Public Finance
and Public Policy

Jonathan Gruber
Massachusetts Institute of Technology

Worth Publishers
Theoretical Tools of Public Finance

Life is going well. After graduating at the top of your college class, you have parlayed your knowledge of public finance into an influential job with your state's Department of Health and Human Services, which oversees, among other things, the Temporary Assistance for Needy Families (TANF) program. This program provides cash payments to single mothers whose income is below a specified level.

Your new job thrusts you into the middle of a debate between the state's governor and the head of your department, the secretary for Health and Human Services. The governor believes that a major problem with the TANF program is that, by only providing income to very low income single mothers, it encourages them to stay at home rather than go to work. To provide incentives for these mothers to work, the governor wants to cut back on these cash benefits. The secretary of the department disagrees. He thinks that single mothers who are home with their children are incapable of finding jobs that pay a wage high enough to encourage them to work. In his view, if the state cuts the cash payments, it will simply penalize those single mothers who are staying home.

The secretary turns to you to inform this debate by assessing the extent to which cutting cash benefits to low-income single mothers will encourage them to work, and by evaluating the net welfare implications for the state if these benefits are cut. Such an evaluation will require that you put to work the economics tools that you have learned in your introductory and intermediate courses. These tools come in two flavors. First are the theoretical tools, the set of tools designed to understand the mechanics behind economic decision making. The primary theoretical tools of economists are graphical and mathematical. The graphical tools, such as supply and demand diagrams and indifference curve/budget constraint graphs, are typically all that you need to understand the key points of theory, but mathematical expositions can also help to illustrate the subtleties of an argument. In the main body of this book, we rely almost exclusively on graphical analysis, with parallel mathematical analysis presented in some chapter appendices.
The second flavor is **empirical tools**, the set of tools that allows you to analyze data and answer the questions that are raised by theoretical analysis. Most students in this course will have had much less exposure to empirical tools than theoretical tools. Yet, particularly over the past two decades, empirical tools have become as important as theoretical tools in addressing the problems of public finance, as both the quality of data and the ability to carefully analyze that data have improved dramatically.

In the next two chapters, you will be introduced to the key theoretical and empirical tools that you need for this course. In each chapter, we first provide a general background on the concepts, then apply them to our TANF example. The discussion in this chapter is intimately related to the first two of the four questions of public finance. The theoretical tools we develop here are the central means by which economists assess when the government should intervene or how it might intervene.

The remainder of this book relies heavily on the microeconomics concepts reviewed in this chapter. This chapter does not, however, substitute for an introductory or intermediate microeconomics course. The goal here is to refresh your understanding of the important concepts that you need to undertake theoretical public finance, not to teach them to you for the first time. If the material in this chapter is very unfamiliar, you may want to supplement this text with a more detailed microeconomics text.

### 2.1 Constrained Utility Maximization

The core of theoretical analysis in public finance is the assumption that individuals have well-defined **utility functions**, a mathematical mapping of individual choices over goods into their level of well-being. Economists assume that individuals then undertake **constrained utility maximization**, maximizing their well-being (utility) subject to their available resources. Armed with this assumption, economists then develop **models**, mathematical or graphical representations of reality, to show how constrained utility maximization leads people to make the decisions that they make every day. These models have two key components: the individual’s preferences over all possible choices of goods and her **budget constraint**, the amount of resources with which she can finance her purchases. The strategy of economic modelers is then to ask: Given a budget constraint, what bundle of goods makes a consumer best off?

We can illustrate how consumers are presumed to make choices in four steps. First, we discuss how to model preferences graphically. Then, we show how to take this graphical model of preferences and represent it mathematically with a **utility function**. Third, we model the budget constraints that individuals face. Finally, we show how individuals maximize their utility (make themselves as well off as possible) given their budget constraints.
Preferences and Indifference Curves

In modeling people's preferences, we are not yet imposing any budget constraints; we are simply asking what people prefer, ignoring what they can afford. Later, we will impose budget constraints to round out the model.

Much of the power of the preferences models we use in this course derives from one simple assumption: non-satiation, or "more is better." Economists assume that more of a good is always better than less. This does not mean that you are equally happy with the tenth pizza as you are with the first; indeed, as we learn later, your happiness increases less with each additional unit of a good you consume. Non-satiation simply implies that having that tenth pizza is better than not having it.

Armed with this central assumption, we can move on to graphically represent a consumer's preferences across different bundles of goods. Suppose, for example, that Figure 2-1 represents Andrea's preferences between two goods, CDs (with quantity $Q_C$) and movies (with quantity $Q_M$). Consider three bundles:

- Bundle $A$: 2 CDs and 1 movie
- Bundle $B$: 1 CD and 2 movies
- Bundle $C$: 2 CDs and 2 movies

Let's assume, for now, that Andrea is indifferent between bundles $A$ and $B$, but that she prefers $C$ to either; she clearly prefers $C$ because of the assumption that more is better. Given this assumption, we can map her preferences across the goods. We do so using an indifference curve, a curve that shows all combinations of consumption that give the individual the same amount of utility.

**FIGURE 2-1**

Indifference Curves for Bundles of CDs and Movies. Andrea is indifferent between consuming 2 CDs and 1 movie (point A) or 1 CD and 2 movies (point B), but she prefers 2 CDs and 2 movies (point C) to both. Utility is the same along a given indifference curve; indifference curves farther from the origin represent higher utility levels.
utility (and so among which the individual is indifferent). In this case, Andrea gets the same utility from bundles $A$ and $B$, so they lie on the same indifference curve. Because she gets a higher level of utility from consuming bundle $C$ instead of either $A$ or $B$, bundle $C$ is on a higher indifference curve.

Indifference curves have two essential properties, both of which follow naturally from the more-is-better assumption:

1. Consumers prefer higher indifference curves. Individuals prefer to consume bundles that are located on indifference curves that are farther out from the origin because they represent bundles that have more of both CDs and movies.

2. Indifference curves are always downward sloping. Indifference curves cannot slope upward because that would imply that Andrea is indifferent between a given bundle and another bundle that has more of both CDs and movies, which violates the more-is-better assumption.

A great example of indifference curve analysis is job choice. Suppose that Sam graduates and is considering two attributes as he searches across jobs: the starting salary and the location of the job. Sam prefers both a higher salary and a higher temperature location because he likes nice weather. We can represent Sam’s preferences using Figure 2-2, which shows the trade-off between salary and weather. Sam has three job choices:

- Bundle $A$: Starting salary of $30,000 in Phoenix, AZ (hot!)
- Bundle $B$: Starting salary of $50,000 in Minneapolis, MN (cold!)
- Bundle $C$: Starting salary of $40,000 in Washington, D.C. (moderate)

Given Sam’s preferences, it may be that he is indifferent between bundles $A$ and $B$—that is, the higher starting salary in Minneapolis is enough to com-

![Figure 2-2](image)

**Figure 2-2**

*Indifference Curve Analysis of Job Choice* - In choosing a job, Sam trades off the two things he cares about, salary and average temperature. On $IC_1$, he is indifferent between a job in Minneapolis, with a high salary and a low average temperature, and one in Phoenix, with a lower salary and a higher average temperature. However, as indicated by its position on $IC_2$, he prefers a job in Washington, D.C., with average salary and average temperature.
pensate him for the much colder weather. But he may prefer C to either: the salary in Washington is higher than in Phoenix and the weather is much better than in Minneapolis. Compromising on salary and location leaves Sam better off than choosing an extreme of one or the other in this example.

Utility Mapping of Preferences
Underlying the derivation of indifference curves is the notion that each individual has a well-defined utility function. A utility function is some mathematical representation \( U = f(X_1, X_2, X_3, \ldots) \), where \( X_1, X_2, X_3, \text{ and so on are the goods consumed by the individual and} f \text{ is some mathematical function that describes how the consumption of those goods translates to utility. This mathematical representation allows us to compare the well-being associated with different levels of goods consumption.}

For example, suppose that Andrea's utility function over CDs and movies is \( U = \sqrt{Q_C} \times Q_M \). With this function, she would be indifferent between 4 CDs and 1 movie, 2 CDs and 2 movies, and 1 CD and 4 movies because each of these bundles would deliver a utility level of 2. But she would prefer 3 CDs and 3 movies to any of these bundles, since this would give her a utility level of 3.

Marginal Utility The key concept for understanding consumer preferences is marginal utility, or the additional increment to utility from consuming an additional unit of a good. The utility function described exhibits the important principle of diminishing marginal utility: the consumption of each additional unit of a good makes an individual less happy than the consumption of the previous unit. To see this, Figure 2-3 graphs the marginal utility, the increment to utility from each additional movie seen, holding the number of CDs

**FIGURE 2-3**

<table>
<thead>
<tr>
<th>Quantity of movies, ( Q_M )</th>
<th>Marginal utility (( Q_C = 2 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>First movie</td>
<td>1.41</td>
</tr>
<tr>
<td>Second movie</td>
<td>0.59</td>
</tr>
<tr>
<td>Third movie</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Diminishing Marginal Utility
Holding the number of CDs constant at 2, with a utility function of \( U = \sqrt{Q_C} \times Q_M \), each additional movie consumed raises utility by less and less.
constant at 2. When Andrea moves from seeing 0 movies to seeing 1 movie, her utility rises from 0 to $\sqrt{2} = 1.41$. Thus, the marginal utility of that first movie is 1.41. When she moves from seeing one movie to seeing a second movie, her utility rises to $\sqrt{4} = 2$. The consumption of the second movie has increased utility by only 0.59, a much smaller increment than 1.41. When she sees a third movie, her utility rises to only $\sqrt{6} = 2.45$, for an even smaller increment of 0.45. With each additional movie consumed, utility increases, but by ever smaller amounts.

Why does diminishing marginal utility make sense? Consider the example of movies. There is almost always one particular movie that you want to see the most, then one which is next best, and so on. So you get the highest marginal utility from the first movie you see, less from the next, and so on. Similarly, think about slices of pizza: when you are hungry, you get the highest increment to your utility from the first slice; by the fourth or fifth slice, you get much less utility per slice.

**Marginal Rate of Substitution** Armed with the concept of marginal utility, we can now describe more carefully exactly what indifference curves tell us about choices. The slope of the indifference curve is the rate at which a consumer is willing to trade off the good on the vertical axis for the good on the horizontal axis. This rate of trade-off is called the **marginal rate of substitution** (MRS). In this example, the MRS is the rate at which Andrea is willing to trade CDs for movies. As she moves along the indifference curve from more CDs and fewer movies to fewer CDs and more movies, she is trading CDs for movies. The slope of the curve tells Andrea the rate of trade that leaves her indifferent between various bundles of the two goods.

The key feature of the MRS is that it is **diminishing**. We can see this by graphing the indifference curves that arise from the assumed utility function $U = \sqrt{Q_C} \times Q_M$. As Figure 2-4 shows, Andrea is indifferent between 1 movie and 4 CDs, 2 movies and 2 CDs, and 4 movies and 1 CD. Along any segment of this indifference curve, we can define an MRS. For example, moving from 4 CDs and 1 movie to 2 CDs and 2 movies, the MRS is $-2$; she is willing to give up 2 CDs to get 1 movie. Moving from 2 CDs and 2 movies to 1 CD and 4 movies, however, the MRS is $-\frac{1}{2}$; she is willing to give up only 1 CD to get 2 movies.

The slope of the indifference curve changes because of diminishing MRS. When Andrea is seeing only 1 movie, getting to see her second-choice movie is worth a lot to her so she is willing to forgo 2 CDs for that movie. But, having seen her second-choice movie, getting to see her third- and fourth-choice movies isn't worth so much, so she will only forgo 1 CD to see them. Thus, the principle of diminishing MRS is based on the notion that as Andrea has more and more of good $A$, she is less and less willing to give up some of good $B$ to get additional units of $A$.

Since indifference curves are graphical representations of the utility function, there is a direct relationship between the MRS and utility: the MRS is the ratio of the marginal utility for movies to the marginal utility for CDs:

$$MRS = -\frac{MU_M}{MU_C}$$
That is, the MRS shows how the relative marginal utilities evolve over the indifference curve: as Andrea moves down the curve, the MU of CDs rises and that of movies falls. Remember that higher quantity implies lower marginal utility, by the principle of diminishing marginal utility. As Andrea moves down the indifference curve, getting more movies and fewer CDs, the marginal utility of CDs rises, and the marginal utility of movies falls, lowering the MRS.

Budget Constraints

If the fundamental principle of consumer choice is that more is better, what keeps folks from simply binging on everything? What stops them is their limited resources, or their budget constraint, a mathematical representation of the combination of goods they can afford to buy given their incomes. For the purposes of this discussion, we make the simplifying assumption that consumers spend all their income; there is no savings. In Chapter 22, we discuss the implications of a more realistic model where individuals can save and borrow, but for now we assume that all income is spent in the period in which it is received. Moreover, for the purposes of this example, let's assume that Andrea spends her entire income on CDs and movies.

Given these assumptions, Andrea's budget constraint is represented mathematically by \( Y = P_C Q_C + P_M Q_M \), where \( Y \) is her income, \( P_C \) and \( P_M \) are the prices of CDs and movies, and \( Q_C \) and \( Q_M \) are the quantities of CDs and movies she buys. That is, this expression says that her expenditures on CDs and on movies add up to be her total income.

Marginal Rates of Substitution • With a utility function of \( U = \sqrt{Q_C \times Q_M} \), MRS diminishes as the number of movies consumed increases. At 4 CDs and 1 movie, Andrea is willing to trade 2 CDs to get a movie (MRS = -2). At 2 CDs and 2 movies, Andrea is willing to trade 1 CD to get 2 movies (MRS = -\( \frac{1}{2} \)).
Graphically, the budget constraint is represented by the line $AB$ in Figure 2-5. The horizontal intercept is the number of movies that Andrea can buy if she purchases no CDs, and the vertical intercept is the number of CDs she can buy if she goes to no movies, and the slope of the budget constraint is the rate at which the market allows her to trade off CDs for movies. This rate is the negative of the price ratio $P_M/P_C$: each extra movie that she buys, holding income constant, must lower the number of CDs that she can buy by $P_M/P_C$.

Figure 2-5 illustrates the budget constraint for the case when $Y = $96, $P_C = $16 and $P_M = $8. At this income and these prices, Andrea can purchase 12 movies or 6 CDs, and each CD she buys means that she can buy two fewer movies. The slope of the budget constraint is the rate at which she can trade CDs for movies in the marketplace, $P_M/P_C = -3 = -\frac{3}{2}$.

**Quick Hint** Our discussion thus far has been couched in terms of “trading CDs for movies” and vice versa. In reality, however, we don’t directly trade one good for another; instead, we trade in a market economy, in which CDs and movies are purchased using dollars. The reason we say “trading CDs for movies” is because of the central economics concept of opportunity cost, which says that the cost of any purchase is the next best alternative use of that money. Thus, given a fixed budget, when a person buys a CD, he forgives the opportuni-
Putting It All Together: Constrained Choice

Armed with the notions of utility functions and budget constraints, we can now ask: What is the utility-maximizing bundle that consumers can afford? That is, what bundle of goods makes consumers best off, given their limited resources?

The answer to this question is shown in Figure 2-6. This figure puts together the indifference curves corresponding to the utility function \( U = \sqrt{Q_c \times Q_m} \) shown in Figure 2-4 with the budget constraint shown in Figure 2-5. In this framework, we can rephrase our question: What is the highest indifference curve that an individual can reach given a budget constraint? The answer is the indifference curve, \( IC_3 \), that is tangent to the budget constraint: this is the farthest-out indifference curve that is attainable, given Andrea's income and market prices. In this example, Andrea makes herself as well off as possible by choosing to consume 6 movies and 3 CDs (point A). That combination of goods maximizes Andrea's utility, given her available resources and market prices.

**Figure 2-6**

Constrained Optimization • Given a utility function of \( U = \sqrt{Q_c \times Q_m} \), an income of $96, and prices of CDs and movies of $16 and $8, respectively, Andrea's optimal choice is 3 CDs and 6 movies (point A). This represents the highest indifference curve she can reach, given her resources and market prices. She can also afford points such as B and C, but they leave her on a lower indifference curve (\( IC_1 \) instead of \( IC_3 \)).
The key to understanding this outcome is the marginal decision Andrea makes to consume the next movie. The benefit to her of consuming another movie is the marginal rate of substitution, the rate at which she is willing to trade off CDs for movies. The cost to her of making this trade is the price ratio, the rate at which the market allows her to trade CDs for movies. Thus, the optimal choice is the one at which:

$$MRS = -\frac{MU_M}{MU_C} = -\frac{P_M}{P_C}$$

At the optimum, the ratio of marginal utilities equals the ratio of prices. The rate at which Andrea is willing to trade off one good for the other is equal to the rate at which the market will let her carry out that trade.

One way to demonstrate that this is the optimal choice is to show that she is worse off with any other choice. Consider point $B$ in Figure 2-6. At that point, the slope of the indifference curve $IC_1$ is higher than the slope of the budget constraint; that is, the $MRS$ is greater than the price ratio. This means that Andrea’s marginal utility of movies, relative to CDs, is higher than the ratio of the price of movies to the price of CDs. Because the $MRS$ is the rate at which Andrea is willing to trade CDs for movies and the price ratio is the rate that the market is charging for such a trade, Andrea is willing to give up more CDs for movies than the market requires. She can make herself better off by reducing her CD purchases and increasing her movie purchases, as happens when she moves from $B$ to $A$.

Now consider point $C$ in Figure 2-6. At this point, the slope of the indifference curve $IC_1$ is less than the slope of the budget constraint; that is, the $MRS$ is lower than the price ratio. Relative to point $B$, Andrea now cares much less about movies and more about CDs, since she is now consuming more movies and fewer CDs, and marginal utility diminishes. At point $C$, in fact, she is willing to give up fewer CDs for movies than the market requires. So she can make herself better off by increasing her CD purchases and reducing her movie purchases, as happens when she moves from $C$ to $A$. Whenever a consumer is at a point where the indifference curve and budget constraint are not tangent, she can make herself better off by moving to a point of tangency.

Quick Hint: Marginal analysis, the consideration of the costs and benefits of an additional unit of consumption or production, is a central concept in modeling an individual’s choice of goods and a firm’s production decision. All optimization exercises in economics are like climbing a hill on a very cloudy day. At any given point, you don’t know yet whether you are at the top, but you do know if you are heading up or heading down. If you are heading up, then you must not yet be at the top; but if you are heading down, then you must have passed the top.

It is the same when you are maximizing your utility (or your firm’s profits). Consider the mountain as your decision on how many movies to buy, and the top as the optimal number of movies given your preferences and budget constraint. Starting from any number of CDs and movies, you consider whether the next movie
has a benefit (MRS) greater than its cost (price ratio). If the benefit exceeds the cost of that next movie, then the next step is upward, and you buy the movie and continue up the optimization mountain. If the benefit is below the cost, then the next step is downward, and you realize that you need to go backward (buy fewer movies) to get back to the top. Only when the benefit equals the cost of the next unit do you realize you are at the top of the mountain.

The Effects of Price Changes: Substitution and Income Effects
The key result from the constrained choice analysis is that \( MU_M / MU_C = P_M / P_C \): Andrea consumes movies and CDs until the ratio of the marginal utility of movies to CDs equals the ratio of their prices. An implication of this result is that when the relative price of a good, such as movies in our example, rises (i.e., \( P_M / P_C \) rises), then the relative quantity of that good demanded falls. This is because, for the equality previously described to hold, when \( P_M / P_C \) rises, then \( MU_M / MU_C \) must also rise. For \( MU_M / MU_C \) to rise, the quantity of movies relative to CDs must fall (since the marginal utility of any good falls as the quantity consumed of that good rises).

This point is illustrated graphically in Figure 2-7. We have already shown that, with an income of $96, and prices of $16 for CDs and $8 for movies, Andrea chooses 6 movies and 3 CDs at point A, the point at which \( BC_1 \) and \( IC_1 \) are tangent. If the price of movies were to rise to $16, for example, the budget constraint would become steeper; it rotates inward from \( BC_1 \) to \( BC_2 \). Because the price of

![Figure 2-7](image-url)

**Figure 2-7**

Substitution and Income Effects - When the price of movies increases, it has two effects. First, holding utility constant, there is a substitution effect, which causes Andrea to demand fewer movies since they are relatively more expensive (moving from point A to point B). Second, holding relative prices constant, there is an income effect, which causes her to demand fewer movies because she is poorer (moving from point B to point C).
CDs hasn’t changed, Andrea can still buy 6 CDs with her $96 income (the vertical intercept of the budget constraint), but because the price of movies has risen to $16, she can only now buy 6 ($96/$16) movies (the horizontal intercept). The slope of the budget constraint rises from $-\frac{3}{2}$ to $-1$, as illustrated by $BC_2$.

With a steeper budget constraint, Andrea can no longer afford to buy the combination at point $A$. The optimal combination becomes point $C$, the point at which $BC_2$ is tangent to a lower indifference curve, $IC_2$. At this point, Andrea can buy 3 CDs, but she can now only buy 3 movies, instead of the 6 she could buy at point $A$. The quantity of movies she demands has fallen, because their price has gone up. She is also now worse off: her budget set, or the set of possible choices she can make given her income, has been restricted (since the budget constraint moved inward from $BC_1$ to $BC_2$). The quantity of CDs she demands has remained constant, but this is simply because of the assumed mathematical form of the utility function; in general, the number of CDs she demands would fall as well.

**Income and Substitution Effects** Imagine that the government could somehow insulate Andrea from the utility she loses when prices rise; that is, suppose the government was somehow able to compensate her enough that she could stay on the same indifference curve ($IC_1$ in our example), even with the new set of prices. Would this mean that the price change will have no effect on her choices? No, it doesn't, because she would still like to choose a different bundle of CDs and movies at this new set of prices.

Figure 2-7 illustrates this point. Despite this price change, the government can hold Andrea’s utility constant at these new prices by giving her a budget constraint $BC_e$, which is parallel to $BC_2$ but tangent to the same indifference curve $IC_1$ that corresponds to her original choice. Graphically, the budget constraint has steepened, but Andrea is on the same indifference curve (the same level of utility). Andrea chooses the bundle represented by point $B$: because movies are relatively more expensive, she chooses to consume fewer movies (4.24) and more CDs (4.24). This effect of a price change is called the **substitution effect**: holding utility constant, a relative rise in the price of a good will always cause a consumer to choose less of that good.

In the real world, when prices rise there is no government agency to hold utility constant. This price rise therefore leads to a second effect on demand: Andrea is now effectively poorer because she has to pay higher prices for movies. She is not poorer in an income sense (her income remains at $96), but she is poorer in a real sense because her $96 can buy fewer goods (in particular, fewer movies). This is the **income effect** of a price change: a rise in any price will make the consumer effectively poorer, causing her to choose less of all goods. The quantity demanded falls because Andrea can buy fewer goods with her income.

---

[substitution effect] Holding utility constant, a relative rise in the price of a good will always cause an individual to choose less of that good.

[income effect] A rise in the price of a good will typically cause an individual to choose less of all goods because her income can purchase less than before.

[normal goods] Goods for which demand increases as income rises.

[inferior goods] Goods for which demand falls as income rises.

We say "typically" here because, in theory, demand for goods can go up or down as your income increases. Most goods are normal goods, for which demand increases as income rises, but some goods are inferior goods, for which demand falls as income rises. Inferior goods are those with better substitutes that might be demanded as income rises. For example, potatoes might be heavily consumed by the poor, but as income rises, fewer potatoes will be consumed as people substitute other goods, such as meat, which they can now afford.
We can measure this income effect by the change from the government-supported budget constraint $BC_0$, to the new budget constraint $BC_2$. This change represents the restriction in Andrea's opportunity set, at the new prices. Since she is poorer, she chooses fewer of all goods, including both movies and CDs, at point C. In this case, the income effect reinforces the substitution effect for movies: both cause the quantity of movies she demands to fall.\(^2\) To sum up, when the price of one good increases relative to another, you choose less of that good for two reasons: because it is relatively less attractive (the substitution effect) and because you are effectively poorer (the income effect).

2.2 Putting the Tools to Work: TANF and Labor Supply Among Single Mothers

In your new position with the state government, you have now reviewed the theoretical concepts necessary to address the concerns of the secretary and the governor. Having reviewed these theoretical concepts, let's turn to the question posed at the start of the chapter: Will reducing TANF benefits increase the labor supply of single mothers? To answer this question, we can apply the tools of utility maximization to the analysis of the labor supply decision.

The TANF program was created in 1996 by a major overhaul of the cash welfare system in the United States. The cash welfare system distributes money from taxpayers to low-income families (as described in much more detail in Chapter 17). TANF provides a monthly support check to families with incomes below a threshold level that is set by each state. In the state of New Mexico, for example, a single mother with two children and no other source of income will receive a monthly check for $389.\(^3\) These checks are largely targeted to single-female-headed households with children, since these families are viewed as having the worst prospects for making a living on their own.

Suppose that Joelle is a single mother who spends all of her earnings and TANF benefits on food for her and her children. By working more hours, she can earn more money for food, but there is a cost to work: she has less time at home with her children (or less time to spend on her own leisure). Suppose that she would prefer time at home to time at work; that is, suppose that leisure is a normal good. With these preferences, more work makes Joelle worse off, but it allows her to buy more food.

How does Joelle decide on the optimal amount of labor to supply? To answer this question, we return to the utility maximization framework, but with one twist relative to the decision to purchase CDs and movies. In that case, we were considering two goods. Now, the single mother is considering one good (food consumption) and one “bad” (labor, since we assume she

\(^2\) They have canceling effects on the demand for CDs, however, which is why demand for CDs doesn't change in Figure 2-7. Note also that if goods are inferior, the income effect would offset the substitution effect, rather than reinforcing it.

\(^3\) U.S. Department of Health and Human Services (2003), Table 12-2.
When the TANF guarantee is reduced to $3,000, Sarah chooses to reduce her leisure since she is now poorer (the income effect), moving to point B on the new budget constraint. At that point, she takes only 1,655 hours of leisure per year, works 345 hours, and earns $3,450. For this mother, the governor is right; the reduction in TANF guarantee has raised her labor supply from 90 hours to 345 hours. Note that because Sarah’s TANF benefits are reduced by half her earnings, her TANF benefits are now $3,000 – 0.5 × $3,450 = $1,275. Thus, her total budget is $4,275; her consumption has fallen by $1,125 from the days of the higher TANF guarantee ($5,450 – 4,275 = $1,125). Her consumption has not fallen by the full $2,000 cut in the guarantee because she has compensated for the guarantee reduction by working harder.

Figure 2-11 illustrates the case of a different single mother, Naomi, with a utility function $U = 75 \times \ln(C) + 300 \times \ln(L)$. Naomi puts a much larger weight on leisure relative to consumption, when compared to Sarah. (Her indifference curves are steeper, indicating that a larger increase in consumption is required to compensate for any reduction in leisure.) For Naomi, the optimal choice when the TANF guarantee is $5,000 is to not work at all; she consumes 2,000 hours of leisure and $5,000 of food (point A). When the guarantee is reduced to $3,000, this mother continues not to work, and just lets her consumption fall to $3,000. That is, she cares so much more about leisure than about consumption that she won’t supplement her TANF guarantee with earnings even at the lower guarantee level. For this mother, the secretary is right; the reduction in TANF guarantee has had no effect on labor supply, it has simply cut her level of food consumption.

Thus, theory alone cannot tell you whether this policy change will increase labor supply, or by how much. Theoretically, labor supply could rise, but it might not. To move beyond this uncertainty, you will have to analyze available data on single mother labor supply, and the next chapter presents the empirical methods for doing so. From these various methods, you will conclude that the governor is right; there is strong evidence that cutting TANF benefits will increase labor supply.

2.3 Equilibrium and Social Welfare

The disagreement we have been discussing is over whether the labor supply of single mothers will rise or not when TANF benefits are cut. As a good public finance economist, however, you know not to stop there. What really should matter to the governor and to the secretary of your department is not a simple fact about whether labor supply of single mothers rises or falls. What should matter is the normative question (the analysis of what should be): Does this policy change make society as a whole better off or not?

To address this question, we turn to the tools of normative analysis, welfare economics. Welfare economics is the study of the determinants of well-being, or welfare, in society. To avoid confusion, it is important to recall that

welfare economics The study of the determinants of well-being, or welfare, in society.
the term "welfare" is also used to refer to cash payments (such as those from the TANF program) to low-income single families. Thus, when referring to cash payments in this chapter, we will use the term TANF; our use of the term "welfare" in this chapter refers to the normative concept of well-being.

We discuss the determination of welfare in two steps. First, we discuss the determinants of social efficiency, or the size of the economic pie. Social efficiency is determined by the net benefits that consumers and producers receive as a result of their trades of goods and services. We develop the demand and supply curves that measure those net benefits, show how they interact to determine equilibrium, and then discuss why this equilibrium maximizes efficiency. We then turn to a discussion of how to integrate redistribution, or the division of the economic pie, into this analysis so we can measure the total well-being of society, or social welfare. In this section, we discuss these concepts with reference to our earlier example of Andrea choosing between movies and CDs; we then apply these lessons to a discussion of the welfare implications of changes in TANF benefits.

**Demand Curves**

Armed with our understanding of how consumers make choices, we can now turn to understanding how these choices underlie the demand curve, the relationship between the price of a good or service and the quantity demanded. Figure 2-12 shows how constrained choice outcomes are translated into the demand curve for movies for Andrea. In panel (a), we vary the price of movies, which changes the slope of the budget constraint (which is determined by the ratio of movie to CD prices). For each new budget constraint, Andrea's optimal choice remains the tangency of that budget constraint with the highest possible indifference curve.

For example, we have already shown that given her income of $96, at a price of $16 for CDs and $8 for movies, Andrea will choose 6 movies and 3 CDs (point A on BC). An increase in the price of movies to $12 will steepen the budget constraint, with the slope rising from $-\frac{1}{2}$ to $-\frac{4}{3}$, as illustrated by BC. This increase in price will reduce the quantity of movies demanded, so that she chooses 3 CDs and 4 movies (point B on BC). A decrease in the price of movies to $6 will flatten the budget constraint, with the slope falling from $-\frac{1}{2}$ to $-\frac{1}{3}$, as illustrated by BC. This decrease in price will increase the quantity of movies demanded, and she will now choose to buy 3 CDs and 8 movies (point C on BC).

Using this information, we can trace out the demand curve for movies, which shows the quantity of a good or service demanded by individuals at each market price. The demand curve for movies, shown in panel (b), maps the relationship between the price of movies and the quantity of movies demanded.

**Elasticity of Demand** A key feature of demand analysis is the elasticity of demand, the percentage change in quantity demanded for each percentage change in prices.

\[
\varepsilon = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in price}} = \frac{\Delta Q/Q}{\Delta P/P}
\]
For example, when the price of movies rises from $8 to $12, the number of movies purchased falls from 6 to 4. So a 50% rise in price leads to a 33% reduction in quantity purchased, for an elasticity of −0.666.

There are several key points to make about elasticities of demand:

- They are typically negative, since quantity demanded typically falls as price rises.
- They are typically not constant along a demand curve. So, in our previous example, the price elasticity of demand is −0.666 when the price of movies rises, but is −1.32 when the price of movies falls (a 25% reduction in price from $8 to $6 leads to a 33% increase in demand from 6 to 8 movies).
- A vertical demand curve is one for which the quantity demanded does not change when price rises; in this case, demand is perfectly inelastic.
- A horizontal demand curve is one where quantity demanded changes infinitely for even a very small change in price; in this case, demand is perfectly elastic.
- Finally, the example here is a special case in which the demand for CDs doesn’t change as the price of movies changes. The effect of one good’s
Supply curve. A curve showing the quantity of a good firms are willing to supply at each price.

Marginal productivity. The impact of a one unit change in any input, holding other inputs constant, on the firm's output.

Marginal cost. The incremental cost to a firm of producing one more unit of a good.

prices on the demand for another good is the cross-price elasticity, and with the particular utility function we are using here, that cross-price elasticity is zero. Typically, however, a change in the price of one good will affect demand for other goods as well.

Supply Curves

The discussion thus far has focused on consumers and the derivation of demand curves. This tells about only one side of the market, however. The other side of the market is represented by the supply curve, which shows the quantity supplied of a good or service at each market price. Just as the demand curve is the outcome of utility maximization by individuals, the supply curve is the outcome of profit maximization by firms.

The analysis of firms’ profit maximization is similar to that of consumer utility maximization. Just as individuals have a utility function that measures the impact of goods consumption on well-being, firms have a production function that measures the impact of firm input use on firm output levels. For ease, we typically assume that firms have only two types of inputs, labor (workers) and capital (machines, buildings). Consider a firm that produces movies. This firm’s production function may take the form $q = \sqrt{K \times L}$, where $q$ is the quantity of movies produced, $K$ is units of capital (such as studio sets), and $L$ is units of labor (such as hours of acting time employed).

The impact of a one-unit change in an input, holding other inputs constant, on the firm’s output is the marginal productivity of that input. Just as the marginal utility of consumption diminishes with each additional unit of consumption of a good, the marginal productivity of an input diminishes with each additional unit of the input used in production; that is, production generally features diminishing marginal productivity. For this production function, for example, holding $K$ constant, adding additional units of $L$ raises production by less and less, just as with the utility function (of this same form), holding CDs constant, consuming additional movies raised utility by less and less.  

This production function dictates the cost of producing any given quantity as a function of the prices of inputs and the quantity of inputs used. The total costs of production, $TC$, are determined by $TC = rK + wL$, where $r$ is the price of capital (the rental rate) and $w$ is the price of labor (the wage rate). For day-to-day decisions by the firm, the amount of capital is fixed, while the amount of labor can be varied. Given this assumption, we can define the marginal cost, or the incremental cost to producing one more unit, as the wage rate times the amount of labor needed to produce one more unit.

For example, consider the production function just described, and suppose that the firm is producing 2 movies using 1 unit of capital and 4 units of labor.

---

4 A good way to see this intuition is to consider digging a hole with one shovel. One worker can make good progress. Adding a second worker probably increases the progress, since the workers can relieve each other in shifts, but it is unlikely that progress doubles. Adding a third worker raises progress even less. By the time there are four or five workers, there is very little marginal productivity to adding additional workers, given the fixed capital (one shovel).
Now, holding the amount of capital fixed, it wants to produce 3 movies. To do so, it will have to increase its use of labor by 5 units (to 9 total units). If the wage rate is $1 per unit, then the marginal cost of raising production from 2 to 3 movies is $5.

The key point of this discussion is that diminishing marginal productivity implies rising marginal costs. To produce a fourth movie would require an increase in labor of 7 units, at a cost of $7; to produce a fifth movie would cost $9. Since each additional unit of production means calling forth less and less productive labor, at the same wage rate, the costs of that production are rising.

Recall that the goal of the firm is to maximize its profit, the difference between revenues and costs. Profit is maximized when the revenue from the next unit, or the marginal revenue, equals the cost of producing that next unit, the marginal cost. In a competitive industry, the revenue from any unit is the price the firm obtains in the market. Thus, the firm’s profit maximization rule is to produce until price equals marginal cost.

We can see this through the type of “hill-climbing” exercise proposed in the Quick Hint on pages 32–33. Suppose the price of movies in the market is $8, the cost of capital is $1 per unit, the cost of labor is $1 per unit, and the firm has one unit of capital. Then, if the firm produces one movie, it will need to use 1 unit of labor, so that total costs are $2. Because revenues on that first unit are $8, it should clearly produce that first movie. To produce a second movie, the firm will need to use 4 units of labor, or an increase of 3 units of labor. Thus, the marginal cost of that second unit is $3, but the marginal revenue (price) is $8, so the second movie should be produced. For the third movie, the marginal cost is $5, as just noted, which remains below price.

But now imagine the firm is producing four movies and is deciding whether to produce a fifth. Producing the fifth movie will require an increase in labor input from 16 to 25 units, or an increase of 9 units. This will cost $9. But the price that the producer gets for this movie is only $8. As a result, producing that fifth unit will be a money loser, and the firm will not do it. Thus, profit maximization dictates that the firm produce until its marginal costs (which are rising by assumption of diminishing marginal productivity) reach the price.

Profit maximization is the source of the supply curve, the relationship between the price and how much producers will supply to the market. At any price, we now know that producers will supply a quantity such that the marginal cost equals that price. Thus, the marginal cost curve is the firm’s supply curve, showing the relationship between price and quantity. As quantity rises, and marginal costs rise, the firm will require higher and higher prices to justify producing additional units.

**Equilibrium**

We have discussed the source of individual demand curves (utility maximization) and firm supply curves (profit maximization). To undertake welfare analysis we need to translate these concepts to their counterparts at the level of the market, the arena in which demanders and suppliers actually interact.
Market Outcome • The supply and demand curves for movies intersect at the equilibrium point $E$, where both consumers and suppliers are satisfied with price and quantity.

(such as the supermarket or a Web site). To do so, we add up the demands of each individual who is demanding goods in this market, and the supplies of each firm that is supplying goods in this market. We horizontally sum these curves. That is, at each price, we add up all the quantities willing to be purchased at that price by demanders to obtain market-level demand, and all the quantities willing to be supplied at that price by suppliers to obtain market-level supply. The result is the market-level supply and demand curves shown in Figure 2-13.

The market-level supply and demand curves interact to determine the market equilibrium, the price and quantity pair that will satisfy both demand and supply. This point occurs at the intersection of the supply and demand curves, such as point $E$ in Figure 2-13. Given the equilibrium price $P_E$, demanders will demand the equilibrium quantity, $Q_E$, and suppliers will be willing to supply that equilibrium quantity. The competitive market equilibrium represents the unique point at which both consumers and suppliers are satisfied with price and quantity.

Social Efficiency
Armed with the analysis thus far, we are now ready to take the final step: to measure social efficiency, or the size of the pie. Social efficiency represents the net gains to society from all trades that are made in a particular market, and it consists of two components: consumer and producer surplus.

Consumer Surplus The gain to consumers from trades in a market for consumers is consumer surplus, the benefit that consumers derive from con-
summing a good above and beyond what they paid for the good. Once we know the demand curve, consumer surplus is easy to measure, because each point on a demand curve represents the consumer's willingness to pay for that quantity.

Panel (a) of Figure 2-14 presents a graphical representation of consumer surplus in the movie market; the shaded area below the demand curve and above the equilibrium price $P_E$ (area $WZX$). This area is consumer surplus because these are units where the willingness to pay (represented by the demand curve) is higher than the amount actually paid, $P_E$. Consumer surplus is largest on the very first unit, since this represents the consumer who most wanted the good. (He is willing to buy the good at a very high price.) For that first unit, consumer surplus is equal to the distance $WX$ on the vertical axis. Consumer surplus then falls as additional consumers derive less and less marginal utility from the good. Finally, for the consumer whose demand (willingness to pay) equals the price (at point $Z$), consumer surplus is zero.

Consumer surplus is determined by two factors: the market equilibrium price and the elasticity of demand. Panel (b) of Figure 2-14 shows the case of a good with very inelastic demand (that is, where quantity demanded is not very sensitive to prices), such as basic foods for a low-income community. In this case, the demand curve is more vertical so the consumer surplus is a very large area. Consumer surplus is large because inelastic demand arises from a lack of good substitutes, so that consumers get enormous surplus out of consuming that particular good. Panel (c) of Figure 2-14 shows the case of a good

\[\text{**FIGURE 2-14**}\]

\[\text{Consumer Surplus • The consumer surplus is the area below the demand curve and above the equilibrium market price, the shaded area } WZX \text{ in all three panels of this graph. This represents the value to consumers of consuming goods above and beyond the price paid for those goods. As demand becomes more inelastic, consumer surplus rises; as demand becomes more elastic, consumer surplus falls.}\]
Producer surplus: The benefit at producers derive from selling a good, above and beyond the cost of producing that good.

with very elastic demand (that is, where quantity demanded is very sensitive to prices), such as going to the movies. In this case, the demand curve is nearly horizontal, so that consumer surplus is a very small area. This is because elastic demand arises from the availability of very good substitutes. Consumers don't derive very much surplus from consuming a good for which there are close substitutes.

**Producer Surplus** Consumers aren't the only ones who derive a surplus from market transactions. There is also a welfare gain to producers, the **producer surplus**, which is the benefit derived by producers from the sale of a unit above and beyond their cost of producing that unit. Like consumer surplus, producer surplus is easy to measure because every point on the supply curve represents the marginal cost of producing that unit of the good. Thus, producer surplus is represented graphically by the area above the supply (marginal cost) curve and below the equilibrium price $P_e$, the shaded area XZY in Figure 2-15. This area is producer surplus because these are units where the market price is above the willingness to supply (the supply curve). Producer surplus is, in effect, the profits made by the producer.

Panels (b) and (c) in Figure 2-15 illustrate the impact on producer surplus of varying the **price elasticity of supply**, the percentage change in supply for each percentage change in market prices. When the price elasticity of supply is very low, so that supply is very inelastic, then the supply curve is more vertical and producer surplus is very large, as in panel (b). When the price elasticity of sup-

---

**FIGURE 2-15**

Panel (a) shows the producer surplus as the area below the equilibrium market price and above the supply curve, the shaded area XZY in all three panels this graph. This represents the profit earned by firms on all units sold at the market price. As supply becomes more inelastic, producer surplus rises; as supply becomes more elastic, producer surplus falls.
Supply is very high so that supply is very elastic, then the supply curve is nearly horizontal and producer surplus is very small, as in panel (c).

**Social Surplus** Total social surplus, also called social efficiency, is the total surplus received by consumers and producers in a market. Figure 2-16 shows the total social surplus for the movie market. The consumer surplus in this market is the shaded area $A + D$, and the producer surplus is the shaded area $B + C + E$. Thus, social surplus for this market is the sum of the shaded areas $A + B + C + D + E$.

**Competitive Equilibrium Maximizes Social Efficiency**

We can use this social surplus framework to illustrate the point known as the **First Fundamental Theorem of Welfare Economics**: the competitive equilibrium, where supply equals demand, maximizes social efficiency. This theorem makes intuitive sense because social efficiency is created whenever a trade occurs that has benefits that exceed its costs. This is true for every transaction to the left of $Q_E$ in Figure 2-16: for each of those transactions, the benefits (willingness to pay, or demand) exceed the costs (marginal cost, or supply).

Doing anything that lowers the quantity sold in the market below $Q_E$ reduces social efficiency. For example, suppose that the government, in an effort to help consumers, restricts the price that firms can charge for movies to $P_R$, which is below the equilibrium price $P_E$. Suppliers react to this restriction by reducing their quantity produced to $Q_R$, the quantity at which the new price, $P_R$, intersects the supply curve: it is the quantity producers are willing to supply at this price. Producer surplus is now area $C$, the total social surplus (social efficiency) is the sum of consumer surplus and producer surplus.

**First Fundamental Theorem of Welfare Economics**

The competitive equilibrium, where supply equals demand, maximizes social efficiency.
area above the supply curve and below price $P_R$. Thus, producer surplus falls by area $B + E$.

On the consumer side, there are two effects on surplus. On the one hand, since a smaller quantity of movies are supplied, consumers are worse off by the area $D$: the movies that are no longer provided between $Q_R$ and $Q_E$ were movies for which consumers were willing to pay more than the cost of production to see the movie, so consumer surplus falls. On the other hand, since consumers pay a lower price for the remaining $Q_R$ movies that they do see, consumer surplus rises by area $B$.

On net, then, society loses surplus equal to the area $D + E$. This area is called **deadweight loss**, the reduction in social efficiency from denying trades for which benefits exceed costs. This part of the social surplus $(D + E)$ has vanished because there are trades that could be made where benefits are greater than costs, but these trades are not being made. Graphically, then, the social surplus triangle is maximized when quantity is at $Q_E$.

**Quick Hint** It is sometimes confusing to know how to draw deadweight loss triangles. The key to doing so is to remember that *deadweight loss triangles point to the social optimum, and grow outward from there*. The intuition is that the deadweight loss from over- or underproduction is smallest right near the optimum (producing one unit too few or one too many isn’t so costly). As production moves farther from this optimum, however, the deadweight loss grows rapidly.

---

**From Social Efficiency to Social Welfare: The Role of Equity**

The discussion thus far has focused entirely on how much surplus there is (social efficiency, the size of the economic pie). Societies usually care not only how much surplus there is but also about how it is distributed among the population. The level of **social welfare**, the level of well-being in a society, is determined both by social efficiency and by the equitable distribution of society’s resources.

Under certain assumptions, efficiency and equity are two separate issues. In these circumstances, society doesn’t have just one socially efficient point, but a whole series of socially efficient points from which it can choose. Society can achieve those different points simply by shifting available resources among individuals and letting them trade freely. Indeed, this is the **Second Fundamental Theorem of Welfare Economics**: society can attain any efficient outcome by a suitable redistribution of resources and free trade.

In practice, however, society doesn’t typically have this nice choice. Rather, as discussed in Chapter 1, society most often faces an **equity-efficiency trade-off**, the choice between having a bigger economic pie and a more fairly distributed pie. Resolving this trade-off is harder than determining efficiency-enhancing government interventions. It raises the tricky issue of
making interpersonal comparisons, or deciding who should have more and who should have less in society.

Typically, we model the government’s equity-efficiency decisions in the context of a social welfare function (SWF). This function maps the set of individual utilities in society into an overall social utility function. In this way, the government can incorporate the equity-efficiency trade-off into its decision making. If a government policy impedes efficiency and shrinks the economic pie, then citizens as a whole are worse off. If, however, that shrinkage in the size of the pie is associated with a redistribution that is valued by society, then this redistribution might compensate for the decrease in efficiency and lead to an overall increase in social welfare.

The social welfare function can take one of a number of forms, and which form a society chooses is central to how it resolves the equity-efficiency trade-off. If the social welfare function is such that the government cares solely about efficiency, then the competitive market outcome will not only be the most efficient outcome, it will also be the welfare-maximizing outcome. In other cases where the government cares about the distribution of resources, then the most efficient outcome may not be the one that makes society best off. Two of the most common specifications of the social welfare function are the utilitarian and Rawlsian specifications.

**Utilitarian SWF** With a utilitarian social welfare function, society’s goal is to maximize the sum of individual utilities:

\[ SWF = U_1 + U_2 + \ldots + U_N \]

The utilities of all individuals are given equal weight, and summed to get total social welfare. This formulation implies that we should transfer from person one to person two as long as the utility gain to person one is greater than the utility loss to person two. In other words, this implies that society is indifferent between one util (a unit of well-being) for a poor person and one util for a rich person.

Is this outcome unfair? No, because the social welfare function is defined in terms of utility, not dollars. With a utilitarian SWF, society is not indifferent between giving one dollar to the poor person and giving one dollar to the rich person; society is indifferent between giving one util to the poor person and one util to the rich person. This distinction between dollars and utility is important because of the diminishing marginal utility of income; richer persons gain a much smaller marginal utility from an extra dollar than poorer persons. With a utilitarian SWF, society is not indifferent between a dollar to the rich and the poor; in general, it wants to redistribute that dollar from the rich (who have a low MU because they already have high consumption) to the poor (who have a low MU). If individuals are identical, and if there is no efficiency cost of redistribution, then the utilitarian SWF is maximized with a perfectly equal distribution of income.

**Rawlsian Social Welfare Function** Another popular form of social welfare function is the Rawlsian SWF, named for the philosopher John Rawls. He
suggested that society's goal should be to maximize the well-being of its worst-off member.\textsuperscript{5} The Rawlsian SWF has the form:

\[ SW = \min(U_1, U_2, \ldots, U_N) \]

Since social welfare is determined by the minimum utility in society, social welfare is maximized by maximizing the well-being of the worst-off person in society.

If individuals are identical, and redistribution does not have efficiency costs, this SWF would call for an equal distribution of income, as does the utilitarian SWF: only when income is equally distributed is society maximizing the well-being of its worst-off member. On the other hand, the utilitarian and Rawlsian SWF do not have the same implications once we recognize that redistribution can entail efficiency costs (and reduce the size of the pie). Suppose all individuals have identical preferences, and equal incomes of $40,000 per year, except for two individuals: Donald, who has income of $1 million per year, and Joe, who has income of $39,999. Now imagine a proposal to tax Donald by $960,000, take $1 of that tax revenue and give it to Joe, and to throw the rest of the money into the ocean. Under a utilitarian SWF this plan would lower social welfare because Donald's utility will fall more from losing $960,000 than Joe's utility will rise from gaining $1. Under a Rawlsian SWF, however, this plan will raise social welfare, since the utility of the worst-off person has increased, and that is all we care about! Thus, in a world of equity-efficiency trade-offs, a Rawlsian SWF will in general suggest more redistribution than will a utilitarian SWF.

\subsection*{2.4 Welfare Implications of Benefit Reductions: The TANF Example Continued}

The equilibrium and social welfare tools developed in Section 2.3 can be applied to evaluate the benefits and costs to society of reducing TANF benefits. The benefits are the improvement in efficiency from removing a barrier to labor supply by single mothers, raising their labor supply and the size of the social surplus. (Relying on the empirical evidence discussed in the next chapter, we assume that labor supply increases when benefits fall.) The costs are the reductions in equity that arise from reducing income support to one of the lowest-income groups in our society. The job of public finance economists is to measure these efficiency and equity consequences. The job of policymakers is to trade them off to decide on appropriate policy choices.

Efficiency We can apply the tools of welfare analysis to model the welfare implications of cutting TANF benefits. Figure 2-17 shows the market for labor services by single mothers. The price of labor, the wage ($W$) is on the vertical axis; the amount of hours worked in aggregate in the market ($H$) is on the horizontal axis.

\begin{footnote}
\textsuperscript{5} See Rawls (1971), pp. 152–157, for arguments about why this should be society's goal.
\end{footnote}
19.1

The Three Rules of Tax Incidence

The goal of determining a tax's incidence is to assess who ultimately bears the burden of paying a tax. Economic tax incidence can be described by three basic rules.

The Statutory Burden of a Tax Does Not Describe Who Really Bears the Tax

The first and most important rule of tax incidence is that tax laws do not accurately identify who actually bears the burden of the tax. The statutory incidence of a tax is determined by who pays the tax to the government. For example, the statutory incidence of a tax paid by producers of gasoline is on those very producers. Statutory incidence, however, ignores the fact that markets react to taxation. This market reaction determines the economic incidence of a tax, the change in the resources available to any economic agent as a result of taxation. The economic incidence of any tax is the difference between the individual's available resources before and after the tax has been imposed.

When a tax is imposed on producers in a competitive market, producers will raise prices to some extent to offset this tax burden, and the producers' income will not fall by the full amount of the tax. When a tax is imposed on consumers in a competitive market, the consumers will not be willing to pay as much for the taxed good, so prices will fall, offsetting to some extent the statutory tax burden on consumers. Technically, we can define the tax burden for consumers as

\[
\text{consumer tax burden} = (\text{posttax price} - \text{pretax price}) + \text{tax payments by consumers}.
\]

For producers the tax burden is

\[
\text{producer tax burden} = (\text{pretax price} - \text{posttax price}) + \text{tax payments by producers}.
\]

For example, suppose that tomorrow the federal government levied a 50¢ per gallon tax on gasoline, to be paid by the producers. Will gas producers receive 50¢ less on each gallon they produce as a result of this tax?

To answer this question, we need to consider the impact of the gas tax on the market for gas, as shown in panel (a) of Figure 19-2. The vertical axis in this graph shows the price per gallon of gas, and the horizontal axis shows billions of gallons of gas. Recall from Chapter 2 that the supply curve shows the quantity that suppliers are willing to sell at any given price. In a competitive market, the supply curve is determined by the firm's marginal cost: the producer will sell any units for which the market price is at or above its marginal cost of producing that unit.

In Figure 19-2, the market is initially in equilibrium at point A: at the market price of $1.50 (P_A), producers will supply 100 billion gallons (Q_A) of gasoline.
Statutory Burdens Are Not Real Burdens: Panel (a) shows the equilibrium in the gas market before taxation (point A). A 50¢ tax levied on gas producers (the statutory burden) in panel (b) leads to a decrease in supply from $S_1$ to $S_2$ and to a 30¢ rise in the price of gas from $P_1$ to $P_2$ (point $D$). The real burden of the tax is borne primarily by consumers, who pay 30¢ of the tax through higher prices, leaving producers to bear only 20¢ of the tax.

Producers are willing to supply 100 billion gallons at $1.50 per gallon because $1.50 is the producers' marginal cost of producing that quantity of gas.

Panel (b) of Figure 19-2 shows the effects of imposing a tax of 50¢ per gallon of gas sold on the producers of gas. For these producers, this is equivalent to a 50¢ per gallon increase in marginal cost. Because firms must pay both their original marginal cost and the 50¢ tax, they now require a price that is 50¢ higher to produce each quantity. To supply the initial equilibrium quantity of 100 billion gallons after the tax is imposed, for example, firms would now require a price of $P_2 = 2.00$ (50¢ higher than the initial $1.50 equilibrium price, at point $B$). Because the tax acts like an increase in marginal cost, the entire supply curve shifts upward by 50¢ from $S_1$ to $S_2$ and the supply of gas falls.

At the initial equilibrium price of $1.50, there is now excess demand for gasoline. Consumers want the old amount of gasoline (100 billion gallons) at $1.50, but with the new tax in place producers are willing to supply only 80 billion gallons (point $C$). At $1.50, there is a shortage of $Q_1$ (point $A$) minus $Q_2$ (point $C$), or 20 billion gallons. Consumers therefore bid up the price as they compete for the smaller quantities of gas that are now available from producers. Prices continue to rise until the market arrives at a new equilibrium (point $D$).
with a market price of $1.80 ($P_3$) and a quantity of 90 billion gallons ($Q_3$). The market price is now 30¢ higher than it was before the tax was imposed.

**Burden of the Tax on Consumers and Producers** The tax has two effects on the participants in the gas market. First, it has changed the market price that consumers pay and producers receive for a gallon of gas; this price has risen by 30¢ from $1.50 to $1.80. Second, producers must now send a check to the government for 50¢ for each gallon sold.

From the producers’ perspective, the pain of the 50¢ tax is offset by the fact that the price the producers receive is 30¢ more than the initial equilibrium price. Thus, the producers have to pay only 20¢ of the tax, the portion that is not offset by the price increase.

From the consumers’ perspective, they feel some of the pain of the tax since they pay 30¢ more per gallon. Even though consumers send no check to the government and producers send a 50¢ check to the government, consumers actually bear more of the tax (30¢ to the producers’ 20¢). The price increase has transferred most of the tax burden from producers to consumers.

These burdens are illustrated in Figure 19-2 by the segments labeled “consumer burden” and “producer burden.” Using the formulas on p. 519, we can compute the burdens on consumers and producers. The consumers’ burden is

\[
\text{consumer tax burden} = (\text{posttax price} - \text{pretax price}) + \text{tax payments by consumers} = P_3 - P_1 + 0 = $1.80 - $1.50 = $0.30.
\]

The producers’ burden is

\[
\text{producer tax burden} = (\text{pretax price} - \text{posttax price}) + \text{tax payments by producers} = P_1 - P_3 + 0.50 = $1.50 - $1.80 + $0.50 = $0.20.
\]

The key insight is that the burden on producers is not the 50¢ tax payment they make on each gallon but some lower number, because some of the tax burden is borne by consumers in the form of a higher price. The sum of these burdens is $0.50, the total tax wedge created by this tax, which is the difference between what consumers pay ($1.80) and what producers receive ($1.30, at point $E$) from a transaction.

**The Side of the Market on Which the Tax Is Imposed Is Irrelevant to the Distribution of the Tax Burdens**

The second rule of tax incidence is that the side of the market on which the tax is imposed is irrelevant to the distribution of the tax burdens: tax incidence is identical whether the tax is levied on producers or consumers. In terms of the previous rule and Figure 19-2, this rule means that whether the 50¢ tax is imposed on producers or consumers, consumers will always end up bearing 30¢ of the tax and the producers will end up bearing 20¢.

\[\text{tax wedge} \quad \text{The difference between what consumers pay and what producers receive from a transaction.}\]
The Side of the Market Is Irrelevant. A 50¢ tax levied on gas consumers (the statutory burden) leads to a decrease in demand from $D_1$ to $D_2$ and to a 20¢ fall in the price of gas from $P_1$ to $P_3$ (with the market moving from the pretax equilibrium at point $A$ to the posttax equilibrium at point $D$). The real burden of the tax is borne primarily by consumers, who pay the 50¢ tax to the government but receive an offsetting price reduction of only 20¢; producers bear that 20¢ of the tax.

Figure 19-3 considers the impact of a 50¢ per gallon tax on consumers of gas. In this case, the tax is collected from consumers at the pump when they pay for their gas rather than from producers, as in Figure 19-2. Recall from Chapter 2 that the demand curve represents consumers' willingness to pay for any quantity of a good. Each point on the demand curve shows the quantity demanded for any market price encountered by consumers. With consumers having to pay a 50¢ tax in addition to the market price at every quantity, they are now willing to pay 50¢ less for each quantity. Thus, because the tax causes a reduction in consumers' willingness to pay, the entire demand curve shifts downward by 50¢, from $D_1$ to $D_2$. Before the tax, consumers were willing to pay a price of $P_1 = 1.50$ for the 100 billionth gallon of gas at point $A$. Now they are only willing to pay a price of $P_2 = 1.00$ for the 100 billionth gallon (point $B$), since they also have to pay the 50¢ tax on each gallon purchased.

At the old market price of $1.50$, there is now an excess supply of gasoline: producers are willing to sell the old amount of gasoline (100 billion gallons, at point $A$), but consumers are only willing to buy 80 billion gallons at that price, at point $C$. There is an excess supply of gasoline of $Q_1 - Q_2 = 20$ billion gallons at the initial equilibrium price of $1.50$ after the demand curve shifts. Producers therefore lower their price to sell their excess supply until the price falls to $1.30$ ($P_3$) at point $D$, with an equilibrium quantity of 90 billion
gallons ($Q_3$). The market price is now 20¢ lower than it was before the tax was imposed.

As in the previous example, this tax has two effects on the participants in the gas market. First, it has changed the market price that consumers pay and producers receive for a gallon of gas; this price has fallen by 20¢ from $1.50 to $1.30. Second, the consumer must now pay the government for 50¢ for each gallon purchased. At the equilibrium price of $1.30, adding the 50¢ tax yields a cost to consumers (price plus tax) of $1.80 at point $E$.

From the consumers' perspective, the pain of the 50¢ check is offset by the 20¢ per gallon decline in the market price. From the producers' perspective, they are feeling some of the pain of this tax since they are receiving 20¢ less per gallon. Even though producers send no check to the government, and consumers send a 50¢ check to the government, both parties bear some of the ultimate burden of the tax, since the price decrease has transferred some of the tax burden from consumers to producers.

These burdens are illustrated in Figure 19-3 by the segments labeled “consumer burden” and “producer burden.” Using our formulas, we can compute the burdens on consumers and producers:

\[
\text{consumer:} \quad p_3 - p_1 + 0.50 = 1.30 - 1.50 + 0.50 = 0.30
\]
\[
\text{producer:} \quad p_1 - p_3 + 0 = 1.50 - 1.30 = 0.20.
\]

Once again, the sum of the burdens on consumers and producers, the difference between what consumers pay ($1.80) and what producers receive ($1.30), is the tax wedge of 50¢.

Note that these tax burdens are identical to the burdens in the previous example. Consumers now have to pay the 50¢ at the pump, but they are facing a lower price ($1.30) to which they have to add that tax. Adding the two together, the consumer pays exactly the same amount ($1.80, price plus tax) as in the previous case. Producers now don’t have to pay a tax, but they receive a lower price for their gas ($1.30 instead of $1.50), so they end up receiving the same amount ($1.30) as well.

**Gross Versus After-Tax Prices** While there is only one market price when a tax is imposed, there are two different prices that economists often track in these types of tax incidence models. The first is the **gross price**, the price paid by or received by the party not paying the tax to the government; it is the same as the price in the market. The second is the **after-tax price**, the price paid by or received by the party that is paying the tax to the government; it is either lower by the amount of the tax (if producers pay the tax) or higher by the amount of the tax (if consumers pay the tax).

When the gas tax is levied on producers, as shown in Figure 19-2, the gross price paid by consumers is $1.80, and the after-tax price received by producers is $1.80 - $0.50 = $1.30. When the gas tax is levied on consumers, in Figure 19-3, the gross price received by producers is $1.30, and the after-tax price paid by consumers is $1.30 + $0.50 = $1.80. The after-tax price is equal to the gross price plus the tax wedge if the tax is on consumers, but is equal to the gross price minus the tax wedge if the tax is on producers.

gross price The price in the market.

after-tax price The gross price minus the amount of the tax (if producers pay the tax) or plus the amount of the tax (if consumers pay the tax).
full shifting. When one party in a transaction bears all of the tax burden.

**Parties With Inelastic Supply or Demand Bear Taxes; Parties With Elastic Supply or Demand Avoid Them**

In the previous example, we described a particular case in which consumers bear more of the burden of a tax than do producers. This is, however, only one of many possible outcomes. The incidence of taxation on producers and consumers is ultimately determined by the *elasticities of supply and demand* on how responsive the quantity supplied or demanded is to price changes.

**Perfectly Inelastic Demand** Consider again the case in which the 50¢ per gallon tax is levied on gasoline producers, but let’s assume this time that consumers have a perfectly inelastic demand for gas, as shown in Figure 19–4. At initial equilibrium, the price for 100 billion gallons is $P_1$ ($1.50). When the tax is levied on producers, they once again treat this as equivalent to a 50¢ rise in marginal cost, raising the price that they require to supply any quantity; supply falls and the supply curve shifts from $S_1$ to $S_2$. The new equilibrium market price is $P_2$ ($2.00), a full 50¢ higher than the original price $P_1$. When demand is perfectly inelastic, the tax burdens are

\[
\text{consumer burden} = (\text{posttax price} - \text{pretax price}) + \text{tax payments by consumers}
\]

\[
= (P_1 + 0.50) - P_1 = 1.50 + 0.50 - 1.50 = 0.50
\]

\[
\text{producer burden} = (\text{pretax price} - \text{posttax price}) + \text{tax payments by producers}
\]

\[
= P_1 - (P_1 + 0.50) = 1.50 - (1.50 + 0.50) = 0.50
\]

When demand is perfectly inelastic, producers bear *none* of the tax and consumers bear *all* of the tax. This is called the **full shifting** of the tax onto consumers.

**Figure 19–4**

- **Inelastic Factors Bear Taxes**
  A tax on producers of an inelastically demanded good is fully reflected in increased prices, so consumers bear the full tax.
**Perfectly Elastic Demand** Contrast that outcome with the case in which consumers’ demand for gas is perfectly elastic, as shown in Figure 19-5. Initially, the market is in equilibrium at \( P_1 = $1.50 \) and \( Q_1 = 100 \) billion gallons. In this case, when a 50¢ tax causes the supply curve to shift from \( S_1 \) to \( S_2 \), the equilibrium price remains at \( P_1, \ $1.50 \), but the quantity falls to \( Q_2, \ 80 \) billion gallons.

When demand is perfectly elastic, the tax burdens are therefore

- **consumer**: \( P_1 - P_1 = $1.50 - $1.50 = 0 \)
- **producer**: \( P_1 - P_1 + 0.50 = $1.50 - $1.50 + $0.50 = $0.50 \).

In this case, producers bear all of the tax and consumers bear none of the tax.

**General Case** These extreme cases illustrate a general point about tax incidence: **parties with inelastic supply or demand bear taxes; parties with elastic supply or demand avoid them.** Demand for goods is more elastic (the price elasticity of demand is higher in absolute value) for goods with many substitutes. For example, the demand for fast food is fairly elastic because higher-quality restaurant meals or home cooking can be substituted for fast food fairly easily. Thus, if the government levied a tax on fast food, fast-food restaurants would find it difficult to raise prices in order to pass all of the tax onto fast-food consumers; if they did, individuals would substitute one of these alternatives for their fast food. Thus, because the demand for fast food is elastic, the producers (the restaurants) bear most of the burden of the tax.

For products with an inelastic demand, the burden of the tax is borne almost entirely by the consumer. For example, the demand for insulin is highly inelastic because it is essential to the health of diabetics. If the government taxes the producers of insulin, they can easily raise their price and completely shift most of the tax burden onto consumers because there are no substitutes available that allow consumers to leave this market because of a higher price.

**Figure 19-5**

![Elastic Factors Avoid Taxes](image-url)

A tax on producers of a perfectly elastically demanded good cannot be passed along to consumers through an increase in prices, so that producers bear the full burden of the tax.
Supply Elasticities  Supply elasticity also affects how the tax burden is distributed. Supply curves are more elastic when suppliers have more alternative uses to which their resources can be put. In the short run, a steel manufacturer has fairly inelastic supply; having invested in the steel plant and expensive machinery to produce steel, there are few alternative choices for production. The plant cannot easily convert from making steel to making plastic pipes or wood furniture. So the supply curve for steel will be fairly inelastic (vertical). The supply of sales from sidewalk vendors (of items such as watches, purses, scarves, and so on) in New York City, in contrast, is very elastic. Since the individuals selling these goods have a very low investment in that particular business, if it is taxed they can easily move to other activities, such as working in a store selling the same items. So the supply curve for sidewalk vendor sales will be very elastic (horizontal).

Compare the incidence of a tax on steel (levied on steel producers) to the incidence of a tax on sidewalk vendors (levied on the vendors) for any given demand curve (assuming that the demand curve is neither perfectly elastic nor inelastic). Panel (a) of Figure 19-6 shows the impact of a tax on steel producers. The steel market is initially in equilibrium at point A. The steel company can only slightly reduce the amount of steel it produces because it is committed to a level of production by its fixed capital investment. As a result, even when the steel company is paying 50¢ to the government for each unit of steel produced, it still wants to produce almost the same amount. Overall, the

![Figure 19-6](image)

- **FIGURE 19-6**

(a) Tax on steel producers (inelastic supply)

(b) Tax on sidewalk vendors (elastic supply)

Elasticity of Supply Also Matters • A tax on producers of an inelastically supplied good (panel (a)) leads to a very small rise in prices, so producers bear most of the burden of the tax. An equal-sized tax on producers of an elastically supplied good (panel (b)) leads to a large rise in prices, so producers bear little of the burden of the tax (and consumers bear most of the burden).
steel company's supply curve shifts inward from \( S_1 \) to \( S_2 \). Price rises only slightly from \( P_1 \) to \( P_2 \), and quantity of steel sold falls only from \( Q_1 \) to \( Q_2 \); the new equilibrium is at point \( B \). Since the price rise is very small, it does not much offset the tax that the steel company must pay. The steel company therefore bears most of the tax, and consumers of steel bear very little (since they don't pay a much higher price).

Panel (b) of Figure 19-6 shows the impact of an equal-sized tax on sidewalk vendors. These vendors are very sensitive to the costs of production in their production decisions, leading to the very elastic supply curve. They are initially willing to provide a quantity \( Q_1 \) of goods at a price of \( P_1 \). If the government makes them pay 50¢ per good they sell, then many vendors will move out of the sidewalk vending business into some more lucrative line of work. The supply curve therefore shifts from \( S_1 \) to \( S_2 \), with prices rising from \( P_1 \) to \( P_2 \), and the quantity of goods sold falling from \( Q_1 \) to \( Q_2 \) (at point \( B \)). The large increase in price in the sidewalk vendors market greatly offsets the taxes the vendor has to pay, so they bear little of the burden of the tax. Consumers of goods sold by sidewalk vendors will see much higher prices for these goods, however, so they will bear most of the tax.

In summary, the incidence of a tax is determined by both demand and supply elasticities. In the appendix to this chapter, we develop the mathematical tax incidence formulas that formalize this intuition.

Reminder: Tax Incidence Is About Prices, Not Quantities

When the demand for gas is perfectly elastic, as in Figure 19-5, we claimed that consumers bore none of the burden of taxation, and yet the quantity of gas consumed fell dramatically. Doesn't this decrease in consumption make consumers worse off? And if so, shouldn't that be taken into account when determining tax incidence?

The answer to both questions is "no" because, at both the old and new equilibria, consumers in this case are indifferent between buying the gas and spending their money elsewhere. Each point on a demand curve represents consumers' willingness to pay for a good. That willingness to pay reflects the value of the next best alternative use of their budget. If the demand curve for gas is perfectly elastic, consumers are truly indifferent, at the market price, between consuming gas and consuming some other good. So if they have to shift to buying more of another good and less gas, they are no worse off.

More generally, when we analyze tax incidence we ignore changes in quantities and only focus on the changes in prices paid by consumers and suppliers. This assumption makes tax incidence analysis simpler.\(^3\)

---

\(^3\) Technically, theoretical tax incidence analysis applies strictly only very small changes in taxes. For those very small changes, the sole consumers who no longer consume the good are those for whom the value of the good is the same as the value of the next best alternative purchase (consumer surplus is zero). Similarly, the only suppliers who no longer sell the good are those for whom the cost of producing that good is the same as the revenues gained from selling it (producer surplus is zero). So, as in the perfectly elastic demand case, there is no implication of changing quantity for the well-being of either consumers or suppliers; only the change in price matters. In practice, we use the same formulas for larger changes in taxes, continuing to ignore any effects of changes in quantities.
19.2 Tax Incidence Extensions

Section 19.1 presented the fundamental rules that will guide tax incidence analysis throughout the rest of this book. To recap:

- The statutory burden of a tax does not describe who really bears the tax.
- The side of the market on which the tax is imposed is irrelevant to the distribution of tax burdens.
- Parties with inelastic supply or demand bear taxes; parties with elastic supply or demand avoid them.

In this section, we apply these rules to cases different from those previously considered, including taxes on factors of production, taxes in markets with imperfect competition, and accounting for (tax-financed) expenditures in tax incidence analysis. As we will see throughout the remainder of this book, the three basic rules of tax incidence are largely all we need to know to understand more complicated cases and issues in taxation.

Tax Incidence in Factor Markets

Our discussion thus far has focused on taxes that are levied in the goods markets, such as the markets for gas or fast food. Many taxes, however, are levied in factor markets, such as the market for labor. The analysis of tax incidence in factor markets is identical to that in goods markets; the only difference is that consumers of the factors are the firms (they demand factors such as labor) and producers of factors are individuals (who provide factors such as labor).

Consider, for example, the market for labor shown in panel (a) of Figure 19-7. Hours of labor supplied in the market are shown on the x axis; the market wage is on the y axis. There is a downward-sloping demand for labor from firms \( (D_L) \) and an upward-sloping supply of labor from individuals \( (S_L) \). The market is initially in equilibrium, before taxes, with a wage \( W_1 \) of $5.15 per hour at point \( A \).

Suppose that the government levies a payroll tax of $1 per hour on all workers. This tax lowers the return to work by $1 at every amount of labor. As a result, individuals require a $1 rise in wages to supply any amount of labor and the supply shifts up from \( S_1 \) to \( S_2 \) in panel (a) of Figure 19-7. With demand remaining at its original level, this shift results in a higher market equilibrium wage of $5.65 \( (W_2) \) at point \( B \). The incidence of the tax is shared by workers (suppliers) and firms (demanders) according to the elasticities of demand and supply. If these elasticities are equal, the burden is shared equally: the wage will rise to $5.65 per hour, and workers will take home $4.65 per hour after paying their $1 per hour tax. The firms and the workers each bear 50¢ of the $1.00 tax, split as indicated on the vertical axis. The firms pay a 50¢ higher wage ($5.65) and the workers receive a 50¢ higher wage ($5.65), but because they must pay $1 an hour in tax, they receive 50¢ less in after-tax wage ($4.65). The gross wage in the market has risen to $5.65, but the after-tax wage of workers has fallen to $4.65.
According to the second rule of tax incidence, what matters for the burdens on workers and employers from this tax is the total tax wedge and the elasticities of supply and demand, not who sends the check to the government. Panel (b) of Figure 19-7 shows the effect on the labor market if the payroll tax in our example were instead paid only by firms and not by workers. In that case, the supply curve would remain at $S_1$, and, because the tax on consumers (the firms) acts like an increase in the price of labor, the demand for labor would fall and the demand curve would shift inward to $D_2$. Market wages would fall by $1.00$ from $5.65$ to $4.65$, the new equilibrium (point $C$), and the burdens of taxation would be unchanged. Firms bear the same 50¢ burden as before; rather than paying a 50¢ higher wage, however, they now pay a wage ($4.65) that is 50¢ lower than the initial equilibrium wage of $5.15$. In addition, firms now must send a $1$ check to the government, so in effect they are paying a wage of $5.65$.

Workers see the same 50¢ burden; rather than receiving a 50¢ higher wage and sending a $1$ check to the government, however, they now receive a 50¢ lower wage ($4.65$). The gross wage in the market has fallen to $4.65$, but the after-tax wage paid by firms is $5.65$.

**Figure 19-7**

Incidence Analysis is the Same in Factor Markets • These figures show the market for labor where firms are the consumers and workers are the producers of hours worked at a wage rate $W$. A $1.00$ tax per hour worked that is levied on workers (panel (a)) leads the supply curve to rise from $S_1$ to $S_2$ and the wage to rise from its initial equilibrium value of $5.15$ (point $A$) to a higher value of $5.65$ (point $B$). A $1.00$ per hour worked tax that is levied on firms (panel (b)) leads the demand curve to fall from $D_1$ to $D_2$ and the wage to fall from $5.15$ to $4.65$ at point $C$. Thus, regardless of who pays the tax, workers and firms each have a burden of 50¢ per hour.
Arnold Harberger, one of the pioneers of the general equilibrium tax incidence model we discussed in Chapter 19, once wrote of his experience in Indonesia, where cars are taxed more heavily than motorcycles. This tax difference provided a great incentive to make motorcycles more carlike. As Harberger reports, "Three-wheel cycles were converted, by artful additions, into virtual buses, or at least taxis. Sometimes a single bench was added, with the passenger looking backward. Other times the cycle was stretched at the back, with two benches going down each side, and maybe even with an extra little running board cutting laterally across the rear (where the rear bumper of the car would be). I must say I was truly astounded when I saw my first eight-passenger motorcycle."1

This example highlights a simple fact: markets do not take taxes lying down. If there is some action that market participants can undertake to minimize the burden of taxation, they will do so. As long as there are substitutes for the consumption of any taxed good, some consumers will shift to those substitutes to avoid the tax, and as long as there are alternatives to the production of taxed goods, some producers will shift into producing those alternatives.

In this chapter we learn how attempts to minimize tax burdens have efficiency costs for society. In the absence of market failures, social efficiency is maximized at the competitive equilibrium without government intervention. When the government taxes market participants, they change their behavior to avoid the tax and move the market away from the competitive equilibrium, thereby reducing social efficiency. Put simply, it is costly for society to transport people in dangerous eight-passenger motorcycles instead of cars.

The remainder of the chapter uses these general lessons to explore the determination of the optimal taxation of commodities (goods such as cereal and cars) and income. We show how the tools of economic theory can be

---

used to describe the ideal tax system across goods or people and discuss how economists use empirical evidence to advise policy makers on constructing more efficient tax systems in the real world.

20.1 Taxation and Economic Efficiency

As we move from discussing the effects of taxation on equity (how the economic pie is distributed among market participants, the topic of Chapter 19) to discussing the effects of taxation on efficiency (how taxes affect the size of the economic pie), we shift our focus from the effect of taxes on market prices to their effect on market quantities. The discussion of tax incidence was about who bore the burden of taxation through tax payments and price changes. This discussion of tax efficiency is about the amount of social efficiency sacrificed by society when trades are impeded by the presence of taxation; the social efficiency effects of taxation are determined by the effect of taxes on quantities.

Graphical Approach

For modeling the efficiency consequences of taxation, it is useful to start with a graphical approach. Figure 20–1 shows the impact of a 50¢ per gallon tax levied on producers of gasoline. Before the tax is imposed, the demand curve for gasoline is $D_1$, the supply curve is $S_1$, and the market is initially in equilibrium at price $P_1 = 1.50$. Re curve supply comp is social and c feit an libriu $t$ they at $1.50$.

The decrease in for line so and th to a bie because consur and origina tion of $t$ $1.50$ curve). is in pl.

This is the con in falls sin leading chases. b AC these it plus is $t$ impos a tax is trades r.

From come co on cr: $1.30 f quantric mined deadw
rium where they intersect at point $A$, with quantity $Q_1$ (100 billion gallons) and price $P_1$ ($1.50 per gallon$).

Recall from Chapter 2 that in a perfectly competitive market the demand curve measures the social marginal benefit of gasoline consumption, and the supply curve measures the social marginal cost of gasoline production. At the competitive market equilibrium (point $A$), all gallons of gasoline that have a social marginal benefit greater than their social marginal cost are produced and consumed. The 100 billionth gallon ($Q_1$) has both a social marginal benefit and social marginal cost of $1.50, so this is the competitive market equilibrium. All previous units sold have a social marginal cost below $1.50 (since they are farther down on the supply curve) and a social marginal benefit above $1.50 (since they are higher up on the demand curve).

The $50c$ tax acts like an increase in producers’ costs, causing them to decrease the quantity supplied at each price, which shifts the supply curve in from $S_1$ to $S_2$. The new equilibrium is at point $B$; the quantity of gasoline sold has fallen from $Q_1$ (100 billion gallons) to $Q_2$ (90 billion gallons) and the price has risen from $P_1$ ($1.50) to $P_2$ ($1.80). The reduction in sales to a level below the competitive equilibrium quantity $Q_1$ means that, because of the tax, trades that would be beneficial to both producers and consumers of gasoline are not being made. The units between 90 billion and 100 billion (between $Q_2$ and $Q_1$) are units for which the social marginal benefit of consumption exceeds the social marginal cost of production; they are valued at more than $1.50 by consumers and cost less than $1.50 to produce (the pretax demand curve is above the pretax supply curve). Yet these units are not being produced and consumed once the tax is in place.

This reduction in quantity creates a deadweight loss of the area $BAC$. Since the competitive equilibrium quantity maximizes social efficiency, the reduction in quantity below $Q_1$ causes social efficiency to fall. Consumer surplus falls since consumers valued the units between $Q_2$ and $Q_1$ above their price, leading to a reduction in consumer surplus of $BAD$ from these forgone purchases. Producer surplus falls since producers could make profit on the units between $Q_2$ and $Q_1$, leading to a reduction in producer surplus of $DAC$ from these forgone sales. The sum of this reduction in consumer and producer surplus is the deadweight loss. Deadweight loss therefore measures the inefficiency of taxation, the amount of consumer and producer surplus society loses by imposing a tax. Deadweight loss is determined by changes in quantities when a tax is imposed, since this change captures the number of socially efficient trades that are not being made.

From the second rule of tax incidence, it should be clear that the outcome of this efficiency analysis would be the same if the tax were imposed on consumers instead of producers. In that case, market prices would fall (to $1.30 per gallon) instead of rising (to $1.80 per gallon), but the change in quantity would be the same. Since the efficiency effects of a tax are determined by the change in quantity, this would not affect the computation of deadweight loss.
Elasticities Determine Tax Inefficiency

Just as the price elasticities of supply and demand determine the distribution of the tax burden among market participants, they also determine the inefficiency of taxation: as demand and supply elasticities rise, the deadweight loss of taxation grows. This lesson is illustrated in Figure 20-2 for a tax on producers in two different markets. In panel (a), demand is relatively inelastic. A tax on producers shifts the supply curve inward from $S_1$ to $S_2$. This leads to a large rise in market prices from $P_1$ to $P_2$ and a relatively small reduction in market quantity from $Q_1$ to $Q_2$. Since deadweight loss is determined by the reduction in socially efficient trades, the deadweight loss in this case (area $BAC$) is small. If the government were to tax insulin, for example, there would be very little effect on the quantity of insulin demanded, and therefore little deadweight loss.

In panel (b), demand is more elastic. Thus, when the tax on suppliers shifts the supply curve from $S_1$ to $S_2$, there is a small rise in market prices from $P_1$ to $P_2$, but a relatively large reduction in market quantity from $Q_1$ to $Q_2$. As a result, the deadweight loss triangle ($BAC$) is much larger because many socially efficient trades (where the pretax demand is above the pretax supply) are not being made. Suppose, for example, that the government levied a tax on a particular fast-food restaurant, McGruber's. This tax would cause a large reduction in demand for McGruber's meals because individuals would shift their consumption to close substitutes (like Gruber King). This change is ineffo-

---

**Figure 20-2**

Deadweight Loss Rises with Elasticities • The deadweight loss of a given tax is smaller when the demand curve is less elastic (panel (a)) than when it is more elastic (panel (b)).
ficient, however, because the fact that individuals were eating at McGruber's before the tax indicates that McGruber's meals are their preferred choice. The large deadweight loss occurs because many individuals move away from their preferred choice in response to the tax.

As these two examples show, the inefficiency of any tax is determined by the extent to which consumers and producers change their behavior to avoid the tax; deadweight loss is caused by individuals and firms making inefficient consumption and production choices in order to avoid taxation. The competitive equilibrium quantity maximizes social surplus. Any change in quantity from the equilibrium point leads to inefficiency because trades that have a benefit larger than their cost are not made. Inefficiency is therefore proportional to the change in quantity induced by the tax. For insulin, there is little change in quantity induced by the tax, so there is little inefficiency. The tax on McGruber's fast-food restaurant induces a large change in quantity, so there is substantial inefficiency. The more elastic the demand or supply of a good is, the larger the change in quantity induced by the tax, and the larger the inefficiency of the tax.

**APPLICATION**

**Tax Avoidance in Practice**

The legendary economist John Maynard Keynes once remarked, “The avoidance of taxes is the only pursuit that still carries any reward.” His comment appears to have been taken to heart by many individuals whose elastic behavior allows them to avoid taxes. Some examples:

1. The British boat designer Uffa Fox lived in a home he constructed from a floating bridge. When the Inland Revenue (Britain’s tax collectors) attempted to collect property tax on the home, Fox began sailing it up and down the river. By the time he was done, Fox had collected so many different addresses that the Inland Revenue gave up their attempts.²

2. An Englishman visiting Cyprus in the early 1980s asked a tour guide why so many of the houses seemed to have steel reinforcement bars jutting out from their top floors. The guide informed him that Cyprus had a building tax that applied only to finished structures. Owners of those houses could thus claim that they were still in the process of finishing the roof. The process, of course, never ended.³

3. The Thai government levies a tax on signs in front of businesses. The tax is levied only on external signs and the rate depends on whether the sign is completely in Thai (low), in Thai and English (medium), or completely in English (very high). A walk around Bangkok thus reveals many businesses hanging English signs with a small amount of Thai writing in the

---

² Angus and Robertson (1978).
Determinants of Deadweight Loss

The appendix to this chapter mathematically derives the formula for deadweight loss as a function of the elasticities of supply and demand and the size of the tax. For convenience, economists often focus on the simplified case where supply is perfectly elastic but demand is not. In this case, the formula for $\text{DWL}$ is

$$\text{DWL} = -\frac{1}{2} \times \eta_d \times \tau^2 \times \left(\frac{Q}{P}\right)$$

where $\eta_d$ is the elasticity of demand and $\tau$ is the tax rate. From this equation, we learn two important lessons. First, deadweight loss rises with the elasticity of demand, $\eta_d$ (and with the elasticity of supply as well, in the general case); the more opportunities market participants have to consume or produce substitutes (the more elastic is demand or supply), the greater the inefficiency they will create by substituting.

As we discuss in the appendix, the appropriate elasticity to use for this calculation is one that reflects substitution effects only, not income effects (called the *compensated* elasticity). This is because any government revenue raising tax has income effects, since income is transferred from individuals to the government, so what determines the inefficiency of a particular tax is how much the tax distorts behavior due to substitution effects. In practice, however, it is typically difficult to distinguish the substitution and income effects of a price change, so we generally rely on the total (or *uncompensated*) elasticity when computing deadweight loss; we use the overall response of quantity to price, not the theoretically appropriate response that reflects substitution effects only.

Second, the deadweight loss rises with the *square* of the tax rate ($\tau^2$), so that the distortion from any given amount of tax is larger if the existing tax base is large. Thus, the distortion from a nickel tax on gas is much greater if it is the last nickel of a 25¢ tax increase than if it is the first nickel of a 5¢ tax increase. The *marginal deadweight loss*, the increase in deadweight loss per unit increase in the tax, rises with the tax rate.

This point is illustrated graphically in Figure 20-3. The gas market is initially in equilibrium at point $A$, with quantity $Q_1$ and price $P_1$. The government then imposes a 10¢ tax on producers, causing the supply curve to shift in from $S_1$ to $S_2$ as producers face a higher cost per unit produced (and so produce less at each price). Quantity falls to $Q_2$ at the new equilibrium point $B$. This tax creates a deadweight loss triangle with area $BAC$.

The government then levies a second tax of 10¢ on producers, which causes the supply curve to shift in even farther to $S_3$. Quantity now falls to $Q_3$ at the new equilibrium point $D$. The additional deadweight loss from this second tax is the trapezoid $DBCE$, which is much larger than the triangle $BAC$. The

---

Marginal deadweight loss from the second 10¢ tax (which brings the total tax to 20¢) is much larger than the marginal deadweight loss from the first 10¢ tax. After both taxes have been levied, the total DWL from the 20¢ tax is $DAE$.

The intuition behind this outcome relates to the Quick Hint about deadweight loss on page 50 of Chapter 2. Small deviations from the competitive market equilibrium are not very costly in terms of lost social surplus, because the transactions made close to the equilibrium are not the ones that generate a lot of social surplus. Indeed, a tax that reduced the quantity sold by only one unit would have approximately zero deadweight loss because the last trade was one for which consumers valued the good at roughly its price (no consumer surplus) and producer costs were roughly equal to price (no producer surplus). The 100 billionth gallon of gas has neither producer nor consumer surplus, so ending the sale of that particular gallon has little consequence for society.

As the market moves farther and farther from the competitive equilibrium, however, the trades that are impeded by taxation (trades for quantities between $Q_2$ and $Q_3$ for the first 10¢ tax, and trades for quantities between $Q_3$ and $Q_4$ for the second 10¢ tax) are trades that have more and more social surplus, as indicated by the widening gap between demand and supply. The loss of these higher-surplus trades means that deadweight loss is larger as the market moves farther from the competitive equilibrium.

**Deadweight Loss and the Design of Efficient Tax Systems**

The insight that the deadweight loss of a tax rises with the square of the tax rate, so that the marginal deadweight loss of taxation is higher at higher tax rates, has several important implications for the design of efficient tax policy.
A Tax System's Efficiency Is Affected by a Market's Preexisting Distortions

The fact that the deadweight loss rises with the square of the tax rate means that preexisting distortions in a market, such as externalities or imperfect competition, are key determinants of the efficiency of a new tax. Consider the two goods markets depicted in Figure 20-4. In the first market (panel (a)), there are no externalities, and the initial equilibrium is at point A, where quantity is Q1. In the second market (panel (b)), there are positive production externalities (like the donut shop and the policemen in Chapter 5). The positive externalities cause the social marginal cost of production (the SMC curve) to be below the private marginal cost (the S1 curve), since the firm does not incorporate the positive benefits for others into its supply decision. The firm chooses to produce at point E, where supply equals demand, but social surplus is maximized at point D, where SMC equals demand. So the firm underproduces at quantity Q2, where the social efficiency maximum is at quantity Q1. Any reduction in production below Q1 is inefficient because units for which social marginal benefit (measured by the demand curve) exceeds social marginal cost (the SMC curve) are not being produced. Thus, there is underproduction of the good, and a deadweight loss of area EDF.

Because to not being

Now st sold on su from S1 to to Q2 in p no externa in panel (b) positive pr area of tra BAC in par come to P production in quantity ring are on social costs.

This poi imperfectly firms alrea the efficient imposing th rium. Of co ition of th rather than in Chapter.

Progressive described in income ass that there c average tax tax systems taxation. th narrow inc indiduals fron up for the tax subset of in excluding so

This point two indi has an hour wages leads supply with labor demand 109
because trades with a positive social surplus (those between $Q_1$ and $Q_2$) are not being made.

Now suppose the government imposes an equal-sized tax of $1 per unit sold on suppliers in both markets. This tax shifts the supply curves inward from $S_1$ to $S_2$ in both panels, and causes the producer to cut back production to $Q_2$ in panel (a) and $Q_3$ in panel (b). For the market in panel (a), which has no externalities, the tax causes a small deadweight loss of $BAC$. For the market in panel (b), the market that already has a pretax deadweight loss because of its positive production externalities, the tax adds a a large deadweight loss (the area of trapezoid $GEFH$). This trapezoid is much bigger than the triangle $BAC$ in panel (a), because the tax in the second market moves the market outcome to $Q_3$, even farther away from the social efficiency maximizing level of production $Q_1$ than is $Q_2$. Once a market is already underproducing, the drop in quantity from a tax is especially costly because the trades that are not occurring are ones for which marginal social benefits significantly exceed marginal social costs.

This point also has important implications for taxation in markets that are imperfectly competitive, such as monopolies. Because imperfectly competitive firms already underproduce their goods relative to competitive equilibrium, the efficiency cost of imposing a tax on them is greater than the cost of imposing the same size tax on a market that is initially in competitive equilibrium. Of course, if there are negative externalities in a market, then the conclusion of the analysis is the opposite: a tax might have no deadweight loss, rather than a small deadweight loss, because it is correcting an externality (as in Chapter 5).

**Progressive Tax Systems Can Be Less Efficient** The insights about $DWL$ described here apply not only to taxation of goods but to taxation of income as well. Another implication of this rule about deadweight loss is that there can be large efficiency costs in moving from proportional (equal average tax rates on all) to progressive (higher average tax rates on the rich) tax systems. Moving to a progressive system means narrowing the base for taxation; the part of the tax that applies to the rich is levied on only the narrow income base of the rich. In general, it is more efficient to tax all individuals at an equal rate (a proportional tax), than to exclude some individuals from taxation and tax other individuals at a higher tax rate to make up for the lost revenues (a progressive tax). The efficiency lost by taxing a subset of individuals more highly is larger than the efficiency gained by excluding some individuals from taxation.

This point is best illustrated through an example. Consider a society of two individuals, one of whom has an hourly wage of $10 while the other has an hourly wage of $20. For both individuals, assume that a 10% rise in wages leads them to supply 10% more labor supply: their elasticity of labor supply with respect to after-tax wages is 1. Assume also that the elasticity of labor demand with respect to wages is $-1$: for each 10% rise in wages, firms demand 10% fewer hours of labor.
The initial equilibrium for each worker is shown in Figure 20-5. Without any taxes, the low-wage worker works 1,000 hours \( (H_1) \) and earns $10 per hour \( (W_1) \), as shown in panel (a), while the high-wage worker also works 1,000 hours \( (H_2) \), but earns $20 per hour \( (W_2) \), as shown in panel (b). The table below panels (a) and (b) shows taxes, hours, and deadweight loss in the initial equilibrium and when the government imposes taxes, as described in the following examples.

Suppose that the government imposes a payroll tax of 20% on all of a worker's earnings (a proportional tax). This tax lowers the income that both

![Figure 20-5](image-url)

**Low Rates Imposed on a Broad Base Are Desirable**: Initially, the government imposes an equal tax on the low-wage worker and the high-wage worker, which results in deadweight losses of triangles \( BAC \) and \( EDF \) in panels (a) and (b). When the government replaces this system with one of no tax on the low-wage worker, there is then no DWL for this worker, but the DWL for the high-wage worker increases by the trapezoid \( GEFI \), resulting in an overall increase in deadweight loss.

<table>
<thead>
<tr>
<th>Low-Wage Worker (panel (a))</th>
<th>High-Wage Worker (panel (b))</th>
<th>Total deadweight loss</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tax rate</strong></td>
<td><strong>Hours of labor supply</strong></td>
<td><strong>Deadweight loss</strong></td>
</tr>
<tr>
<td><strong>below $10,000</strong></td>
<td><strong>hours worked, ( H_1 )</strong></td>
<td><strong>from taxation</strong></td>
</tr>
<tr>
<td>No tax</td>
<td>1,000 (( H_1 ))</td>
<td>0</td>
</tr>
<tr>
<td>Proportional tax</td>
<td>894 (( H_2 ))</td>
<td>$115.71 (area ( BAC ))</td>
</tr>
<tr>
<td>Progressive tax</td>
<td>837 (( H_3 ))</td>
<td>$566.75 (area ( GDF ))</td>
</tr>
</tbody>
</table>

\[ \text{Total deadweight loss} = 0 + (347.13 + 566.75) = 913.88 \]
workers receive from work, making them less willing to supply labor, as reflected in the shift in the labor supply curve from $S_1$ to $S_2$ in both panels. For the low-wage worker, labor supply falls from 1,000 hours ($H_1$) to 894 hours ($H_2$), at a higher pretax wage of $11.18. As a result, the tax has created a deadweight loss of area $BAC$, an area of $115.71$. For the high-wage worker, the quantity of labor supplied also falls from 1,000 hours to 894 hours at a higher pretax wage of $22.36, creating a deadweight loss of area $EDF$, an area of $231.42$.

Quick Hint: Why is the deadweight loss larger for the higher-wage worker, despite the same reduction in hours worked? Because in a competitive labor market, the wage rate equals the marginal product of labor, so the high-wage worker has a higher marginal product of labor. As a result, society loses more efficiency when the high-wage worker reduces her hours (at a marginal product of $20$ per hour) than when the low-wage worker reduces her hours (at a marginal product of $10$ per hour).

Suppose now that society decides to switch to a progressive tax schedule: no tax on the first $10,000$ of earnings, but a $60\%$ tax rate on any additional earnings. This progressive tax system will raise almost exactly the same amount of revenues as the proportional $20\%$ tax on all earnings. Yet these two different tax systems have very different efficiency consequences.

Figure 20-5 illustrates this point. In panel (a), the change to the progressive tax schedule raises the benefits of working for the low-wage worker (since she is no longer taxed) and causes her labor supply curve to shift back to $S_1$. She is now back to her original optimum at point $A$: 1,000 hours of work at $10$ per hour. The deadweight loss associated with the low-wage earner falls from $115.71$ to zero. In panel (b), the change in the tax schedule further reduces the benefits of working to the high-wage worker and shifts her supply curve in to $S_3$, so that her hours fall from 894 to 837 ($H_3$). This change increases the size of the deadweight loss triangle to $GDI$, by adding the area of the trapezoid $GEFI$. The $DWL$ in panel (b) now has a total area of $566.75$.

This shift in the tax schedule has reduced $DWL$ by $115.71$ for the low-wage worker and raised it by $335.33$ for the high-wage worker. On net, $DWL$ has increased from $347.13$ to $566.75$, an increase of $63\%$ ($219.62$).

The large increase in $DWL$ arises because this more progressive tax is levied on a smaller tax base. The proportional tax is levied on a total of $30,000$ of earnings ($10,000$ of earnings for the low-wage worker and $20,000$ of earnings for the high-wage worker). The progressive tax is levied on a total of only $10,000$ of earnings (the second $10,000$ of earnings of the high-wage worker). In order to raise the same amount of revenues on this smaller tax base, the progressive tax must impose a higher tax rate, and a higher tax rate means a higher marginal deadweight loss. The low-wage taxpayer sees a reduction in her marginal tax rate from $20\%$ to zero, while the high-wage taxpayer sees an increase in her marginal tax rate from $20\%$ to $60\%$. Since the
deadweight loss rises with the square of the tax rate, the efficiency gains from reducing the low-wage tax rate by 20% are smaller than the efficiency costs of raising the high-wage tax rate by 40%.

This example illustrates a larger point: the more taxes are loaded on one source, the faster deadweight loss rises. By that logic, the most efficient tax systems are those that spread the burden of taxation the most broadly, so that the tax rate, the driver of deadweight loss, can be minimized. The guiding principle for efficient taxation is to create a broad and level playing field rather than taxing some groups or goods particularly highly and others not at all.5

Governments Should “Smooth” Tax Rates Over Time The fact that the deadweight loss rises with the square of the tax rate implies that governments should not raise and lower taxes as they need money but should instead set a long-run tax rate that will meet their budget needs on average, using deficits and surpluses to smooth out its short-run budget fluctuations. For example, suppose that a nation has a tax rate of 20% that finances its revenue needs. Now also suppose that the nation decides to enter a one-year war, which it estimates will double its revenue needs for one year, after which these will return to normal. The government should not finance its needs by raising the tax rate to 40% next year, and then lowering the rate back to 20% in the year after. Rather, the government should raise its tax rate by a small amount in all future years, for example by 1% for 20 years, to finance this war.

This course of action is suggested by the fact that the marginal deadweight loss rises with the tax rate. A tax of 40% in one period and 20% in the next causes more deadweight loss than a tax of 21% for 20 years, because the marginal deadweight loss associated with the increase in rates from 20% to 40% in one year is larger than the savings in deadweight loss going from a rate of 21% to a rate of 20% for 20 years. Just as individual utility is maximized by full consumption smoothing, government efficiency in taxation over time is maximized by tax smoothing, by having a relatively constant tax rate over time rather than high taxes in some periods and low taxes in others.

The Deadweight Loss of Taxing Wireless Communications

Hausman (2000) estimated the deadweight loss from a particularly dynamic sector of our economy: wireless communications services, those communications carried out with cell phones, PCs, and other wireless devices. In 1999, the state and federal tax burden on wireless communication in the typical state was 14

New Y
taxatio
raised j

This
tions is
-0.51.
because
wireles
many 1
service
ation. F:
square
loss is 7
This
margin
This d
grows:
mated
wireles
states,
almost

Opti

Secti
live
of how
the tax
ferent t
goods.

Ramse
The th
twentei
a gover
ferent t
Ramsey
How ca
tion? In
commen
ing its t
The
solution
was 14.5%, although the rate was 25% in high-tax states such as California, New York, and Florida. Hausman estimated that the deadweight loss from this taxation averaged 53¢ per dollar of revenue: for every dollar the government raised in taxes, social welfare was reduced by 53¢.

This figure is high for three reasons. First, demand for wireless communications is fairly price elastic; Hausman estimated a price elasticity of demand of -0.51. Second, there is already a large preexisting distortion in this market because wireless prices are well above marginal cost; the marginal cost of wireless services is only 5¢ per minute, while the typical wireless plan charges many multiples of that. Thus, there is already underproduction of wireless services relative to the competitive equilibrium, which is exacerbated by taxation. Finally, the taxes are fairly high, and the deadweight loss rises with the square of the tax rate; in California, New York, and Florida, the deadweight loss is 70¢ per dollar raised.

This is only the average deadweight loss. Hausman also computed the marginal deadweight loss from additional increments to wireless taxation. This deadweight loss is higher than the average, since deadweight loss grows as we move farther from the undistorted equilibrium. Hausman estimated that the marginal deadweight loss caused by an additional tax on wireless services ranged from 72¢ to 90¢ per dollar raised. Thus, in high-tax states, for every additional dollar in revenue raised, society would lose almost another dollar in efficiency losses.

20.2 Optimal Commodity Taxation

Section 20.1 has provided us with the necessary tools to turn from the positive question of how to measure deadweight loss to the normative question of how the existence of DWL should be taken into account in the design of the tax system. We address this normative question with reference to two different types of taxes. This section looks at commodity taxation, the taxation of goods. The next section discusses the taxation of income.

Ramsey Taxation: The Theory of Optimal Commodity Taxation

The theory of optimal commodity taxation began with the early-twentieth-century economist Frank Ramsey, who considered the problem of a government with a given budgetary requirement and the ability to set different tax rates for different commodities (food, clothing, tobacco, and so on). Ramsey formulated the problem of optimal taxation by asking the question: How can we raise a given amount of revenue with the least amount of distortion? In other words, how should a government set its tax rates across a set of commodities to minimize the deadweight loss of the tax system while meeting its budgetary requirement?

The appendix to this chapter presents the mathematics of Ramsey’s elegant solution. Here, we discuss the key lesson of his model: The government should
set taxes across commodities so that the ratio of marginal deadweight loss to marginal revenue raised is equal across commodities:

**Ramsey Rule:** set commodity taxes such that

\[
\frac{MDWL_i}{MR_i} = \lambda
\]

where \(MDWL\) is the marginal deadweight loss from increasing the tax on good \(i\), \(MR\) is the marginal revenue raised from that tax increase, and \(\lambda\) is the value of additional government revenues. This constant measures the value of having another dollar in the government’s hands relative to its next best use in the private sector. If \(\lambda\) is large, it implies that additional government revenues are quite valuable relative to keeping the money in private hands; if \(\lambda\) is small, then additional government revenues have little value relative to the value private individuals place on having that money.

This rule states that the deadweight loss per dollar of tax revenue associated with an additional dollar of taxes on commodity \(i\) should be equal for all commodities. If the tax on good \(A\) has an \(MDWL/MR\) that is higher than the \(MDWL/MR\) from taxing good \(B\), taxing good \(A\) causes more inefficiency per dollar of revenue raised than does taxing good \(B\). Recall that \(MDWL\) is a positive function of the tax rate; as discussed earlier, higher taxes lead to a higher marginal deadweight loss because they move the market farther from the competitive equilibrium (the deadweight loss is determined by the square of the tax rate). Therefore, to minimize inefficiency in the market, the government should reduce taxation of good \(A\), thus reducing its \(MDWL\), and raise the tax on good \(B\), increasing its \(MDWL\). These adjustments should continue until the \(MDWL/MR\) ratios for both goods are equal to \(\lambda\), so that both goods have the same efficiency cost per dollar of revenue raised.

If \(\lambda\) is large, then additional resources to the government have a high value, so the \(MDWL/MR\) should be large for all commodity taxes (tax rates should be high); if \(\lambda\) is small, then additional resources to the government have a low value, so the \(MDWL/MR\) should be small for all commodity taxes (tax rates should be low). In other words, the government should be willing to have potentially inefficient (high \(MDWL\)) taxes when it has large budgetary needs. This tells the government to set the marginal cost of taxation (\(MDWL/MR\)) equal to its marginal benefit (\(\lambda\)).

**Inverse Elasticity Rule**

It is convenient to express the Ramsey result in a simplified form that allows us to relate it to elasticities of demand. As we show in the appendix, if we assume that the supply side of commodity markets is perfectly competitive (elasticity of supply is infinite), then the Ramsey result implies that

\[
\tau_i^* = -\frac{1}{\eta_i} \times \lambda
\]

where \(\tau_i^*\) is the optimal tax rate for commodity \(i\), and \(\eta_i\) is the elasticity of demand for commodity \(i\). This equation indicates that the government should set taxes so that the tax rate on each commodity is proportional to \(1\) over the elasticity of demand; elastic goods (a higher value of \(\eta_i\)) should be taxed less and inelastic goods taxed more.
This formulation of Ramsey's rule shows that two factors must be balanced when setting optimal commodity taxes:

- **The elasticity rule**: When elasticity of demand for a good is high, it should be taxed at a low rate; when elasticity is low, the tax rate should be high. The deadweight loss from any tax rises with the elasticity of demand, so efficiency is enhanced by taxing inelastic goods more than elastic goods.

- **The broad base rule**: It is better to tax a wide variety of goods at a moderate rate than to tax very few goods at a high rate. Because the deadweight loss from a tax rises with the square of the tax rate, the government should spread taxes across a large number of commodities and not tax any one commodity at a very high rate.

To balance these two recommendations, the government should tax inelastically demanded goods at a higher rate, but should not look to collect all its taxes from these goods unless the price elasticity of demand is perfectly inelastic. If a government cared only about the elasticity rule, it would find the most inelastic good and raise all revenues from taxing that good. The broad base rule, however, tempers that tendency. Thus, while the government should tax inelastic goods more highly, it should tax other goods as well.

**Equity Implications of the Ramsey Model**

This inverse elasticity formulation of the Ramsey model highlights the fairly nasty equity implications of the Ramsey approach. Imagine that the government had only two goods it could tax, cereal and caviar. The elasticity of demand for caviar is much higher than that for cereal, so the inverse elasticity rule would suggest that the government tax cereal much more highly than caviar. This would mean imposing a tax on a good consumed exclusively by higher-income groups that was much lower than the tax imposed on a good consumed by all. This outcome, while efficient, might violate a government's sense of tax fairness across income groups (vertical equity).

An optimal commodity tax framework can address equity concerns by taking into account not only the elasticity of each commodity but also the income distribution of its consumers. Goods that are disproportionately consumed by higher-income consumers could have a tax rate above that implied by the inverse elasticity rule, and goods that are disproportionately consumed by lower-income consumers could have a tax rate below that implied by the inverse elasticity rule. How much of this "reweighting" of optimal taxes across commodities should be done is a function of the extent to which governments want to trade off efficiency for equity. As the government moves away from the Ramsey efficiency rule by bringing in equity issues, the tax system becomes less efficient but more equitable.

Perhaps because of these distributional concerns, there is relatively little reliance on commodity taxation in the United States. Most of our tax revenues come from taxing individual incomes, which we discuss after the following application.