his colleague Richard Friedrich Johannes Pfeiffer (standing) investigate a cholera outbreak in India.

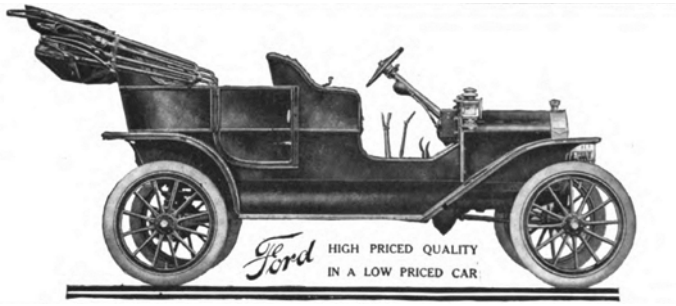
By Wellcome Collection gallery CC-BY-4.0, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=36038902

research institute ARPA (Advanced Research Projects Agency), it reached consumers in the 1990s with the development of the web browser. Today the Internet has transformed entertainment, shopping, media, education, and politics, and some of the largest companies in the United States—including Apple, Microsoft, Google, and Amazon—all depend on it.

The Germ Theory of Disease

The development of the germ theory of disease in the mid to late 1800s by scientists such as the French scientist Louis Pasteur (1822–95) and the German microbiologist Robert Koch (1843–1910) gave us a powerful new way to fight disease, and thereby lower mortality rates, especially for infants and children. Beginning in the late 1800s and early 1900s, we set up three lines of defense against infectious disease.

First, we killed germs in the environment, or prevented them from entering the body. Water filtration and chlorination, and better handling of sewage, eliminated



A 1908 advertisement for a Model T. The automobile was one of several key inventions of the second Industrial Revolution.

germs in our water. Refrigeration and better food handling practices eliminated them in our food and drink. Control of pests, especially mosquitoes, eliminated the diseases borne by them, such as malaria. And antiseptic practices in hospitals made surgery and childbirth much safer.

Second, we gave ourselves and our children immunity through vaccines. Vaccines were essential to the elimination of smallpox and polio and have greatly reduced the incidence or mortality of diseases such as measles, the flu, and covid.

Finally, for when disease does strike, we created effective treatments. The biggest achievement in disease treatment was the invention of antibiotics, such as penicillin, which can fight bacterial diseases such as pneumonia and tuberculosis.

These three lines of defense have reduced infectious disease mortality in the United States by more than 90 percent since 1900.[32]

ORIGINS AND DIFFUSION OF TECHNOLOGY-DRIVEN GROWTH

These developments were not created all at once, and they were not implemented in every country or region of the world at the same time. The first phase of technology-driven economic growth, sometimes called the "first Industrial Revolution," began in Britain in the mid-1700s and later spread to the United States, France, and Germany. Major inventions of this phase included steam engines, textile machinery, horse-drawn agricultural machinery, and the locomotive. The inventions of this era were based on only a modest amount of scientific theory, and many were purely mechanical devices.

Workers at a computer chip manufacturing facility. Workers making computer chips today are more productive than workers making them in 1970 because today's chips are more powerful and therefore more valuable.

By the late 1800s, however, science had advanced to the point where it was providing the foundation for technological breakthroughs—the "second Industrial Revolution." Electromagnetism was the foundation for electric power and for electronic communications; advances in chemistry allowed for improved metallurgy, refined petroleum, new materials such as plastics, and synthetic products such as fertilizer; and discoveries in microbiology led to improved sanitation, new vaccines, and antibiotics. Key inventions of this era included the electric light bulb, motor, and generator; the internal combustion engine, automobile, and airplane; the telephone, radio, and television; and the system of mass production, including the assembly line. During this time, industrial leadership passed from Britain to the United States, especially after the devastation of the World Wars.

From the late 1900s to the present, the greatest pace of technological progress has been seen in digital technologies: computers and the Internet. This is sometimes called the "third Industrial Revolution." Also in these recent decades, the higher standard of living made possible by industrial development has begun to be achieved by many societies around the world, including Japan, South Korea, Taiwan, and China.

KEY CONCEPTS IN ECONOMIC GROWTH

The capsule history of technological and industrial development that follows illustrates some key economic concepts.

Technology and Technological Progress

Before we go any further, we have to answer a key preliminary question: what is technology anyway? For economists, "technology" has a specific technical meaning: a technology is a way of transforming one set of inputs into outputs. For example, our technology for agriculture tells us how to convert inputs like labor, soil, seeds, water, fertilizer, tractors, and so on into outputs like corn or wheat. Technological progress can come in at least two flavors: process and product innovation.

Process innovations consist of the discovery of new and more efficient ways to transform inputs into outputs. For example, Henry Bessemer discovered a way to make steel that required far less time-intensive human labor. Or process innovation may allow us to improve the quality of outputs, without requiring more inputs. Following Bessemer, the quality of steel was steadily improved with further new methods. **Product innovations**, in contrast, are discoveries of ways to transform inputs into entirely new outputs, such as during the creation of new synthetic materials like plastic.

Labor Productivity and Capital

Because, for economists, technology is about how inputs are transformed into outputs, a key measure of technological progress is how much output we can get from inputs. **Labor productivity** is the amount of output produced per worker or per worker-hour. Since people can ultimately only be paid out of what is produced, increases in labor productivity are required for increases in per capita income.

Sometimes it is possible to increase labor productivity simply by changing how workers use their existing tools and resources—as, for example, when manufacturing workers are organized in an assembly line. In other cases, labor is directly multiplied by mechanization and automation. A worker using a machine can produce many times more than the same worker using only hand tools. To enable these increases in labor productivity often (though not always) requires **capital equipment**: machines, factories, office buildings, power plants, vehicles, roads, bridges, ports, electric grids, oil pipelines, etc.

Finally, it should be noted that labor productivity is also increased when the quality of the product increases. Workers making computer chips today are more productive than workers making them in 1970 because today's chips are more powerful and therefore more valuable.

Then U.S. President Bill Clinton presents economist Robert Solow with the National Medal of Science in 1999.

Solow and TFP

As we have seen, capital equipment is an important component of labor productivity. This raises the question: is the only difference between a rich economy and a poor economy the amount of capital they have? More generally, can an economy grow rich by investing in more and more capital equipment? After all, more capital makes workers more productive, including more productive at making capital! Surprisingly though, this doesn't work—capital investment alone cannot account for the Great Enrichment.

This was illustrated in the 1950s by the work of the American economist Robert Solow (b. 1924), who developed a model of economic growth that shed light on the drivers of labor productivity. To model the overall economy, Solow wrote a formula for economic output called a *production function*.[33] Solow wrote a production function of the form:

$$Y = AK^{1/3}L^{2/3}$$

FIGURE 60

Capital (*K*)	GDP	GDP per Capita
1,000	100,000	100
2,000	125,992	126
3,000	144,225	144
4,000	158,740	158
5,000	170,997	171
6,000	181,707	182
7,000	191,293	191
8,000	200,000	200

GDP and GDP per Capita when L = 1,000 and A = 100

In this equation, *Y* stands for total economic output, or GDP. *K* and *L* stand for capital and labor. *A* is a multiplier representing technical factors, such as new inventions or processes, or better ways of organizing factories. Today this multiplier is called "**total factor productivity**," or TFP. This model captures some important intuitions about economic production. For example, it indicates that if you doubled all the people and machines used to produce GDP, then you would also double GDP.

To see this, let's imagine $A = 100$, $K = 1000$, and $L = 1000$. If we plug these values into the equation, it tells us GDP will be equal to 100,000. If we then doubled all the workers and all the tools and factories they used to produce output, so that $L = 2{,}000$ and $K = 2{,}000$, then this equation tells us that GDP doubles to 200,000.

However, notice that if we divide GDP by the number of workers, in both cases, we get the same thing: 100,000 divided by 1,000 is 100, and 200,000 divided by 2,000 is also 100. If we want to increase income, we need to increase how much each worker produces, not just how much is produced in total. One way we can do that is by adding capital while keeping the number of workers fixed.

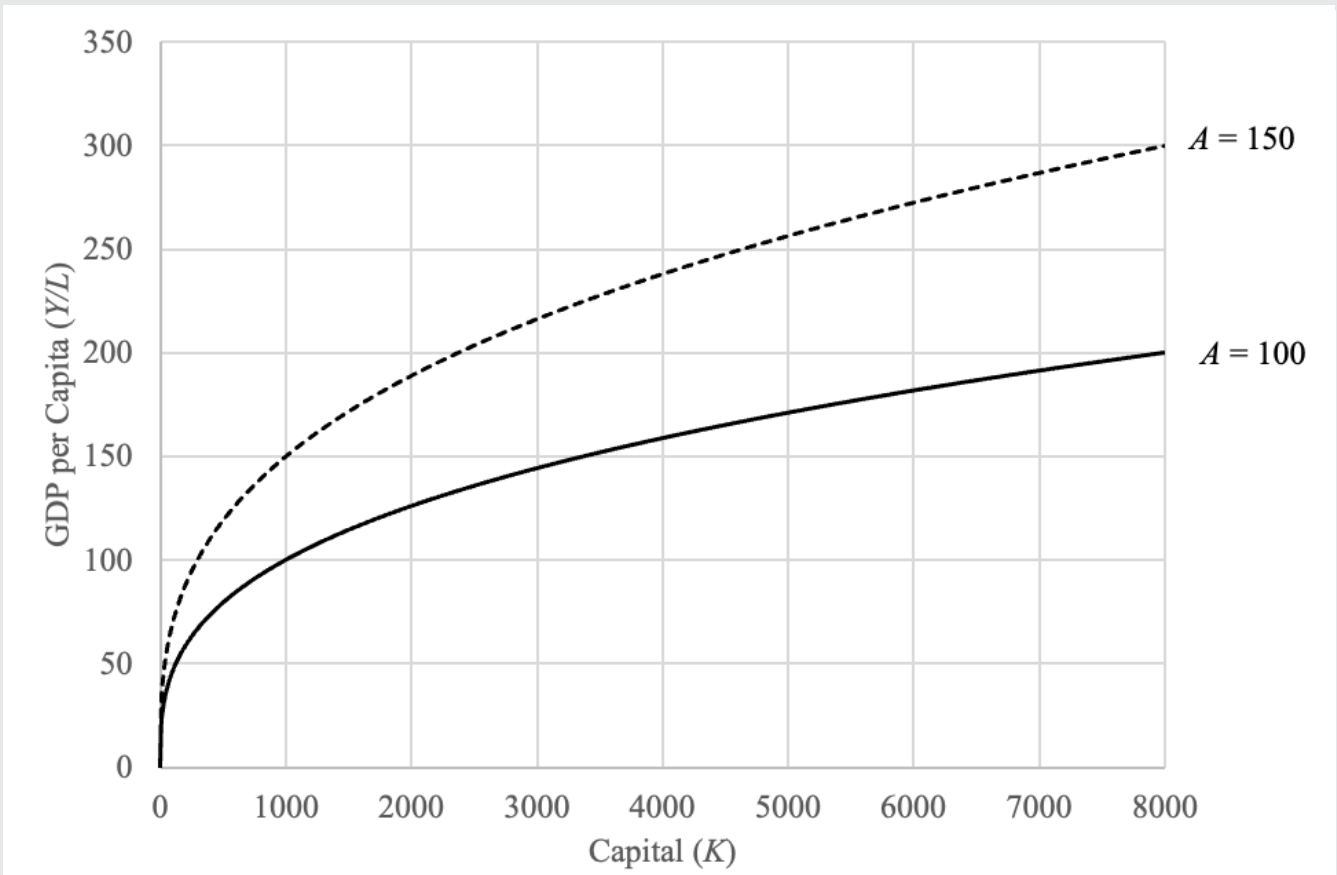
Let's suppose we double capital to 2,000, but we keep $L = 1{,}000$. If we plug this value into the equation, GDP will be 125,992, or 126 per capita. That's more than 100—but notice that doubling the capital did not double GDP per capita. The table shown in Figure 60 illustrates how every additional 1,000 units of capital increases GDP per capita by less and less.

The first additional 1,000 capital increased GDP per capita by 26. The last 1,000 increased it by just 9. Although more capital, in this model, will always increase worker productivity, the diminishing returns to productivity impose limits to this process. At some point, it's not worth adding more capital—especially if capital has ongoing costs, such as maintenance. An important conclusion of Solow's model is that capital accumulation alone cannot sustain long-term increases in per-capita income—at some point, productivity and incomes will plateau.

However, there is another way to increase productivity, and that is by improvements to the technology multiplier, TFP. When TFP increases, it directly increases labor productivity by its multiplier effect on output. Furthermore, by increasing *capital* productivity, it enables more capital to be deployed before hitting diminishing returns. Thus, sustained increases in TFP can create sustained increases in both capital accumulation and labor productivity.

Figure 61 illustrates the two ways a society can increase GDP per capita. In the solid black line, we see that increasing capital (which is measured on the horizontal axis) increases GDP per capita (measured on the vertical axis), as described in Figure 60. This line is bent down, capturing the notion that as we add

FIGURE 61

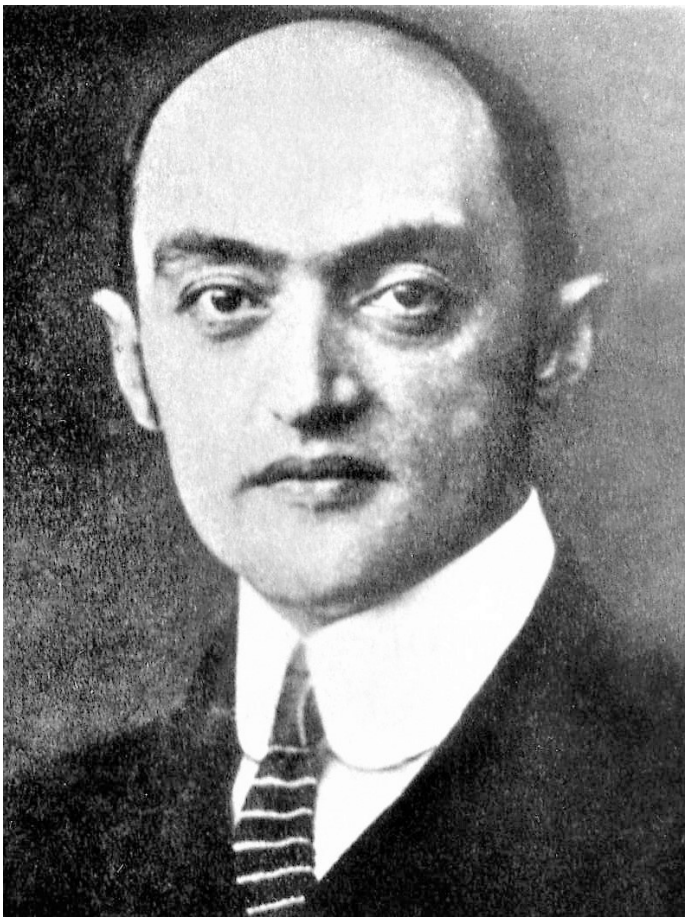


The relationship between capital and GDP per Capita when L = 1,000 and A = 100 or A = 150.

more and more capital, GDP per capita increases by a smaller and smaller amount. The dashed line illustrates the same thing, but when A is 150, instead of 100. Note that for any given value of capital, the dashed line is higher than the solid one, illustrating one way that GDP per capita increases when A increases. But also, when A increases, it becomes worth investing in more capital. For example, when $A = 100$, if we move from $K = 1{,}000$ to $K = 8{,}000$, GDP per capita increases by 100 from 100 to 200. However, when $A = 150$, the same move from $K = 1{,}000$ to $K = 8{,}000$ increases GDP per capita from 150 to 300, which is an increase of 150. An economy with higher values of A finds that investing in capital is more attractive.

Solow went further and tried to see which was the case for understanding differences in GDP per capita across time. From his vantage point in the 1950s, he wondered if GDP per capita growth in the United States was mostly because of steadily increasing capital per worker or growth in TFP. He gathered data on the size of the U.S. labor force and capital stock over time and then tried to statistically see if differences in GDP per capita could be mostly accounted for by differences in the amount of capital per worker or not.[34] Surprisingly, he found very little role for capital! Solow concluded that most of the growth in GDP per capita across time was due to changes in the A of his production function: total factor productivity, which among other things captures differences in technology.

The upshot of Solow's findings is that improvements in technical factors, not capital accumulation, are what drive sustained growth. That is, in the long run, we can have material progress if and only if we have technological progress.



The economist Joseph Schumpeter coined the term "creative destruction" to describe the way in which old methods of doing things are replaced by new and more productive ways.

CREATIVE DESTRUCTION AND TECHNOLOGICAL UNEMPLOYMENT

Even though technological progress is the main driver of the Great Enrichment, the way this happens is not without ups and downs. The economist Joseph Schumpeter (1883–1950) coined the term "**creative destruction**" to describe the way in which old methods of doing things are replaced by new and more productive ways. This means technological progress can be bad for old industries that produce technologies that are rendered obsolete by progress, even if it is (usually) good for society as a whole. For example, while the invention of the personal computer was bad news for the typewriter industry, the availability of computers probably made society better off overall.

A related concept is technological unemployment. Throughout history there have been fears that automation might lead to rising unemployment, as workers with specific skills are replaced by machines that can do their jobs better or at a lower cost. In practice, unemployment rates do not seem to have risen markedly as a consequence of technological progress though.

This is because technological progress creates new jobs, even as it destroys old ones. For example, new technologies usually require new kinds of skills to manufacture and maintain them. But more broadly, technological progress also creates jobs by making society richer, so that people can afford to spend money on things they would not previously have paid someone to do. For example, a richer society is more willing to spend money on fancy coffee and tea, which creates new jobs associated with providing those goods.

While new jobs have been created at a fast enough rate to replace older ones so far, this does not mean we do not have to worry about technological unemployment. The future is under no obligation to resemble the past. Rapid advances in machine learning and artificial intelligence are kindling new worries about technological unemployment. More generally, even if new jobs are created, the workers whose skills are rendered obsolete by technological progress may not have an easy time switching into the new jobs, especially if they require very different skills. Both issues are reasons why there may be an important role for the government to help manage those who lose out from creative destruction—for example by helping workers retrain for new jobs or perhaps even providing people with a universal basic income.

INNOVATION TODAY

Measuring the Rate of Progress

We can try to more formally measure the rate of technological progress using some of the economic concepts introduced earlier in this guide. As discussed in section III of this guide, GDP is a way to measure the size of the economy by adding up the value of all final goods and services produced within an economy during a time period. *Real* GDP is a measure that does this in a way that tries to make the prices for goods and services comparable over time, so that it tracks how *much* is produced, rather than inflation. Between the 1950s and 2020, annual real GDP per capita growth in the United

A mechanic works on a motorcycle. Allowing workers to access the jobs that best use their skills aids in economic growth.

By Seattle Municipal Archives - Seattle Municipal Archives, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=26185091

States averaged 2 percent per year.[35] We now live in a world where economic growth is taken to be the norm.

Not all of that growth is due to better technology though. Some comes from increased investment in capital (though, as just noted, technology can indirectly affect this by making it worth investing in more capital), some from a greater share of the population working, and some from increases in the average education of workers.[36] Another portion comes from matching workers to the jobs that best use their skills—for example, by allowing talented women and minorities to enter professions from which they were previously blocked. These are things we can measure (sometimes with great difficulty), but they do not fully account for the growth we have observed.

The leftover that remains, after we have tried our best to measure all the elements contributing to growth, is usually attributed to new and better technologies.[37] In the end, economists estimate technological progress directly contributed to about 1 percent annual GDP per capita growth per year between the 1950s and 2020.[38]

This part of growth is important because technological progress can continue as long as we can find new things worth inventing, whereas the other drivers of growth will eventually run out. Of course, there is no guarantee that we will always be able to do that. On the one hand, there appears to be no shortage of new inventions every year and no shortage of new ideas to explore. But on the other hand, some economists have pointed to a troubling trend: invention seems to be getting harder. Scientists must spend more and more time in training before they begin to make big discoveries,[39] and more and more R&D (research and development) effort seems to be necessary to maintain the same rates of technological progress.[40]

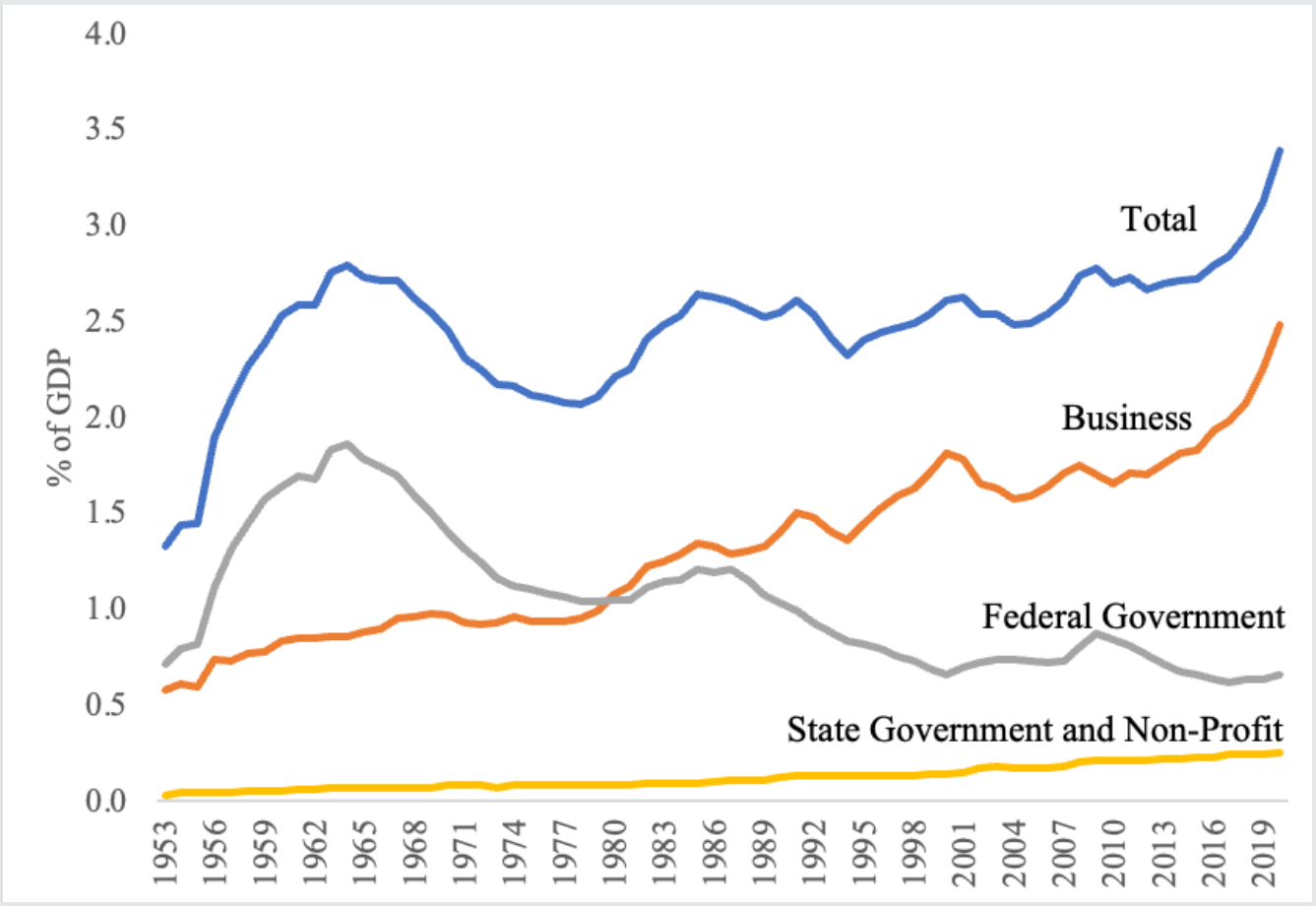
Solow's work showed that differences in technology, across time and places, are the most important reason for differences in GDP per capita, but it had little to say about how and why technological progress happens (or doesn't). It simply assumed technological progress happened, somehow, with its exact processes hidden inside a metaphorical black box. In this section, we will open the box and try to understand the factors that drive technological progress.

Paying for Progress

Creating new and improved technologies is not free. Since the 1950s, an average of about 2.5 percent of GDP is classified as spending on research and development (R&D) in the United States. By research and development, technically we mean activities meant to add to our stock of knowledge about the world (research) or to develop new applications of technology (development), though the two aren't necessarily so easily separated in practice. While 2.5 percent isn't the highest share in the world, it is above average, and since the United States is the largest economy in the world (at current exchange rates), the United States is the single largest funder of R&D in the world. As of 2019, the United States does 27 percent of global R&D. China and Europe each fund 22 percent of global R&D, followed by Japan at 7 percent.[41]

In the United States, R&D funding comes from three

FIGURE 62



Sources of R&D funding, % of GDP

Source: Figure RD-3 of NSF 2022

major sources: private industry, the government, and non-profit organizations. Today, private industry is the source of about 73 percent of the total R&D spending in the United States, followed by the federal government at 20 percent, with state governments, non-profit foundations, and universities rounding things out at 7 percent. Since the 1950s, the share of R&D conducted by the private sector has gradually increased from 44 percent in 1950 to 73 percent in 2020. This has mostly been balanced by a decline in funding from governments, from a high of 67 percent in 1964 to 20 percent today.[42]

Recall that earlier we asserted that from 1950 to 2020, technological progress probably accounted for 1 percent of the annual increase in real GDP per capita. But we have just learned the United States has spent 2.5 percent of GDP per year on R&D. Spending 2.5 percent of GDP per year in order to do R&D that only increases GDP per capita by 1 percent per year might seem like a bad investment. We spend more on R&D in a year than we get back in a year! But this argument makes an error. R&D is a one-time expense that generates new ideas and new technologies that raise income for as long as we use them. When we increase TFP by 1 percent, TFP is 1 percent higher in *every* following year, whether we keep doing R&D or not. In less than three years, the extra GDP enabled by R&D more than pays for itself!

One reason innovation can have such high returns is because new ideas are *non-rival*. A rival good is one that can't be shared between more than one person without diminishing it in some way. Most goods are rival goods. In contrast, non-rival goods are infinitely extendable; multiple people can use them at the

same time. For example, specific copies of most of the technologies that enriched the world—reaping machines, steam engines, and automobiles—are rival, in the sense that if two people want to use the same machine, they can each use it only half the time. But the *ideas* for reaping machines, steam engines, and automobiles are nonrival. Once invented, the idea can be enjoyed equally well by everyone in the world.

The nonrivalry of ideas is one reason innovation has such high returns. In fact, economists estimate on average every dollar of R&D we spend probably produces more than $3.60 in value in the long run, possibly much more.[43] Doing research is like having access to a machine where you put a dollar in, and it gives back several dollars. That presents its own question though: if you had access to such a machine, why would you only put in $2.50 out of every $100?[44] And why did it take so long for economies to begin investing in progress?

APPROPRIATION AND THE VALUE OF R&D

The Challenge of Appropriation

To answer that question, we need to think about the decision to invent. American Economist Paul Romer (b. 1955) won the 2018 Nobel Prize in economics[45] for his research on *endogenous growth theory*, which brought the decision to invent into economic models of growth, instead of merely assuming technological progress just happens. Economists primarily look at the decisions of entrepreneurs and inventors through the lens of selfish profit maximization. We assume they choose to invent if, from a self-interested perspective, it's the most profitable use of their time and resources. This isn't necessarily the whole story, but it's an illuminating place to start.

Viewed as an economic decision, one reason inventors may choose not to invent, even if invention pays off handsomely on average, is that invention is a uniquely risky activity. If an inventor knew an invention would work, it probably wouldn't count as an invention! Instead, they may get a higher-than-average benefit sometimes, but often they will fail to invent anything new that works. So, one reason inventors may fail to invest in R&D projects that offer a good return on average is that they can't be guaranteed the "average" benefits, and they dislike the inherent risk.

Economist Paul Romer won the 2018 Nobel Prize in economics for his research on endogenous growth theory.

By Bengt Nyman from Vaxholm, Sweden - EM1B6039, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=74934767

But there is something else special about invention that makes it less attractive as an economic activity: inventors do not capture most of the value created by their inventions. As we've seen, a lot of the value of innovations comes from the fact that they are non-rival and thus can be used by many people at once. But all that value is hard to capture because ideas are also partially *non-excludable*.

An **excludable good** is a good that you can prevent other people from using. For example, if Alice has a newly invented automobile, her automobile is excludable if she can prevent others from driving it. A non-excludable good is one where it is impossible to prevent people from accessing it. If Alice invents the automobile and starts selling automobiles, then the idea of the automobile is non-excludable because anyone can buy one, take it apart, and then learn how it works.

The challenge with innovation is that new ideas and new technologies are at least partially non-rival and non-excludable. Non-rivalry means a lot of the value of an idea will come from people besides the inventor

Thomas Edison's first successful model of a light bulb, which was used in a public demonstration at Menlo Park, December 1879.

using it. Non-excludability means it is difficult to sell access to the idea. That makes R&D a lot less valuable to an individual than it is to society.

To illustrate, suppose Thomas invests $1 million into R&D and succeeds in inventing a lightbulb that can be manufactured for $10 per lightbulb. Thomas hopes to sell 1 million lightbulbs for $12 each, so that he can make $2 in profit on each lightbulb sold. This will earn him $2 million, which is enough to pay off his $1 million in R&D expenses and keep $1 million in profit.

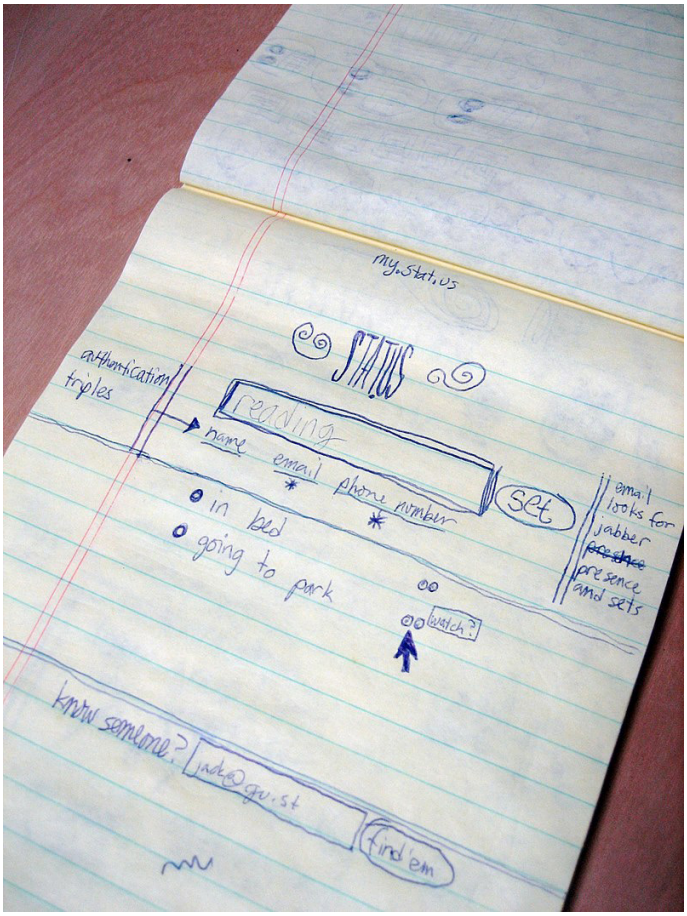
However, let us suppose competitors can purchase Thomas' new lightbulbs and figure out how they work. Now, as soon as he invents a lightbulb, other competitors will figure out how to manufacture them without paying $1 million in R&D and can set up their own manufacturing operations to sell lightbulbs alongside Thomas. When many firms compete in the market to sell the same good, the price will tend to fall to the level of the average cost of producing the good. In this case, the price will converge to $10, the cost of manufacturing a lightbulb. If Thomas tries to sell his lightbulbs at $12, no one will buy them since the exact same design can be had from competitors for $10. But if he sells at $10, he will only ever cover the costs of manufacturing lightbulbs. At a $10 price, he will never earn back his original $1 million in R&D costs, much less turn a profit.

In this case, the investment in the lightbulb was worth it from the perspective of society. If 1 million people would have been willing to pay $12 for a lightbulb, but instead purchased one for $10, then each earns $2 in consumer surplus. If 1 million people each enjoy $2 in consumer surplus, total consumer surplus is $2 million, which is much more than the $1 million cost of R&D. However, in this case, the value of the innovation was not captured by Thomas, but rather by consumers. Thomas, therefore, has no incentive to invest in the R&D in the first place.

We call Thomas' ability to capture some share of the social value of his innovation the **appropriability** of the innovation. As just described, one reason it can be hard to appropriate the value of an innovation is that non-excludability and non-rivalry mean competitors may be able to learn how to produce a new invention without doing any R&D themselves. This competition can lower the price of a new technology below the prices that would allow the original innovator to break even. Fortunately, as we have seen, private companies have found ways to (at least partially) solve this problem. As we have seen, they actually perform the majority of R&D. There are four main ways firms try to appropriate the value of innovation: secrecy, lead time, complementary assets, and intellectual property.

Secrecy

An obvious way to stop competitors from stealing an idea is to keep it secret. This strategy can work quite well for inventions that are difficult to reverse engineer. In our lightbulb example, we assumed competitors could simply purchase the lightbulb and figure out how it works. But this is not possible for all inventions. In many cases, the methods of manufacturing a product are just as important as the product itself: think of Henry Bessemer's new method of producing



A sketch, c. 2006, by Jack Dorsey, co-founder and former CEO of Twitter, Inc., envisioning an SMS-based social network. Social media platforms such as Twitter exemplify network effects as they derive much of their value from their large user base.

By Jack Dorsey - twttr sketch, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=5150577

steel, for example. While it is impossible to prevent rivals from purchasing and carefully studying a new product, it is usually a lot easier to keep them out of the factory where it is produced. In fact, in surveys of manufacturing firms, respondents said secrecy was effective at protecting about half of innovations.[46]

Secrecy has many advantages for a firm. As long as a firm can keep its secrets, it can enjoy a competitive advantage over its rivals. However, secrecy also has limits. Not all inventions can be easily protected by secrecy nor does secrecy prevent a rival from engaging in its own R&D to independently re-invent a technology first discovered elsewhere.

Lead Time

In some cases, it is not possible to keep an idea secret forever, but it still takes time for rival firms to learn how to produce a new technology. During this window, the original inventor can charge a price above manufacturing cost and potentially recoup their R&D expenses. Alternatively, the inventing firm can continue to invest in R&D to improve a product during this time, so that it keeps its lead over rivals and can continue to charge a higher price.

As with secrecy, **lead time** is more effective in some settings than others. Sometimes it is easy for rivals to learn how to produce a new technology. In other cases, it's not easy to figure out how to improve a product and keep a lead over rivals. For example, we don't see companies racing to invent successively better versions of the chair. In either case, an inventor can't rely on lead time to charge a price above its manufacturing costs. In practice, manufacturing firms report lead time is effective at helping them appropriate the value of their inventions about 50 percent of the time for new products.[47]

Complementary Assets and Related Factors

Both secrecy and lead time depend on preventing rival firms from learning about a new invention. Another strategy is to make sure simply learning about the idea isn't enough, so that it doesn't matter if competitors learn how to manufacture the invention. One way to do this is for a firm to have a **complementary asset** that makes the technology more valuable to them alone.

For example, as mentioned earlier, concrete can be improved by mixing in clay or volcanic ash. If you don't have access to any clay or volcanic ash, then knowing how to make concrete with it doesn't help you actually make it. Clay and volcanic ash are complementary assets in this case. But complementary assets can take many other forms. Maybe a firm has a great reputation, so that customers will trust that its inventions are high quality. Maybe it has special skills in manufacturing that are hard to learn. Maybe it has a large sales force. In all these cases, a firm with these assets may be able to charge more for a new technology than rivals who lack them. In surveys, firms report 30 to 50 percent of inventions are well protected by complementary manufacturing, sales, or service assets.[48]

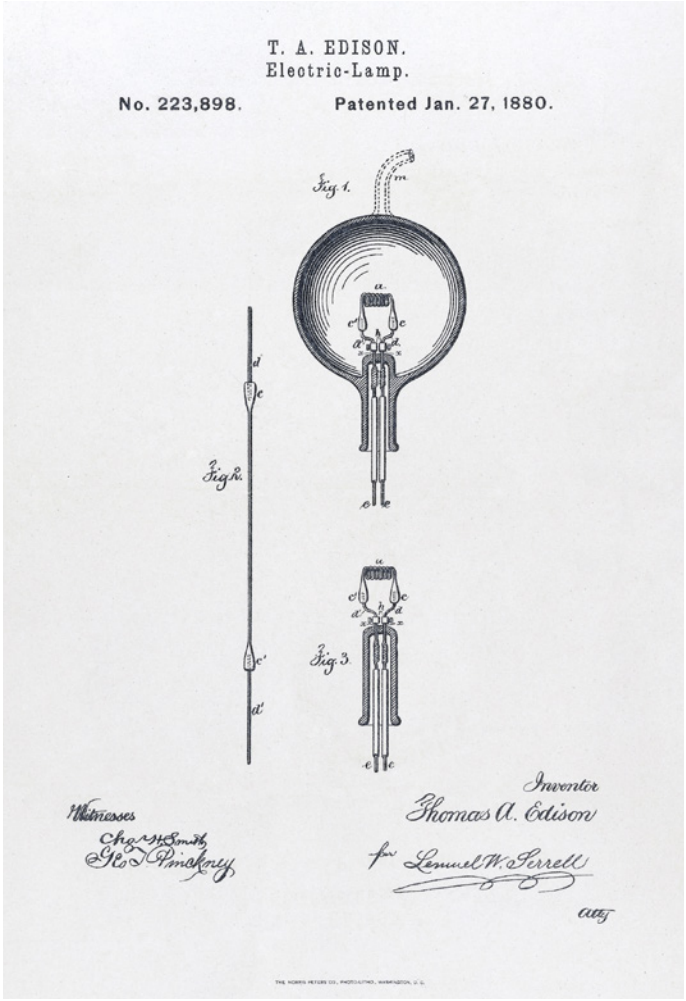
And there might be other reasons it's hard for a new inventor to profit from a freely available idea. One example is **network effects**; these are technologies that

become more valuable the more users they have. Social media platforms, to take one example, derive much of their value from a large user base. Even if a rival could easily reverse engineer the code for something like Facebook or Twitter, the value of these platforms comes from the presence of other users. It is hard to convince users to switch to a new platform if there are no other users with whom to interact! Other technologies derive their value from an ecosystem of related technologies. For example, much of the value of a mobile operating system comes from the apps that run on it, most of which are created by other companies. This also makes it hard for a competitor to copy the platform because they would struggle to attract buyers without all the apps that run on the original platforms. But without users, they also cannot easily convince companies to make apps for their platform!

Another example is **economies of scale**. For many kinds of product, the average cost of producing goods falls as you produce more. For example, because they require such specialized and precise equipment, factories that manufacture the microcircuits inside most modern computing devices cost billions of dollars to build.[49] If a factory costs $1 billion, then if it produces 1 billion microcircuits, it needs to sell each one for $1 to break even. But if it only makes 1 million microcircuits, it will need to sell each one for $1,000 to break even! Economies of scale mean big producers can afford to sell at a lower price than small rivals and still earn a profit. That can be another competitive advantage over copycats; even if the idea is free to copy, it may be too expensive for a copycat to produce at a large enough scale to compete on cost with the originator of an idea.

Intellectual Property

Lastly, in some cases the legal system can make a non-excludable good excludable. This is called **intellectual property** rights. Common forms include patents, trademarks, and copyright. Most relevant for innovation are **patents**, which allow an inventor to decide how their patented invention can be used for a period of up to twenty years. This allows an inventor to charge a price above manufacturing cost for twenty years, if they wish, and to prevent competitors from copying the invention. Alternatively, a patent-holder could decide to allow competitors to produce the invention but require them to pay a fee for the right to do so.


T. A. EDISON.
Electric-Lamp.
No. 223,898. Patented Jan. 27, 1880.
Fig. 1.
Fig. 2.
Fig. 3.
Witnesses
Inventor
Thomas A. Edison
for Lemuel W. Serrell
atty

U.S. Patent #223898: Electric-Lamp, issued January 27, 1880, to Thomas Edison. A patent grants an inventor temporary property rights.

Patents attempt to balance two competing goals. On the one hand, to encourage innovation, it is important to provide ways for companies to appropriate the value of their inventions. But on the other hand, we would like newly invented discoveries to be used as broadly as possible since it is only when an invention is used, that it can provide any social value. As we will discuss in the next section, we would also like other inventors to be able to build on and improve new technologies. A patent manages this tradeoff by granting the inventor *temporary* property rights. While the property rights are in force, the inventor can try to appropriate the value of their invention and recoup R&D costs. But after it expires, the invention is available for anyone to produce or improve on.

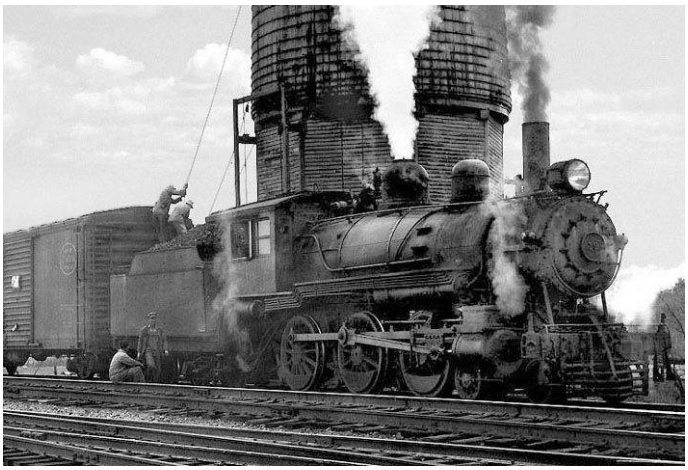
Patents also have a unique advantage over other means of appropriating the value of inventions. Throughout

this discussion, we have conflated two activities that are actually distinct: the invention of some new technology and the building of a successful business around it. The latter activity, which can entail just as much uncertainty and creative problem solving as the original invention, is typically the domain of the entrepreneur, at least for inventions discovered outside of existing businesses. Secrecy, lead time, and complementary assets only work if the inventor is involved in manufacturing and selling their invention. But inventing and running a business can be quite different sets of skills. Few people are truly excellent at both. Patents allow an inventor to transfer the rights over an invention to an entrepreneur or organization that is better able to manufacture and sell the invention. This lets inventors earn a reward for their R&D effort without needing to have the skill to successfully run a business.

Although patents are quite well known, they are actually among the least common methods used to protect innovations. Manufacturing firms say only 20 to 35 percent of their inventions are well protected by patents,[50] though bigger firms are significantly more likely to use patents.[51] One reason patents are not more common is they can be costly, especially for small firms, and enforcing patent rights requires a firm to discover and then sue rivals who infringe on a patent. That can be costly and time-consuming, and there is no guarantee of winning in court. Meanwhile, securing a patent requires giving up an invention's secrets, which might allow rivals to incorporate the ideas into products that compete with you, but which are different enough so as not to infringe on the patent.

OTHER IMPORTANT DRIVERS OF INNOVATION

Even with all these ways of appropriating value, firms usually only capture part of the social value they create when they do R&D. A significant reason for this is that ideas tend to build on each other in unpredictable ways. For example, the historian David Wootton has argued that scientists' invention of the experimental method, where you systematically change just one thing at a time and carefully observe the results, was later adapted by the inventors of the first steam engines to learn exactly what designs worked best. Eventually, steam engines became so efficient and so sturdy that they could be used to run railroads across the world.



A steam locomotive photographed in the 1940s. It's unlikely the inventors of the steam engine, working in the mid-1700s, could have envisioned that steam engines would someday traverse railways across the world, as ideas tend to build on each other in unpredictable ways.

But it's unlikely the scientists working out the best way to understand nature thought their work would one day result in locomotives!

When research conducted by one organization enables another to make new discoveries or invent other technologies, we call this a **knowledge spillover**. Knowledge spillovers are a special kind of positive externality where R&D by one group has positive effects on another. The total value of R&D comes not only from the technologies that are directly invented (and which might be copied), but also from the many unpredictable new technologies that may be invented in the future, and which build on the earlier discoveries.

Knowledge spillovers account for quite a large share of the value of R&D—more than half by some estimates.[52] But because the value of knowledge spillovers is captured by organizations that did not do the original R&D, private companies usually capture only a tiny part of the value they create. Accordingly, companies may very well decide not to invest in some R&D projects, even if the social value outweighs the costs, because most of the value spills over to other companies. The net effect is complicated by the fact that positive spillovers might also encourage *other* firms to do R&D they wouldn't normally do since part of the R&D cost has been absorbed by someone else. However, in general, we may be concerned that the positive externality of knowledge spillovers means the private sector on its own would not do enough R&D, relative to the social optimum. For example, one famous study found

inventors, on average, were only able to appropriate 2.2 percent of the social value of their inventions![53] This is one rationale for the government and other non-profit organizations to also invest in R&D.

Non-Profit Support of R&D

As noted earlier, 27 percent of R&D today is funded by either the government or other non-profit organizations. Since non-profit organizations do not need to maximize profits (by definition!), they do not need to worry about their ability to appropriate the value of innovations when they decide whether to fund R&D. In fact, it usually makes sense for non-profit organizations *not* to try to appropriate the value of their R&D investments. This is because appropriating the value of an invention usually means finding a way to exclude others from using it. That limits the ability of other groups to make inventions available to society at a lower price, or use knowledge spillovers to improve on the ideas with their own R&D investments.

It also makes sense for non-profit organizations to focus their R&D efforts on the kinds of innovation where it is hardest for private firms to appropriate value or where trying to appropriate value has the biggest negative impacts. For example, a discovery that has very many knowledge spillovers would be a particularly bad discovery to keep secret since it would prevent many other new inventions from being discovered.

A very important kind of research that has this property is science. Scientific knowledge can usually lead to many different technologies; just think of how many technologies rely on electricity! But precisely because science can lead to many different kinds of inventions, if a private company invests in scientific research, most of the benefits of the knowledge are likely to go to other firms creating other inventions. This doesn't mean scientific research won't sometimes occur at private companies since the company doing the research might get some benefit too. However, in practice, most scientific research in the United States is not done by private companies. Although the federal government and non-profit sectors only fund 27 percent of all R&D, they fund 67 percent of all scientific research.[54]

As noted earlier, economists estimate that on average, every dollar spent on R&D returns at least $3.60 in social value. This suggests we should be doing more R&D than we are, but as we've seen, it's hard for an inventor to capture the value of their R&D. One way

One way governments try to encourage the private sector to do more R&D is by funding science since better scientific knowledge can make it easier to invent.

governments try to encourage the private sector to do more R&D is by funding science since better scientific knowledge can make it easier to invent.

The government tries to encourage firms to invent in various other ways as well. The government subsidizes the education and training of new scientists through scholarships. The government also grants tax breaks on R&D investments, so that firms that do more R&D can pay less in taxes than otherwise similar firms that do less R&D. The government may also give loans or grants to companies for doing R&D. Lastly, the government can create and enforce a system of intellectual property rights, such as patents. These policies do not always encourage R&D because a government may administer them poorly or corruptly (see, for example, the discussion of institutions and innovation at the end of this section), but when done well, they can facilitate more R&D projects with social value exceeding their cost.

The Size of the Market and R&D

Another factor that affects how much and what kinds of R&D private companies do is the size of the market for different kinds of inventions. As noted earlier, in the preindustrial era, markets tended to be small, which made innovation less profitable. When international trade becomes easier, the ability to sell into larger markets can spur more innovation. Matthew Boulton, who partnered with the inventor James Watt to sell the first steam engines, once wrote: "It is not worth my while to manufacture your engine for three countries only, but I find it very worth my while to make it for all the world."[55]

The size of a market also affects what kinds of things are invented. For example, when certain diseases become

During the COVID-19 pandemic, the U.S. government paid for a large quantity of vaccines and then made them freely available.

more prevalent in the United States, pharmaceutical companies increase their R&D on drugs to treat those diseases.[56] Unfortunately, this can also lead to problems when the size of the market is different from the social value of a specific innovation. For example, in 2017 roughly one out of every six people was infected by one or more tropical diseases.[57] The social value of better medical treatments for these diseases would therefore be very high. However, since most of these people live in low-income countries, their ability to pay for treatment is low, and so the market value of better medical treatments is comparatively low. Since firms need to recoup their R&D investments, they have little incentive to invest in R&D for these diseases. In fact, out of more than 1,200 new medicines licensed worldwide between 1975 and 1997, only four targeted these kinds of tropical diseases.[58]

Some government policies try to close the gap between the social and market value of inventions. For example, in 2007, a coalition of governments and the Gates foundation got together and proposed to pay for vaccines for a disease common in low income countries, agreeing to pay a price that was above what people in that country would normally be able to pay.[59] Similar policies were used by the U.S. government during the COVID-19 pandemic, where the U.S. government agreed to pay for a large quantity of vaccines and then to make them freely available.[60]

Cities and Innovation

Many important non-economic factors can also have a large impact on innovation. For one, where an inventor lives seems to matter quite a lot. Cities, for example,

Cities tend to be responsible for a larger share of invention than would be expected given their share of the population.

By Rickmouser45 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=93429418

tend to be responsible for a larger share of invention than would be implied simply by their share of the population. There seems to be something about living in a city full of other inventors that is good for inventing.

There are a lot of ideas about why this is the case. One theory is that living close to other inventors might make it easier for them to learn from each other. Living and working near other inventors helps inventors create professional networks that let them pass new ideas around more easily. We can see evidence of this local learning, for example, in the citations inventors make to other patents, which are much more likely to go to the patents of local inventors.[61] But there are also other reasons that might be important too. Cities tend to have higher wages, and so they can attract talented workers. Being around lots of other talented inventors can make it easier to learn from the best or to assemble

teams of inventors that are unusually strong. A city full of inventors is also more likely to have people who are willing to lend money to inventors, so they can do R&D. The innovation boost that cities bring about is a form of **agglomeration benefit**, a term that refers more generally to the ways that living in cities can make workers more productive.

Institutions and Innovation

Where an inventor lives can matter for other, more fundamental, reasons as well. Throughout this discussion, we have assumed that an inventor can use one or more of the strategies discussed above to appropriate part of the value their invention creates. But, in fact, this is not guaranteed. For one, being able to benefit from an invention in the ways we've discussed requires that the inventor has some property rights and can participate in markets. For instance, throughout history enslaved people have had little reason to believe they would be able to benefit much from invention and so probably did not bother with it much. And under a socialist or planned economy, markets may not be used at all, which makes it impossible to benefit by selling access to a new invention as described above.

But not only access to markets matters. Even if they live in a country where they can enjoy property rights and use markets, inventors may also be dissuaded from bothering with the hard work of invention if they are not protected by the rule of law. As we discussed earlier, technological progress tends to happen via a process of creative destruction, where new and better technologies may make older ones obsolete. These older industries may very well try to resist this process, and compared to the inventor of a new technology, they might have more wealth and political connections, which they could use to stop new inventions from becoming available. If an inventor thinks that existing businesses and industries will probably succeed in stopping anyone who tries to challenge their way of business, then that inventor might decide trying to invent new things is not worth the effort.

A country's economic and political system, the extent of rule of law, and so on are all part of what economists call **institutions**. Institutions are "the rules of the game" for a given society, and they shape the benefits and costs of different economic decisions (and other kinds of decisions as well). The concept is quite broad, but one way to get a bit more specific is to divide the concept into two main categories: *extractive institutions* and *inclusive institutions*. A society living under extractive institutions is one where most people are excluded from political decision making and the rules of society are fixed to extract the wealth of the many to benefit the few in power. Think, for example, of a country ruled by a king with absolute power, who can claim the wealth of anyone (including a successful innovator).

In contrast, inclusive institutions allow and encourage everyone to participate in both the economy and political decision-making. Here, we might think of a democracy with a strong rule of law and broadly accessible education. It's easy to imagine that under extractive institutions few people will try to invent new technologies since they (rightfully) believe they are unlikely to be able to benefit from success. In contrast, under inclusive ones many potential innovators will be willing to try to invent, confident the only risks they face are the ones inherent to invention and that if they succeed, they will be able to reap the rewards of their effort. At a broad historical level, the move away from extractive institutions to more inclusive ones is seen as one major cause of the Great Enrichment.

Culture and Innovation

Lastly, it may be that the beliefs, values, and ideas of the community in which you live matter. The economic historian Joel Mokyr has argued that the key factor leading to the Great Enrichment was the invention of the very *idea* of progress and growth.[62] Perhaps, for most of history, most people believed that inventing better ways of doing things was simply impossible. That wouldn't be a crazy thing to believe since economic progress really was very, very slow for most of history. So long as people believe progress is not possible, they are unlikely to put much effort into inventing. Under this theory, once progress got going, these beliefs began to change because people could see that it was possible to invent better ways of doing things. From that point on, the factors we have discussed influenced how much effort people put into invention. But for any of them to matter in the first place, people had to believe progress was possible.

SECTION IV SUMMARY

- Life in the preindustrial world was hard. Most people were poor, lived in small and uncomfortable homes, worked difficult jobs, ate monotonous diets, rarely traveled long distances, could not read or write, and suffered from disease.
- Since around 1800, industrialized nations have seen great advances in material living standards, connectedness through transportation and information networks, and health and safety—a "Great Enrichment."
- One of the primary drivers of these economic advances has been a revolution in technology. This includes the mechanization of factories and farms, better materials for manufacturing and construction, greater agricultural productivity, new sources of energy, powered transportation, electronic communications, and better health through sanitation, vaccines, and antibiotics.
- The United States spends about 2.5 percent of GDP on research, which makes it the largest supplier of research in the world. Economists estimate this leads to about 1 percent annual growth in total factor productivity, a common way of trying to measure technological progress. Because R&D leads to permanently higher living standards, it is estimated that every $1 spent on R&D results in at least $3.60 in social value.
- Private companies may underinvest in R&D because rivals can copy their ideas and because R&D has positive externalities in the form of knowledge spillovers. Businesses try to appropriate part of the value they create in a variety of ways, including the use of secrecy, lead time, complementary assets, and intellectual property rights, but these are generally imperfect.
- Innovation is also supported by a number of other factors, such as government and non-profit investment in research (especially scientific research), a big market in which to sell inventions, and the presence of many other local inventors.
- Institutions affect how much inventors expect to be able to able to benefit from invention. Institutions, and the very idea of progress itself, are key determinants of whether potential inventors become actual inventors.

Glossary

Agglomeration benefit – the increase in productivity associated with production that occurs in places with a denser population, such as cities

Aggregate demand curve – a graphical depiction of the relationship between the level of desired expenditures in an economy and the price level

Aggregate supply curve – a graphical depiction of the relationship between the quantity of goods and services firms wish to supply and the price level

Appropriability – how much of the social value of an innovation the inventor can capture for themselves

Average labor productivity – total output divided by the quantity of labor employed in its production

Bank run – a sudden rush of depositors seeking to withdraw funds from the banking system

Barriers to entry – conditions that prevent firms from freely entering or exiting a market

Business cycle – fluctuations in aggregate economic activity

Capital – one of three factors of production; in classical economics, capital refers to money or physical assets. Plows or mature tree crops may be considered forms of capital in this context.

Capital equipment – material goods that increase productivity, such as machines, factories, vehicles, and infrastructure

Capital goods – long-lived goods that are themselves produced and are used to produce other goods and services, but are not used up in the production process

Cartel – a group of firms that collude in a given market to restrain competition, often making quota arrangements among themselves

Coase Theorem – the proposition that if private parties can bargain without cost over the allocation of resources, then they can solve the problem of externalities on their own

Comparative advantage – the ability to produce a good or service at a lower opportunity cost than other producers

Competitive market – a market with many buyers and sellers trading a homogenous good or service in which each buyer and seller is a price taker

Complementary asset – a possession that allows a firm to obtain more value from an invention than rivals

Complements – two goods for which a rise in the price of one leads to a decline in the demand for the other

Consumer Price Index (CPI) – an index constructed by comparing the cost of purchasing a fixed basket of goods at different times

Consumer surplus – the difference between the amount that a buyer would be willing to pay for a good or service and the price actually paid

Consumption – spending by households on goods and services, with the exception of the purchase of new housing

Creative destruction – the process by which new technologies make some older ones obsolete

Crowding out – the decrease in private investment that occurs as a result of a reduction in government saving or an increase in government borrowing

Currency – coins and bills in the hands of the public

Cyclical unemployment – unemployment caused by deviations of output from its potential level

Deadweight loss – the reduction in total surplus that results from a market distortion such as a tax

Demand curve – a graphical representation of the quantity of a good or service demanded as a function of the price

Demand schedule – a table showing the relationship between the price of a good or service and the quantity demanded

Depression – a severe recession

Diminishing returns to scale – the property whereby each additional increase in inputs results in a smaller increase in the quantity produced

Discount rate – the interest rate that the Federal Reserve charges banks when they must borrow reserves from it

Economic profit – the difference between the revenue realized by a producer and the opportunity cost of production

Economies of scale – cost advantages gained by a company when output increases; the average cost per unit decreases as the number of units produced increases.

Elasticity – the percentage change in quantity demanded or supplied as a result of a one percent change in price

Entrepreneur – an individual who takes on the risk of attempting to create new products or services, establish new markets, or develop new methods of production

Equilibrium – a situation in which the forces in a system are in balance so that the situation is stable and unchanging

Excludable good – a good that an individual can prevent another individual from using

Excludability – the ability to prevent buyers from enjoying the benefits of consuming a good or service without paying for it

Expansion – a period between a trough and a peak in economic activity

Externality – when the action of one person affects the well-being of someone else, but where neither party pays nor is paid for these effects

Federal funds rate – the rate that banks charge other banks when they lend reserves

Final goods – goods or services that are purchased by their ultimate user

Financial markets – the institutions through which individuals with savings can supply these funds to persons or firms that wish to borrow money to purchase consumption goods or invest in physical capital

Fiscal policy – the use of taxes and spending to influence aggregate demand and through it the level of overall economic activity

Fixed cost – a cost of production that is independent of the quantity produced

Foreign direct investment – when a company or individual acquires assets in a foreign country that they will manage directly

Frictional unemployment – unemployment that results because it takes time for workers to search for the jobs that are best suited to their tastes and skills

Gains from trade – the benefits that both individuals or nations realize from mutually beneficial exchange

Government purchases – spending on goods and services by federal, state, and local governments

Gross Domestic Product (GDP) – the market value of final goods and services produced in an economy during a specified period of time

Gross Domestic Product (GDP) per capita – estimate of national output (gross domestic product), divided by the population; its key advantage as a measure of economic performance is in giving an average level of income per person, which can be compared between countries.

Human capital – skills and experience that are acquired through education, training, and on-the-job experience that increase a worker's productivity; considered an important factor in facilitating improvements in productivity and economic growth

Imperfect competition – the case of a market with a small number of sellers, so that sellers have market power

Industrial Revolution – a historical period during which some economies (most notably the United Kingdom and the United States) first made a transition from "from an agrarian and handicraft economy to one dominated by industry and machine manufacturing," roughly 1760–1840

Inferior good – a good for which the quantity demanded falls as buyers' income increases

Inflation – a general increase in prices

Institutions – the "rules of the game" in a society, which shape the costs and benefits of economic decisions

Intellectual property – exclusive rights over the control of an idea; common examples are trademarks, copyright, and patents.

Intermediary – a third party who acts as a link between two others who wish to transact business

Intermediate good – a good or service that is used in the process of producing other goods and services

Investment – spending on capital equipment, inventories, and structures, including household purchases of new housing

Keynesian model – a model of short-run aggregate economic fluctuations inspired by the analysis of British economist John Maynard Keynes, which attributes short-run deviations in output from potential to variations in the level of aggregate demand or aggregate supply

Knowledge spillover – a positive externality that occurs when research by one group is useful for other groups that did not do the R&D

Labor force – the sum of those individuals who are employed and those who are seeking paid work but have not found it

Labor force participation rate – the fraction of the working-age population who are in the labor force

Labor productivity – the amount of output produced per worker or per worker-hour

Law of demand – holding other things equal, the quantity demanded is negatively related to the price

Law of supply – holding other things equal, the quantity supplied is positively related to the price

Lead time – a period of time between when an invention is discovered and when rivals learn how to manufacture it

Liquidity – the ease with which a nonmonetary asset may be converted into money

Logrolling – the practice of elected officials trading votes

Marginal cost – the additional cost of production associated with a small increase in the quantity produced

Marginal revenue – the additional revenue resulting from a small increase in the quantity produced

Market failure – any situation in which a market does not do what market theorists believe it should—allocate goods and services efficiently; externalities and monopoly/oligopoly are two commonly discussed failures; a situation where the allocation of goods and services is not efficient, for instance when too much of a good (e.g., pollution) or too little of a good (e.g., clean air) is provided by the free market

Market power – a situation in which one firm, or a group of them acting as a cartel, can control prices in a market, often by restricting output, and thus have market power; in a theoretical, purely competitive market, this is not possible

Monetary base – the quantity of currency plus bank reserves

Monetary policy – the use of the supply of money in the economy by the Federal Reserve to influence the level of aggregate demand

Money – an asset that is a medium of exchange, unit of account, and store of value

Money multiplier – the ratio of the money supply to the monetary base

Money supply – the quantity of money available to the economy

Monopolistic competition – a market in which there is free entry or exit, but every producer supplies a differentiated product and faces a downward-sloping demand curve

Monopoly – a market in which there is a single producer

Natural rate of unemployment – the level of unemployment that would exist if the economy were producing at its potential output

Net capital outflow – the difference between the purchases of foreign assets by domestic residents and the purchases of domestic assets by foreign residents

Net exports – the difference between the value of goods and services sold to foreigners and the value of goods and services purchased from foreigners

Network effects – products whose value to an individual user increases as more people use them

Neutrality of money – the proposition that in the long run, changes in the quantity of money affect the price level but do not affect any real quantities

Nominal GDP – the production of goods and services valued at current prices

Normal good – a good or service for which demand is

positively related to the buyer's income

Normative economics – economic analysis used to guide decisions about what should be as opposed to what is the case

Okun's law – a relationship identified by Arthur Okun between the output gap and the level of cyclical unemployment

Oligopoly – a market in which there are just a few producers

Open market operations – a tool used by the Federal Reserve to adjust the money supply by buying or selling U.S. government bonds in the financial market

Opportunity cost – the cost of any choice is what must be given up by making that choice

Output gap – the difference between actual output and potential output

Pareto efficiency – describes an allocation in which the only way to make any individual or group of individuals better off would require making at least one other person worse off

Patents – intellectual property rights over inventions that confer twenty years of exclusive control over novel, non-obvious, and useful new inventions if the inventor discloses a description of how the invention works

Per capita – literally per head, used to denote an average value for a population

Portfolio investment – the purchase of shares of stock or bonds

Positive economics – the use of the tools of economic analysis to describe and explain economic phenomena and to make predictions about what will happen under particular circumstances

Potential output – the quantity of output that would be produced by an economy if all of its resources were being employed at normal rates

Price discrimination – when a business sells the same product to different buyers at different prices

Price elasticity of demand – the amount by which demand for a given product changes in response to changes in price; specifically, the percentage change in demand that corresponds to a one percent change in the price

Process innovation – a technological innovation that makes it feasible to produce more (or higher quality) outputs using the same number of inputs

Producer surplus – the difference between the price that producers receive for supplying a good and their marginal cost of producing it

Product innovation – a technological innovation that makes it feasible to produce a new kind of output

Production possibility frontier (PPF) – a graphical depiction of the combinations of output that can be produced by an economy

Public good – a good or service for which it is not possible to establish individual property rights

Rationality – when individual choices are made by comparing the benefits and costs of different actions and then selecting the action that produces the greatest benefit

Real GDP – the production of goods and services valued at constant prices

Recession – a period between a peak and a trough in economic activity

Rent seeking – using political influence to increase one's economic profits at the expense of others

Reserve requirement – the amount of reserves that the Federal Reserve requires banks to hold

Reserves – the fraction of deposit liabilities that banks hold to meet depositor withdrawals

Rival good – a good that cannot be used by more than one person at once without being diminished

Savings – the difference between a person's disposable income and their expenditures

Scarcity – an inescapable fact of human existence that results from the fact that the available resources are always less than our limitless desires

Structural unemployment – unemployment that results from the mismatch in skills, locations, or other important characteristics between job seekers and the available jobs

Substitutes – two goods for which an increase in the price of one leads to an increase in the demand for the other

Supply curve – a graphical representation of the quantity of a good or service supplied as a function of the price

Supply schedule – a table showing the relationship between the price of a good or service and the quantity supplied

Technology – a process for transforming a set of inputs into an output

Total factor productivity – a measure of how much output can be produced from a given supply of inputs; often interpreted as a measure of technological progress, though other non-technological factors can also impact it

Total revenue – the total revenue received by a supplier

Total surplus – the sum of consumer and producer surplus

Tragedy of the commons – the depletion of a common resource due to overuse

Unemployment – the state of actively seeking paid work but being unable to find it

Unemployment rate – the number of unemployed workers as a fraction of the total labor force

Variable cost – a cost of production that depends on the quantity produced

Velocity of money – the ratio of nominal GDP to the money supply; in effect, the average number of transactions supported by each dollar of the money supply

Wealth – the total value of assets used as a store of value

Notes

1. "Supermarket Facts: Industry Overview 2008," Food Marketing Institute, 26 June 2009 <http://www.fmi.org/facts_figs/?fuseaction=superfact>.
2. The auctioneer may be an actual person, or the process of matching sellers and buyers may be accomplished by means of a computer network.
3. Many consumers object to genetically modified foods such as milk produced using BGH. The use of BGH by U.S. dairy farmers has led to an ongoing dispute between the U.S. and the European Union in the World Trade Organization (WTO) concerning hormone-fed beef.
4. Formally, the elasticity measures the percentage change in quantity demanded caused by a one percent change in price at a specific point on the demand curve. When we measure changes over finite distances, the results will depend on the position we take as our starting point. To avoid this problem, the convention is that we calculate the percentage change with reference to the midpoint of the initial and final values. If the price changed from P1 to P2, then we would calculate the percentage change as:

$$\text{pct change} = 100 \times \frac{P2 - P1}{(\frac{1}{2}) \cdot (P2 + P1)}.$$

5. In reality, farmers can choose to produce organic milk, which consumers understand is BGH free. Because some consumers prefer milk produced without BGH, farmers who choose to produce in compliance with the requirement to label their product as organic can command a higher price for their product.
6. Bob's costs of producing a quantity q are $C(q) = 300 + 2q + q^2/300$, and the marginal costs corresponding to this equation are $MC(q) = 2 + q/150$.
7. N. Gregory Mankiw, *Principles of Economics*, 4th ed. (Mason, OH: Thomson Southwestern, 2007) 345–46.
8. Rachel Holliday Smith, "How Does Congestion Pricing Work? What to Know About the Toll System Taking Manhattan," The City, 15 September 2021, Accessed 5 November 2021 <https://www.thecity.nyc/2021/9/15/22674371/how-does-congestion-pricing-work-toll-system-in-manhattan>.
9. "Real Gross Domestic Product (GDP) of the United States of America from 1990 to 2019 (in billion chained (2012) U.S. dollars)," Statista, Accessed 15 September 2020 <https://www.statista.com/statistics/188141/annual-real-gdp-of-the-united-states-since-1990-in-chained-us-dollars/>.
10. International comparisons of the sort presented in FIGURE 32 are sensitive to the prices that are used to compare production across the different countries. The comparisons made here use current exchange rates to convert national GDP figures into dollars, a practice that results in an understatement of the standard of living in lower-income countries. Using an alternative approach that better reflects actual purchasing power in the different countries would perhaps double or triple income levels in countries like Ghana or Nigeria. While this would narrow the gap in living standards relative to the U.S., the gap still remains huge.
11. The official series probably greatly overstates the economic growth of World War II. See, for example, Robert Higgs, "Wartime Prosperity? A Reassessment of the U.S. Economy in the 1940s," *Journal of Economic History*, 52, no. 1 (March 1992).
12. See Richard Sutch, "National Income and Product," *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*, Eds. Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright. New York: Cambridge University Press, 2006.
13. For a more extensive discussion of these issues, see Clifford Cobb, Ted Halstead, and Jonathan Rowe, "If the GDP is Up, Why is America Down," *Atlantic Monthly* (October 1995) 59–78. Notes 2010–2011 § Economics Resource Guide 127.
14. Michael Boskin, et al., "Toward a More Accurate Measure of the Cost of Living," *Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index*, December 1996, Social Security Online, 20 Aug. 2009 <http://www.ssa.gov/history/reports/boskinrpt.html>.
15. We assume here that the government compensates for the income lost as a result of the tax credit by increasing taxes on some other transactions. If this were not the case, then government savings would be reduced, and the tax credit would have no net effect on the economy.
16. See <http://www.federalreserve.gov/faqs/about_12591.htm>.
17. This number is adjusted for inflation (given in 2011 dollars) and for purchasing power. From Roser and Ortiz-Ospina (2013). Students should be aware that the dividing line between "preindustrial" and "industrial" is not sharp, and it occurs in different parts of the world at different times. It might be put at mid- to late 1700s in the UK, early 1800s in the U.S., roughly the same for the rest of Western Europe, and not until the second half of the twentieth century in much of Asia (and maybe in the future for parts of Africa).
18. Roser (2013a)
19. The remaining 4 percent were clerical workers, sales workers, and other service workers. Gordon (2017). Table 8-1, p. 330
20. Hindle and Lubar (1986), p. 219
21. Roser and Ortiz-Ospina (2016)
22. Estimates can vary widely, depending on the source. Some historians suggest a figure as high as 60 percent. See for instance: Benedictow (2004).
23. Roser, Ritchie and Dadonaite (2013)
24. Roser (2013b)
25. Giattino, Ortiz-Ospina and Roser (2020)
26. In the U.S., for instance, in the twentieth century alone, infectious disease mortality declined from 797 deaths per 100,000 in 1900 to 36 in 1980: Armstrong, Conn and Pinner (1999)
27. Roser, Ortiz-Ospina and Ritchie (2013)
28. Ritchie and Roser (2018)
29. Throughout this overview, the pronoun "we" is often used in the grand sense referring to "humanity" to indicate human accomplishments. Although of course many accomplishments were driven by individual scientists, entrepreneurs, government programs, etc., for the purposes of this summary, the focus remains largely on the impact of the advancements rather than the details of who pushed them forward.
30. Ritchie and Roser (2013)
31. Bernstein (2008), p. 330
32. Armstrong, Conn, Pinner (1999)
33. Solow (1956)

34. Solow (1957)
35. Jones (2021)
36. Solow's production function does not include education, as written, but more realistic versions do.
37. Note we don't actually measure technology directly. We measure everything else and assume the rest comes from technology, which we can't measure. Our measure might be biased if there are many non-technological things we are failing to measure properly.
38. Jones (2021)
39. Jones (2010)
40. Bloom et al. (2020)
41. Calculations derived from data in Figures RD-5 and RD-6 of NSF (2022).
42. All calculations derived from data in Figure RD-3 of NSF (2022).
43. Jones and Summers (2021)
44. Some readers may wonder if there is a simple explanation for this, namely that it might be that although the *average* dollar spent on R&D returns $3.60 in value, this is not true of the *marginal* dollar. In other words, maybe we don't invest more because we've already invested in all the good ideas and the ones that are left don't generate more than a dollar in value for each dollar spent. As best as we can tell, that doesn't seem to be the whole story: for example, studies have found the last innovative companies and the last proposals for scientific research that qualify for government funding usually generate a lot of value for the money spent (Howell 2017, Myers and Lanahan 2021, Azoulay et al. 2019). We wouldn't expect that to be the case if all the best projects are easily identified and the last ones are barely worth funding at all.
45. Technically, The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel
46. Cohen, Nelson, and Walsh (2000)
47. Ibid.
48. Ibid.
49. Clark (2022)
50. Cohen, Nelson, and Walsh (2000)
51. Mezzanotti and Simcoe (2022)
52. Myers and Lanahan (2021)
53. Nordhaus (2004)
54. NSF (2022), table RD-3, using basic research as a measure of science.
55. Scherer (1989), p. 13
56. Acemoglu and Linn (2004)
57. Mitra and Mawson (2017)
58. Glennerster, Kremer, and Williams (2005)
59. Kremer, Levin, and Snyder (2020)
60. Gross and Sampat (2021)
61. Jaffe, Trajtenberg, and Henderson (2000)
62. Mokyr (2018)

Bibliography

Acemoglu, D., and J. Linn. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry." *The Quarterly Journal of Economics* 119, no. 3 (August 1, 2004): 1049–90. https://doi.org/10.1162/0033553041502144.

Armstrong, G.L., L.A. Conn, and R.W. Pinner. "Trends in Infectious Disease Mortality in the United States During the 20th Century" *JAMA* 281:1 (1999): 61–66. doi:10.1001/jama.281.1.61.

Atack, J. and Passell, P. *A New Economic View of American History*, 2nd ed. New York: W. W. Norton, 1994.

Azoulay, Pierre, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat. "Public R&D Investments and Private-Sector Patenting: Evidence from NIH Funding Rules." *The Review of Economic Studies* 86, no. 1 (January 1, 2019): 117–52. https://doi.org/10.1093/restud/rdy034.

Benedictow, Ole Jørgen. *The Black Death 1346–1353: The Complete History*. United Kingdom, Boydell Press, 2004.

Bernstein, William J. *A Splendid Exchange: How Trade Shaped the World*. United Kingdom, Atlantic Books, 2008.

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb. "Are Ideas Getting Harder to Find?" *American Economic Review* 110, no. 4 (April 1, 2020): 1104–44. https://doi.org/10.1257/aer.20180338.

Boskin, Michael, et al. "Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index," December 1996. Online resource http://www.ssa.gov/history/reports/boskinrpt.html Accessed 20 August 2009.

Brown, E. C. "Fiscal Policy in the Thirties: A Reappraisal." *American Economic Review* 46 (1956).

Carter, Susan B., Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright, eds. *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*. New York: Cambridge University Press, 2006.

Clark, Don. "The Huge Endeavor to Produce a Tiny Microchip." *New York Times*, April 8, 2022. https://www.nytimes.com/2022/04/08/technology/intel-chip-shortage.html.

Cohen, Wesley, Richard Nelson, and John Walsh. "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)." Cambridge, MA: National Bureau of Economic Research, February 2000. https://doi.org/10.3386/w7552.

Cobb, Clifford, Ted Halstead and Jonathan Rowe. "If the GDP is Up, Why is America Down." *Atlantic Monthly*, October 1995.

Edelstein, Michael. "War and the American Economy in the Twentieth Century." *The Cambridge Economic History of the United States.* Edited by Stanley L. Engerman and Robert E. Gallman, Vol III. New York: Cambridge University Press, 2000.

Eichengreen, Barry. *Golden Fetters: The Gold Standard and the Great Depression, 1919–1939.* New York: Oxford University Press, 1992.

Federal Reserve, "Frequently Asked Questions." Internet resource: http://www.federalreserve.gov/generalinfo/faq/faqfrs.htm#return. Accessed 10 October 2009.

Field, Alexander J. *A Great Leap Forward: 1930s Depression and U.S. Economic Growth.* New Haven, CT: Yale University Press, 2012.

Food Marketing Institute, "Supermarket Facts," internet resource: http://www.fmi.org/facts_figs/?fuseaction=superfact. Accessed 26 June 2009.

Friedman, Milton and Anna Schwartz. *A Monetary History of the United States, 1867–1960.* Princeton: Princeton University Press, 1963.

Giattino, Charlie, Esteban Ortiz-Ospina, and Max Roser. "Working Hours." *Our World in Data* (2020): https://ourworldindata.org/working-hours.

Glennerster, Rachel, Michael Kremer, and Heidi Williams. "The Price of Life." Prime Numbers, June 2005.

Goodwin, Richard W. *Remembering America*. Boston: Little, Brown & Company, 1988.

Gordon, Robert J. *The Rise and Fall of American Growth: The U.S. Standard of Living Since the Civil War*. United States, Princeton University Press, 2017.

Gordon, Robert and Robert Krenn. "The End of the Great Depression 1939–41: Policy Contributions and Fiscal Multipliers." *NBER Working Paper Series*, September 2010.

Gross, Daniel, and Bhaven Sampat. "Crisis Innovation Policy from World War II to COVID-19." Cambridge, MA: National Bureau of Economic Research, June 2021. https://doi.org/10.3386/w28915.

Higgs, Robert. *Crisis and Leviathan Critical Episodes in the Growth of American Government*. Oakland: The Independent Institute, 1987.

Hindle, Brooke, and Stephen Lubar. *Engines of Change: The American Industrial Revolution, 1790–1860*. United Kingdom, Smithsonian, 1986.

Howell, Sabrina T. "Financing Innovation: Evidence from R&D Grants." *American Economic Review* 107, no. 4 (April 1, 2017): 1136–64. https://doi.org/10.1257/aer.20150808.

Hughes, Jonathan and Louis P. Cain. *American Economic History,* 7th ed. Boston: Pearson, 2007.

Jaffe, A. B., M. Trajtenberg, and R. Henderson. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *The Quarterly Journal of Economics* 108, no. 3 (August 1, 1993): 577–98. https://doi.org/10.2307/2118401.

Johnston, Louis D. and Samuel H. Williamson. "What Was the U.S. GDP Then?" *MeasuringWorth.* 2008. Internet Resource: http://www.measuringworth.org/usgdp/. Accessed: 10 October 2009.

Jones, Benjamin F. "Age and Great Invention." *Review of Economics and Statistics* 92, no. 1 (February 2010): 1–14. https://doi.org/10.1162/rest.2009.11724.

Jones, Benjamin, and Lawrence Summers. "A Calculation of the Social Returns to Innovation." Cambridge, MA: National Bureau of Economic Research, September 2020. https://doi.org/10.3386/w27863.

Jones, Charles I. "The Past and Future of Economic Growth: A Semi-Endogenous Perspective." *Annual Review of Economics* 14, no. 1 (August 12, 2022): 125–52. https://doi.org/10.1146/annurev-economics-080521-012458.

Keynes, John Maynard. *The Economic Consequences of the Peace*. New York: Harcourt, Brace, and Howe, Inc., 1919.

Kremer, Michael, Jonathan Levin, and Christopher Snyder. "Advance Market Commitments: Insights from Theory and Experience." Cambridge, MA: National Bureau of Economic Research, February 2020. https://doi.org/10.3386/w26775.

Kudlow, Lawrence and Brian Domitrovic. *JFK and the Reagan Revolution: A Secret History of American Prosperity.* New York: Penguin, 2016.

Lebergott, Stanley. *The American Economy: Income, Wealth and Want*. Princeton: Princeton University Press, 1962.

Leuchtenberg, William E. *The FDR Years: On Roosevelt and His Legacy*. New York: Columbia University Press, 1995.

Maddison, Angus. *Contours of the World Economy, 1–2030 AD.* Oxford: Oxford University Press, 2007.

Mankiw, N. Gregory. *Principles of Economics,* 4th ed. Mason, OH: Thomson Southwestern, 2007.

Meyer, Bruce D. and James X. Sullivan. 2013. "Winning the War: Poverty from the Great Society to the Great Recession." NBER Working Paper 18718.

Mezzanotti, Filippo, and Timothy Simcoe. "Innovation and Appropriability: Revisiting the Role of Intellectual Property." *Center for Economic Studies* 22–09 (n.d.).

Mitra, Amal, and Anthony Mawson. "Neglected Tropical Diseases: Epidemiology and Global Burden." *Tropical Medicine and Infectious Disease* 2, no. 3 (August 5, 2017): 36. https://doi.org/10.3390/tropicalmed2030036.

Mokyr, Joel. *A Culture of Growth: The Origins of the Modern Economy*. First paperback printing. The Graz Schumpeter Lectures. Princeton Oxford: Princeton University Press, 2018.

Morse, Jarvis M. *Paying for a World War: The United States Financing of World War II.* United States Department of Treasury, U.S. Savings Bonds Division, 1971.

Myers, Kyle R., and Lauren Lanahan. "Estimating Spillovers from Publicly Funded R&D: Evidence from the US Department of Energy." *American Economic Review* 112, no. 7 (July 1, 2022): 2393–2423. https://doi.org/10.1257/aer.20210678.

Nordhaus, William. "Schumpeterian Profits in the American Economy: Theory and Measurement." Cambridge, MA: National Bureau of Economic Research, April 2004. https://doi.org/10.3386/w10433.

NSF. "Research and Development: U.S. Trends and International Comparisons." National Science Board, National Science Foundation, 2022. https://ncses.nsf.gov/pubs/nsb20225/.

Officer, Lawrence F. "The Annual Consumer Price Index for the United States, 1774–2008." *MeasuringWorth*. 2009. Internet Resource: http://www.measuringworth.org/uscpi/. Accessed: 17 October 2009.

Okun, Arthur M. 1970. *The Political Economy of Prosperity.* Washington, D.C.: The Brookings Institution.

Olney, Martha. *Buy Now, Pay Later: Advertising, Credit, and Consumer Durables in the 1920s*. Chapel Hill: University of North Carolina Press, 1991.

Parker, Randall. "An Overview of the Great Depression." *EH.Net Encyclopedia*. Robert Whaples, editor (March 16, 2008). Internet Resource http://eh.net/encyclopedia/article/parker.depression. Accessed 17 October 2009.

Peppers, Larry. "Full Employment Surplus Analysis and Structural Change: the 1930s." *Explorations in Economic History* 10 (1973).

Reardon, Terry. *Winston Churchill and Mackenzie King: So Similar, So Different.* A. J. Patrick Boyer Boo, 2012.

Record, Jeffrey. "Japan's Decision for War in 1941: Some Enduring Lessons." *Stategic Studies Institute.* February 2009. http://www.strategicstudiesinstitute.army.mil/pubs/summary.cfm?q=905 (accessed August 13, 2015).

Ritchie, Hannah, and Max Roser. "Causes of Death." *Our World in Data* (2018): https://ourworldindata.org/causes-of-death.

Ritchie, Hannah, and Max Roser. "Crop Yields." *Our World in Data* (2013): https://ourworldindata.org/crop-yields.

Rockoff, Hugh. *America's Economic Way of War: War and the US Economy from the Spanish-American War to the Persian Gulf War.* Cambridge University Press, 2012.

Rockoff, Hugh. "The United States: from ploughshares to swords." *The Economics of World War II: Six Great Powers in International Comparison.* Mark Harrison, ed. Cambridge: Cambridge University Press, 2000.

Roser, Max. "Economic Growth." *Our World in Data* (2013b): https://ourworldindata.org/economic-growth.

Roser, Max. "Employment in Agriculture." *Our World in Data* (2013a): https://ourworldindata.org/employment-in-agriculture.

Roser, Max, and Esteban Ortiz-Ospina. "Global Extreme Poverty." *Our World in Data* (2013): https://ourworldindata.org/extreme-poverty.

Roser, Max, and Esteban Ortiz-Ospina. "Literacy." *Our World in Data* (2016): https://ourworldindata.org/literacy.

Roser, Max, Hannah Ritchie, and Bernadeta Dadonaite. "Child and Infant Mortality" *Our World in Data* (2013): https://ourworldindata.org/child-mortality.

Roser, Max, Esteban Ortiz-Ospina, and Hannah Ritchie. "Life Expectancy." *Our World in Data* (2013): https://ourworldindata.org/life-expectancy.

Scherer, Frederic M. *Innovation and Growth: Schumpeterian Perspectives*. 3. print. Cambridge, Mass.: MIT Press, 1989.

Solow, Robert M. "A Contribution to the Theory of Economic Growth." *The Quarterly Journal of Economics* 70, no. 1 (February 1956): 65. https://doi.org/10.2307/1884513.

———. "Technical Change and the Aggregate Production Function." *The Review of Economics and Statistics* 39, no. 3 (August 1957): 312. https://doi.org/10.2307/1926047.

Sutch, Richard. "National Income and Product." *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition.* Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright,

eds. New York: Cambridge University Press, 2006.

Tassava, Christopher. "Eh.net." February 10, 2008. http://eh.net/encyclopedia/the-american-economy-during-world-war-ii/ (accessed August 2, 2015).

Temin, Peter. *Lessons from the Great Depression.* Cambridge: MIT Press, 1989.

Tucker, Spencer C. "Overview of the Vietnam War." *The Encyclopedia of the Vietnam War: A Political, Social, and Military History. 2nd* ed. Santa Barbara, CA: ABC-CLIO, 2011.

United Nations, *Human Development Reports*, 2005. Internet Resource: http://hdr.undp.org/sites/default/files/reports/266/hdr05_complete.pdf. Accessed 10 October 2009.

United States, Department of Labor, Bureau of Labor Statistics, Employment Situation Summary, Sept. 4, 2009. Internet resource: http://www.bls.gov/news.release/empsit.toc.htm. Accessed: 20 September 2009.

Walton, Gary M. and Hugh Rockoff. *History of the American Economy.* South-Western College Publishers, 2013.

Wicker, Elmus. *Federal Reserve Monetary Policy, 1917–1933.* New York: Random House, 1966.

Wimer, Christopher, Liana Fox, Irwin Garfinel, Neeraj Kaushal, and Jane Waldfogel. 2013. "Trends in Poverty with an Anchored Supplemental Poverty Measure." Columbia Population Research Center Working Paper.