7 Multipart Tariffs

7.1 Introduction and Definitions

A tariff is an algorithm for determining the bill to the customer for consumption of a firm's products. In the trivial case of one good with one price, the tariff is simply that price: the customer's bill is the price times the quantity consumed. This is called a one-part tariff.[1] For many regulated firms, tariffs are more complex, consisting of several billing components. For example, many local phone companies charge a monthly fee for service plus a charge for each call. The tariff in this case consists of the monthly fee and the per-call charge. Similarly, electricity providers often charge one price for consumption up to a certain number of kilowatt-hours during a month and then charge a different price for consumption beyond this amount. The tariff consists of the two prices and the consumption level, called the threshold, at which the price changes. Tariffs with several billing components are called multipart.

Multipart tariffs have important welfare implications. Perhaps the most relevant is the fact that a regulator, by applying an appropriately designed multipart tariff, can induce a natural monopolist to operate closer to the first-best outcome than would be possible with only one price. In fact, in some cases the first-best outcome can be attained exactly. These findings are demonstrated in the sections to follow. But first, some definitions, distinguishing various types of multipart tariffs, are required.

A *usage/access tariff* consists of a fixed fee that does not depend on level of consumption (called the access charge) and a per-unit price

1. The term "one-part tariff" is seldom used, because it is simply a price. The terminology serves to distinguish multipart tariffs.

for consumption (called the usage charge). Local phone service is usually subject to a usage/access tariff: the monthly fixed fee is the access charge and the price per call is the usage charge. The fixed monthly fee is called an "access fee" because by paying this fee, the consumer has access to the phone network. In a sense, the access fee is charged for the right to make phone calls (that is, for potential calls), and the usage fee is for actual calls.

With local phone service, a distinction is often made between *flat* and *measured* rates. Under flat-rate service, the consumer pays a fixed monthly fee and is not charged for local calls, while under measured service the consumer pays a monthly fee and is also charged for each local call. It is important to note that both flat and measured rates are types of access/usage tariffs. They differ only in whether the usage charge is zero or strictly positive.

Block rates are tariffs under which the price for each additional unit of consumption changes when the total level of consumption reaches certain thresholds. Electricity is usually charged in this way. One such tariff is represented in figure 7.1 and consists of: 7 cents per kilowatt-hour for up to (and including) 1,000 kWhs in a month, 5 cents for each kWh between 1,000 and 2,500 in the month, and 4 cents for each kWh over 2,500. Under this tariff, a consumer who uses 1,500 kWhs of electricity in a month is charged $95 ($70 for the first 1,000 kWhs, which are priced at 7 cents apiece, plus $25 for the 500 kWhs over 1,000, which are priced at 5 cents each).[2] The term "block rates" arises from the fact that the pricing algorithm, when graphed as in figure 7.1, looks like a series of blocks. The consumption interval under which one price applies (that is, from one threshold to the next) is called a *block*. The tariff in figure 7.1 therefore consists of three blocks: zero to 1,000, 1,000 to 2,500, and 2,500 and over.

Block rates can be *declining* or *inverted*, depending on whether the price for additional units decreases or increases as total consumption

2. Note that each price applies only for the portion of consumption within the range for that price. That is, when consumption is 1,500, the 5 cent price applies only to the kWhs over 1,000, not to the entire 1,500. It is possible to define a tariff that consists of charging one price for all units of consumption, with that price depending on the total level of consumption. If the same prices and thresholds are used as in figure 7.1, then, under this alternative definition, the consumer of 1,500 kWhs would be billed $75 (5 cents for each of the 1,500 kWhs). However, this type of tariff has the peculiar feature that, at the thresholds, the consumer's bill can decrease when consumption increases. In our example, consumption of 1,000 kWhs costs $75 (1,000 times 7 cents each), while consumption of 1,001 kWhs costs $50.05 (1,001 times 5 cents each).

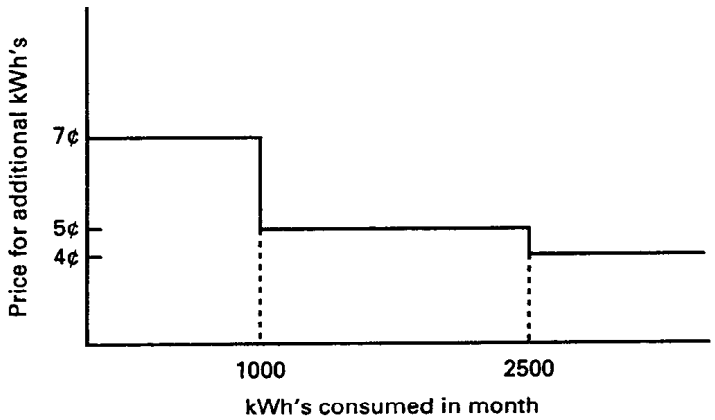

Figure 7.1
A block-rate tariff: declining blocks

increases. Figure 7.1 depicts declining blocks. An example of an inverted blocks tariff is given in figure 7.2, consisting of a charge of 3 cents per kWh for up to 800 kWhs, 5 cents for each kWh between 800 and 3,000, and 10 cents for kWhs over 3,000.

When discussing an individual consumer under a block-rates tariff, a distinction is made between *marginal* and *inframarginal* prices. The marginal price is the price the consumer pays for additional consumption given the total consumption of the consumer, and inframarginal prices are the prices that apply to lower levels of consumption. For example, a customer that consumes 1,500 kWhs under the tariff in figure 7.1 faces a marginal price of 5 cents and an inframarginal price of 7 cents. If, instead, this customer had used 2,600 kWhs, then the marginal price would be 4 cents and there would be two inframarginal prices: 7 and 5 cents.

Because the consumer is paying different prices for different units, a distinction between average and marginal prices is also necessary. Average price is the total dollar outlay of the consumer divided by the total consumption. (In the case of the consumer of 1,500 kWhs, this average price is 6⅓ cents: $95 divided by 1,500.) If total consumption does not exceed the first block, average and marginal price are the same. However, for total consumption beyond the first block, marginal price is below average price under declining block rates (because price declines as consumption increases) and above average price under inverted block rates (because price increases with consumption).

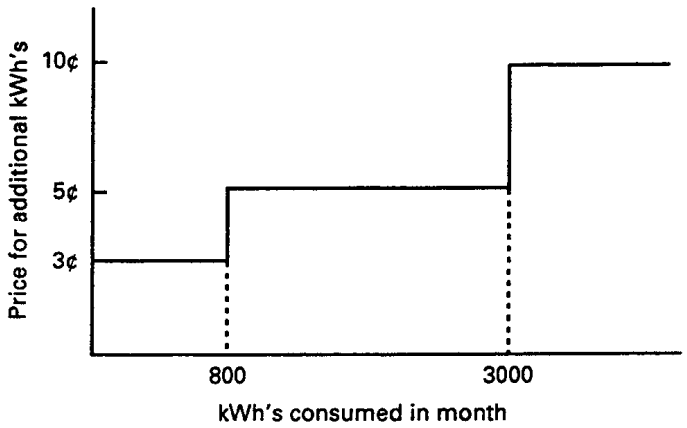

Figure 7.2
An inverted block-rate tariff

Block-rate tariffs and access/usage tariffs are not mutually exclusive concepts. In particular, the usage charge under access/usage tariffs can consist of block rates. In this case, a fixed monthly fee is charged and usage is priced in blocks. Local phone service is sometimes billed in this way. For example, it is increasingly common for local phone companies to charge a fixed monthly fee, allow a certain number of local calls per month (called an allowance) at no extra charge, and then levy a fee for each call above this allowance. In this case, usage is charged in inverted blocks, with the price in the first block being zero.

In much of the theoretical literature on multipart tariffs, an access/usage tariff is considered to be a type of block-rate tariff. Because the consumer cannot consume any of the product unless the access fee is paid, the price of the first unit of consumption can be viewed as the access fee plus the usage charge for one unit. For example, a tariff that consists of a $10 access fee and 5 cent usage charge can be viewed as a block-rate tariff with a charge of $10.05 for the first unit of consumption and 5 cents for each unit over one. This tariff is depicted in figure 7.3.

This equivalence between an access/usage fee and a block-rate tariff with two blocks is meaningful as long as the consumer would never choose to have access without consuming any units. Under an access/usage tariff the consumer is charged the access fee even if consumption is zero, while under a block-rate tariff the bill for zero consumption is necessarily zero. This distinction is important, for example, in

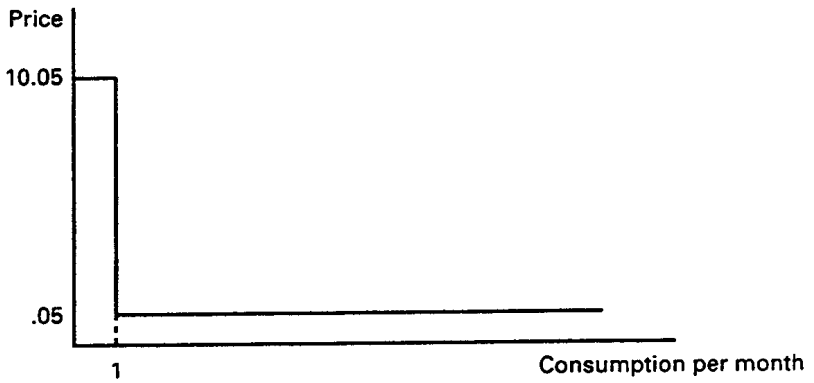

Figure 7.3
Access/usage tariff represented as a block-rate tariff

the case of phone service. A person might choose to have phone service (purchase access) and yet not make any calls. Any of several reasons could explain this choice: (1) having phone service allows the customer to receive incoming calls, even though no outgoing calls are made; (2) the customer might not know, when deciding to obtain service, whether it will want to make any calls during the upcoming period; (3) the customer might know that it will make no calls in the upcoming month (for example, the household might be going on vacation), but the cost of discontinuing service and then reconnecting after service is needed again is sufficiently high (in terms of time and any connect/disconnect charges that the phone company might levy) to prevent the customer from taking such action.

Generalizing from these reasons allows us to determine the conditions under which an access/usage tariff is equivalent to a block-rate tariff. The first reason given above for obtaining phone service without placing any calls is a simple case of externalities: the phone service provides an unpriced benefit, namely incoming calls. The second reason consists of the customer not knowing its demand; if the customer knew that its demand would be zero, it would (aside from the other reasons) choose not to have service and save the access fee. The third reason reflects transaction costs; if connecting and disconnecting were costless and effortless, then the customer that knew its demand and did not care about the externalities would choose service only for those periods when it knew it would use the service. The theoretical literature on multipart tariffs usually assumes (consistent with standard economic theory) that there are no externalities or

transaction costs and that the customer knows its demand. Under these conditions, an access/usage tariff is behaviorally the same as a block-rate tariff. The fact that a consumer with access but no usage would be billed the access fee under the access/usage tariff and not billed at all under the block-rate tariff is irrelevant under these conditions, because no consumer would do such a thing. However, in situations in which any of these conditions does not hold, an access/usage tariff differs from a block-rate tariff: a consumer could reasonably choose access without having any usage in a period.

The following sections describe the welfare gains that can be obtained with multipart tariffs, the optimal prices under these tariffs, and regulatory mechanisms that induce the firm to charge these prices. Some concepts and results are more evident or relevant for access/usage tariffs, and others for block-rate tariffs. To exploit this fact, the two types of tariffs are discussed separately. Similarities, which are numerous and illuminating, are identified along the way.

The findings of the chapter can be summarized as follows:

1. When access demand is fixed, the optimal access/usage tariff consists of a usage charge equal to the marginal cost of usage and an access fee that is sufficiently high to allow the firm to break even. First-best optimality is achieved. This is called the "Coase result."

2. When access demand is fixed, the surplus subsidy schemes of chapter 6 can be used to attain first-best optimality through an access/usage tariff without direct subsidy. The firm is allowed to choose the usage price. The regulator sets an access fee based on the surplus the firm generates at that price. Essentially, the access fee becomes the method by which the regulator subsidizes the firm. Over time, the firm moves to a usage price equal to marginal cost, and the access fee is set to allow the firm to break even.

3. When access demand is price sensitive, the optimal access/usage fees are determined by the Ramsey rule. Usage is priced above its marginal cost (except in very rare circumstances), and the access fee is lower than it would be if access demand were fixed. Second-best optimality is attained.

4. With price-sensitive demand for access, the V-F mechanism described in chapter 5 can be used to attain the optimal access/usage tariff.

5. Under traditional assumptions about consumer behavior, the con-

sumer acquires access only if the surplus from usage exceeds the access fee.

6. For any N block tariff with prices in excess of marginal cost, an $N + 1$ block tariff can be designed that Pareto dominates the N block tariff. (An N block tariff is a tariff that consists of N number of blocks, such as a two-block tariff or a three-block tariff.)

7. The optimal (second-best) $N + 1$ block tariff attains greater surplus than the optimal (second-best) N block tariff if prices under the N block tariff exceed marginal cost. This result implies that surplus can be increased by increasing the number of blocks in the tariff until first-best optimality is attained.

8. Except in very rare circumstances, the optimal (second-best) block-rate tariff consists of declining blocks.

7.2 Access/Usage Tariffs

7.2.1 Fixed Access Demand: The Coase Result

Consider first a situation in which demand for access is fixed independent of the access fee (at least within the range of possible fees under consideration). For example, all households in a region might choose to have phone lines at any of the prices that the local phone company and its regulator are considering. In this case, the optimal access and usage charges are easy to determine. Recall that the first-best outcome results from pricing at marginal cost and subsidizing the firm for whatever losses it incurs at these prices when producing efficiently. Coase pointed out in 1946 that an access/usage tariff can be designed that "mimics" this first-best pricing and subsidy scheme. In particular: charge a usage fee that equals the marginal cost for usage, and set the access fee at whatever level is needed for the firm to break even when it minimizes costs. With usage priced at marginal cost, the first-best consumption levels are attained. And the access fee provides sufficient revenue to the firm such that it breaks even when minimizing costs. Because the firm would lose money if it does not minimize costs, it chooses to produce efficiently.

To be concrete, consider, for example, a phone company that faces fixed costs of $1 million (to construct a phone network) and a constant marginal cost of 5 cents (to provide a call on the network). If the number of customers that obtain phone service is fixed at 100,000, the

Coase result states that the optimal access fee is $10 (the $1 million cost of the network divided by 100,000 customers) and the optimal usage charge is 5 cents. The access fee covers the cost of the network and the usage charge covers the cost of the calls on the network.[3]

This access/usage tariff is equivalent to charging marginal cost for usage and subsidizing the firm for its fixed costs. The access fee is the mechanism for providing this subsidy. With access demand fixed, the access fee is simply a transfer of funds with no consumption implications. The fact that the access fee does not affect consumption and therefore constitutes a pure transfer is the basic reason that first-best optimality can be attained.

7.2.2 Price-Sensitive Access Demand

Usually, access demand is not fixed, but rather varies, at least somewhat, on the basis of price. This fact is the basis for much of the pricing structure of telecommunication services, certainly prior to deregulation. Since the original establishment of the Bell system, one of the primary goals of local regulators has been to promote "universal service," which is expressed colloquially as "a phone in every home." This goal has been pursued through pricing strategies for local and long-distance service that maintain relatively low monthly charges for local access, so that as many households as possible are able and willing to purchase access.[4] Implicit in this approach is the idea that access demand is price sensitive: low monthly charges for access increases the number of households with phone service; conversely, raising the monthly charge reduces the demand for access.[5]

---

3. In the more general situation, marginal cost is not constant but varies with the level of production. In these cases, the optimal usage fee is the fee at which the marginal-cost curve intersects the demand curve for usage, and the optimal access fee is the firm's losses at this price divided by the number of customers.

4. Because of deregulation, the term "local access" has also been used to describe long-distance carriers' access to the local network. In this chapter, however, the term refers consistently to access from the customer's perspective, such that local access means acquiring of a phone line.

5. Given the importance of the concept of universal service in the regulation of local phone service, it is not surprising that a considerable body of empirical work has been directed toward determining the price elasticity of phone lines (that is, access). See for example the survey by Taylor (1980). Although the estimated elasticities vary over time and geographic regions (as one would expect), the general finding seems to be that price elasticity for access, while not zero, is fairly low. This result implies that, while

When access demand is not fixed, the access fee cannot be treated simply as a subsidy mechanism with no consumption implications. Raising the access charge involves a surplus loss because some consumers will be induced to forego phone service in the face of the higher price. This fact needs to be incorporated into the determination of optimal prices.

Identifying the optimal access and usage charges when access demand is price sensitive is actually quite straightforward. Instead of thinking of the firm as providing one good (e.g., phone service), we can think of the firm as providing two goods: access and usage. For example, in the case of local phone service, the phone company provides phone lines and calls on these phone lines. A person can choose whether or not to purchase a phone line and, given the line, whether or not to place calls. The two goods have separate, though interrelated, demands; and there is a marginal cost associated with each good separately. For example, the demand for phone lines is different from the demand for calls, though the two are related (because calls can be made only if a line is purchased and a line will be purchased more readily if the consumer anticipates making calls). From the company's perspective, there is a marginal cost for adding a phone line to the local phone network, and there is a marginal cost of providing a call on the network.

First-best optimality is attained by setting the access fee equal to the marginal cost of access and the usage fee equal to the marginal cost of usage. Generally, however, this would result in losses for the natural monopolist. The question becomes: what are the optimal access and usage fees when the firm is required to break even?

Chapter 4 identified the second-best prices for a two-good natural monopolist when the firm cannot be subsidized: Ramsey prices. The optimal access and usage fees that allow the firm to remain solvent are therefore those consistent with the Ramsey rule. Letting subscripts a denote access and u denote usage, the Ramsey rule in this situation becomes (see section 4.5)

$$\left(\frac{P_a - MC_a}{P_a}\right)\cdot(\epsilon_a - \epsilon_{ua}) = \left(\frac{P_u - MC_u}{P_u}\right)\cdot(\epsilon_u - \epsilon_{au}), \tag{7.1}$$

where

in actuality access demand is price sensitive, treating access demand as fixed and using the Coase result for pricing might be acceptably accurate.

ϵ_a is the elasticity of access with respect to the access fee;
ϵ_{ua} is the elasticity of usage with respect to access fee;
ϵ_u is the elasticity of usage with respect to usage fee; and
ϵ_{ua} is the elasticity of access with respect to usage fee.[6]

The Ramsey rule in this situation states that the percent by which the access fee is raised above the marginal cost of access, multiplied by the "net" elasticity for the access fee, is the same as for the usage fee. This rule can be used to calculate the second-best access and usage fee, with the methods described in section 4.4.

In special situations, the Ramsey rule reduces to a simpler expression. For example, when access demand is fixed, the Ramsey rule reduces to the Coase result. That is, the Ramsey rule under this condition states that the optimal usage fee is the marginal cost of usage and the optimal access fee is whatever fee is necessary to allow the firm to break even. To see this, rewrite equation (7.1) with all the elasticities relating to access being zero, to represent the fact that access demand is fixed independent of price:[7]

$$\begin{aligned}((P_a - MC_a)/P_a)\cdot(0) &= ((P_u - MC_u)/P_u)\cdot(\epsilon_u), \text{ or,} \\ 0 &= (P_u - MC_u)/P_u)\cdot(\epsilon_u).\end{aligned} \tag{7.2}$$

The only way this equation can hold is for the percent increase in usage fee over its marginal cost to be zero; that is, for P_u to be set equal to MC_u. With the usage fee at marginal cost, the firm breaks even only if the access fee is raised sufficiently. Note that equation (7.2) holds no matter how high the access fee is raised, because the percent increase in the access fee is multiplied by zero. In short, the Ramsey rule implies the Coase result when access demand is fixed.

---

6. The cross-elasticities can be zero, but in general are not. For example, because calls can be made only if a phone line is acquired, the monthly charge (that is, access fee) indirectly affects the number of calls placed through its impact on the number of lines acquired. Similarly, with a high enough per-call charge, the customer might realize that it will make no calls and hence need not pay to acquire a phone line. Hence the usage fee can affect the demand for access.

7. The same-price elasticity ϵ_a is clearly zero. The cross-price elasticity ϵ_{ua} is also zero, because, if access demand is fixed, changes in the usage fee do not affect access demand. The fact that ϵ_{au} is also zero is less evident. Though it has not been stated so far, the elasticities in the Ramsey rule relate to compensated demand rather than uncompensated demand. A basic result from microeconomic theory is that cross-elasticities of the compensated demands for two goods are equal. Consequently, ϵ_{au} is necessarily zero if ϵ_{ua} is zero. This distinction between compensated and uncompensated demand has not been made in the text because it adds a layer of complication that we have found does not translate into sufficiently greater insight to warrant burdening readers.

When access demand is not fixed but rather is price sensitive, the Ramsey rule requires that (except in rare cases)[8] the usage fee be set above its marginal cost. With the usage fee above marginal cost, revenues from marginal usage exceed the cost of providing that usage; as a result, the additional revenues needed for the firm to break even are not entirely generated by the access fee. The access fee can therefore be lower than in the Coase situation. Stated succinctly: when access demand is price sensitive, the optimal access fee is lower and the optimal usage fee is higher than when access demand is fixed.

The first-best outcome is not attained when access demand is price sensitive and the firm is required to break even, unlike the situation with fixed access. The reason for this is clear. When access demand is price sensitive, the access fee cannot serve simply as a subsidy mechanism; it also affects access demand and, indirectly, usage. This fact gives additional insight to the Ramsey rule. Ramsey pricing results in the second-best outcome in general; this second-best outcome becomes the same as the first-best outcome when the demand for one good is fixed.

A result opposite to that of Coase is obtained when usage demand is fixed, rather than access demand. In this case, the Ramsey rule becomes

$$((P_a - MC_a)/P_a)\cdot\epsilon_a = 0. \tag{7.3}$$

This equation holds only if $P_a = MC_a$, that is, if the access fee is set equal to the marginal cost of access. The usage fee is then set sufficiently high to allow the firm to break even.

This result has some important implications. Regulated monopolists in many settings do not charge access fees; electricity and natural gas are examples. This practice of charging usage fees that allow the firm to break even but no access fee is optimal only if the demand for usage is fixed and the marginal cost of access is zero. However, usage demand can nearly always be expected to be more price sensitive than access demand, simply because usage is conditional on access. Surplus gains can therefore be expected from moving toward more reli-

8. It is interesting to note that the optimal usage price can actually be below marginal cost in particular circumstances. If $\epsilon_u - \epsilon_{au}$ has an opposite sign from $\epsilon_a - \epsilon_{ua}$, then equation (7.1) holds when $P_u - MC_u$ is negative. This event occurs if the cross-elasticity of demand between access and usage exceeds (in magnitude) the access elasticity but not the usage elasticity. This event is highly unlikely in reality.

ance on access fees, a practice that is starting, for commercial and industrial customers at least, in electricity sales.[9]

7.2.3 Regulatory Mechanisms for Optimal Access/Usage Tariffs

Chapter 5 identified a regulatory mechanism that induces the firm to charge Ramsey prices in equilibrium, namely, Vogelsang-Finsinger (V-F) regulation. Because the second-best access and usage fees are simply the Ramsey prices for these two goods, V-F regulation can be imposed to induce the firm to charge these access and usage fees in equilibrium. The regulation takes the form of allowing the firm in each period to charge whatever usage and access fees it chooses as long as these fees, when multiplied by the past period's access and usage quantities, equal the last period's costs. That is, in period t, the firm is allowed to charge any access and usage fees that satisfy

$$P_a^t Q_a^{t-1} + P_u^t Q_u^{t-1} \leq C^{t-1}, \tag{7.4}$$

where Q_a is the quantity of access (for example, number of phone lines), Q_u is the quantity of usage (for example, number of phone calls), and C is the cost of providing both access and usage.

As shown in chapter 5, the application of this type of regulation induces the firm to move over time to the second-best access and usage fees. While the firm might waste during its movement to equilibrium, once equilibrium has been reached the firm will produce efficiently at the second-best prices.

It is important to note that V-F regulation is applicable whether access demand is fixed or price sensitive. Recall that when access demand is fixed, the Ramsey rule implies the same prices as the Coase result, namely, usage priced at its marginal cost and access demand priced sufficiently high for the firm to break even. First-best optimality is attained, with the access fee providing the necessary subsidy for the firm. V-F regulation, which results in Ramsey prices in equilibrium, therefore achieves the Coase result and first-best optimality when access demand is fixed. When access demand is price sensitive, the Ramsey prices, and hence V-F regulation, attain second-best optimality.

---

9. Charging access fees for electricity might prevent some consumers from being able to purchase any electricity. While surplus will still increase, the hardship imposed on these consumers might be considered sufficiently inequitable to warrant foregoing these gains.

If access demand is fixed, other regulatory mechanisms can also be used by the regulator to induce the first-best outcome. In chapter 6, several surplus subsidy schemes are described that induce the first-best outcome in equilibrium. These procedures (at least those of Loeb/Magat and Sappington/Sibley) bring the firm to equilibrium more quickly than V-F regulation and with no waste on the way. With demand for access fixed, access/usage tariffs can provide a means for implementing these procedures.

Recall that under the surplus subsidy schemes, the regulator allows the firm to set its own price and then subsidizes the firm on the basis of the price that the firm charges. The need to subsidize the firm often constitutes a practical barrier to the implementation of these schemes. In particular, there is often no mechanism by which the regulator can subsidize a firm. Traditionally, regulators are not legally able to tax the population and use the generated revenues to subsidize public utilities. More fundamentally, if regulators were allowed to raise revenues for subsidies, they nevertheless would face the problem of how to generate revenues without distorting consumption patterns.

The dilemma can be solved by the use of access/usage tariffs. Recall that when access demand is fixed, the access fee serves simply as a transfer of funds from consumers to the firm, with no effect on consumption. The access fee can therefore provide the means by which the regulator subsidizes the firm for a surplus subsidy scheme.

In particular, the regulator can mandate that customers be billed under an access/usage tariff, with the firm free to choose the usage price in each period but not the access fee. Using the firm's chosen usage price, the regulator calculates the subsidy that the firm should be paid under whichever surplus subsidy scheme the regulator wishes to apply. The regulator can then set the access fee in each period to be this subsidy divided by the number of customers demanding access. The access fee generates the subsidy for the firm through the firm's bills to its customers, without the regulator needing to raise funds itself or pay a subsidy directly. That is, the regulator, within its traditionally defined role of approving tariffs, can institute a regulatory mechanism that is equivalent to subsidizing the firm.

When demand for access is fixed, this use of access fees to generate subsidies is appropriate, because the level of the access fee does not affect consumption patterns. If, however, access demand is not fixed, then the access fee does not constitute a pure transfer and the consumption implications of changing the access fee must be considered.

7.2.4 Behavior of Consumers under Classic Assumptions

In the previous sections, we state that access and usage can be considered two goods with separate but interrelated demands. Under certain conditions it is possible to describe more precisely the relation between the demand for access and the demand for usage. This analysis provides insight to the consumer's choice process even in situations in which the conditions do not hold.

Consider a situation in which the assumptions of traditional economic theory, mentioned in section 7.1, hold. In particular, assume that the customer knows its demand for usage and that there are no externalities or transaction costs involved in access. Our task is to determine, for any given usage and access fees, whether or not a customer will choose to acquire access and its level of usage conditional on this choice.

Figure 7.4 depicts a case in which the consumer's demand for calls is linear (to make the arithmetic simple). The y-intercept of the demand curve is 90 cents, meaning that, if calls were charged at 90 cents each, the consumer would choose not to make any calls. Suppose the phone company charges \$15 per month (the access fee) and 10 cents per call. The question is: will the consumer choose to have a phone line and, if so, how many calls will be placed? Consider the two options available to the customer:

1. *Acquire Service*
If the customer acquires a phone line, the customer will place 150 calls a month, as indicated by the demand curve at 10 cents per call. These calls provide surplus to the customer, because for all but the last call the customer is willing to pay more than 10 cents for the call. The total consumer surplus from the calls is the shaded area in the graph, which equals \$60 ($150 \cdot (.9-.1) \cdot (1/2)$). The phone line costs \$15 per month: the customer must pay this in order to make calls and receive the \$60 of surplus from calling. On net, the customer would obtain surplus of \$45 from acquiring a phone line (the \$60 surplus from the calls minus the \$15 cost of the phone line).

2. *Do Not Acquire Service*
If the customer does not acquire a line, the customer will incur no costs but will also obtain no surplus, because no calls can be placed.

Given the choice of \$45 extra surplus with a line and no gain with-

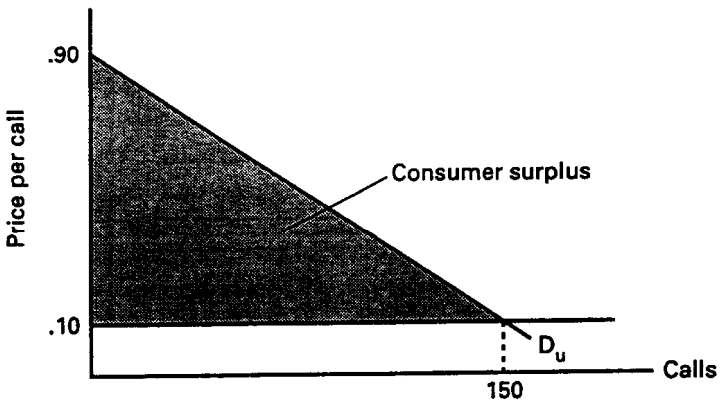

Figure 7.4
Interrelated demand for access and usage

out a line, the customer in this case will choose to acquire the line. Stated succinctly, the consumer acquires access if the surplus generated by usage, given the usage fee, exceeds the access fee, and foregoes access otherwise.

These concepts can be used to compare the behavior of consumers with different usage demands and to clarify the impacts of changes in access and usage fees. Consider, for example, a customer with lower demand, as depicted in figure 7.5. At 10 cents per call, this customer would make only forty calls in a month. The surplus from these calls would be only $12. Because the cost of acquiring a line ($15) is greater than the surplus the customer would obtain from making calls on the line, this customer will choose not to acquire a line.[10]

Consider now the impact of changes in usage and access fees on the demand for access and usage, using the customer depicted in figure 7.4 as illustration. If the access fee is raised, but not above the $60 of surplus the consumer obtains from calls, then the consumer continues to acquire service and place 150 calls. That is, usage is unaffected by a change in the access fee as long as the access fee is below the consumer's surplus from usage. The only effect of the access fee in this range is a transfer of funds from the customer to the

10. Recall that we have assumed that there are no externalities or transaction costs and that the customer knows its demand. If any of these assumptions does not hold, the consumer in this situation might nevertheless choose to acquire a line. For example, the consumer might expect to obtain more than $3 worth of benefit from incoming calls, such that the surplus from usage and the value of the externalities (incoming calls) exceeds the cost of acquiring the line.

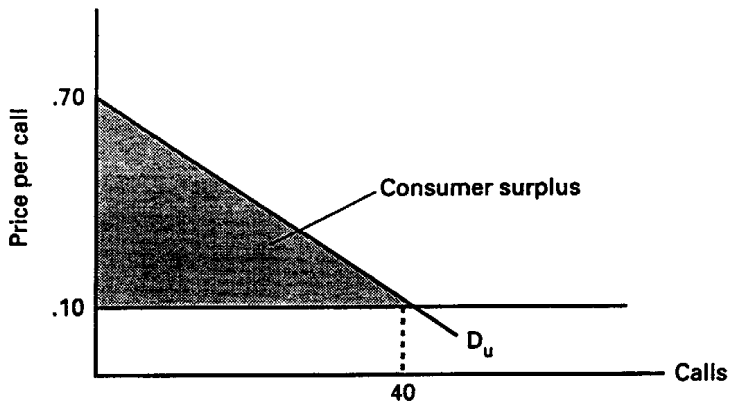

Figure 7.5
Consumer chooses not to acquire line

firm.[11] However, if the access fee is raised above $60, then the customer chooses to forego service and make no calls. Consumption drops to zero when the access fee exceeds the surplus obtained from usage.

The fact that the customer foregoes service when the access fee is too high creates a discontinuity in the relation between the access fee and the quantity of usage demand. Usage is positive and constant as long as the access fee is below the consumer's surplus from usage and is zero for any higher access fee. In our example, usage demand is 150 for access fees below $60 and zero for access fees above $60.

Consider now the usage fee. Raising the usage fee decreases, of course, the quantity of usage demanded. For example, as shown in figure 7.6, if the per-call charge is raised from 10 to 18 cents, then the customer will place only 135 calls instead of 150. The higher usage charge, however, also affects the surplus from usage and thereby can affect whether the customer acquires service. With a usage fee of 18 cents, the surplus from usage is $48.60 ($135 \cdot (.90 - .18)/2$). Because this surplus exceeds the access fee of $15, the customer will continue to acquire service when the usage charge is raised to 18 cents.

However, if the usage charge is raised high enough, the surplus from usage will shrink sufficiently that the customer will no longer choose to acquire service. At a per-call charge of 50 cents, the surplus that the customer obtains from usage is exactly $15 ($75 \cdot (.90 - .50)/2$).

11. Raising the access fee decreases the consumer's available income. If the consumer's usage demand depends on income, usage demand will be affected by the change in access fee. However, this is a second-order, and usually very small, effect.

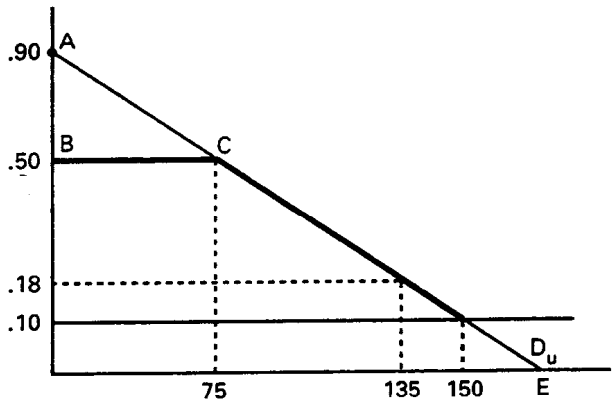

Figure 7.6
Impact of raising usage fee

Therefore, if the usage fee is raised above 50 cents, the customer will choose to forego service and make no calls. The demand for calls is therefore zero for any usage price above 50 cents.

Because of this phenomenon, two curves are relevant for the demand for usage. One curve is the demand for usage *conditional* on service being acquired. This curve is the line *ACE* in figure 7.6 and represents the quantity of usage that would be demanded at any usage fee provided the consumer chooses to acquire access. The other curve is the demand for usage given the consumer's choice process regarding access and the fee for access. This curve is the kinked line *ABCE:* zero for usage fees above 50 cents (because the consumer chooses not to acquire service when the usage fee is this high) and the same as the conditional demand curve for usage fees below 50 cents (because the consumer chooses to acquire service at these prices).

The situation of access demand being fixed can be viewed in the context of this discussion. Access demand is fixed only if surplus from usage is sufficiently large for each consumer that changes in access and usage fees (at least within the relevant range) do not induce any consumer to forego service. In this case, the access fee can be raised as much as necessary without the detrimental effects of consumers foregoing service. The optimal prices therefore consist of the usage fee at marginal cost and the access fee at the level needed for the firm to break even.

7.2.5 Empirical Example of Access and Usage Charges

Local phone companies have traditionally charged a fixed monthly fee for a phone line and allowed an unlimited number of local calls to be placed on the line at no extra charge. As stated in section 7.1, this type of access/usage tariff, under which the usage charge is zero, is called flat-rate service.

We know, however, that the optimal usage fee equals the marginal cost of usage if access demand is fixed, and (except in rare cases) exceeds marginal cost if access demand is price sensitive. Because the marginal cost of a local call is above zero, flat-rate service results in an inefficiently large amount of calling. With the price of a call below marginal cost, consumers make calls whose value is less than the marginal cost of the call. Furthermore, if access demand is price sensitive, flat-rate service results in fewer consumers obtaining phone lines than is optimal, because the price of access is higher than optimal.

Because of these and other considerations, local phone companies in many areas are introducing measured service, under which local calls are charged on a per-call and/or per-minute basis. The revenues generated by these usage charges allow the firm to provide service at a lower monthly access fee.

In considering flat-rate and measured service, the central issue is essentially empirical: does a shift from flat-rate to measured service actually increase consumer surplus, and, if so, by how much? Theory implies that surplus increases.[12] However, theory cannot determine the size of the gain. The gain might not be sufficiently large to warrant the disruption caused by the change in tariff.

Train (1989) investigated empirically the impact that a shift from flat-rate to measured service for local phone calls would have on consumer and producer surplus. Because both flat-rate and measured service are access/usage tariffs, the study is essentially an analysis of the surplus impact of moving from a usage fee of zero and a high access fee to the optimal usage and access charges. The investigation provides estimates of the potential surplus gains that can be obtained

12. If there are costs associated with measurement (for example, the cost to the phone company of keeping records on local calls to use in billing customers), surplus need not increase: the gains from pricing nearer marginal cost might be less than the cost of billing for usage.

from moving from zero to positive usage fees when the standard economic assumptions are satisfied.

Train's analysis is restricted in several ways. (l) The empirical demand model and calculation of surplus are specific to residential customers of the phone company in one particular area and are not necessarily generalizable to other areas. (2) The implementation of measured service requires that the phone company keep track of each customer's usage so that the customer can be billed appropriately. The company must incur costs for this accounting that it would not have to incur under flat-rate service. Train's analysis assumes that these costs are sufficiently small as not to affect the calculation of optimal prices under measured service. (3) Demand for access is considered fixed, such that the shift from flat-rate to measured service is assumed not to affect the number of consumers that acquire phone service. This assumption was necessitated by limitations of the empirical demand model, which predicts demand conditional on the consumer having a phone line and does not examine the choice of whether to acquire a line.

Because of these restrictions, the numerical results are not to be taken as prescriptions for particular rates. Rather, the analysis is useful in elucidating the types of trade-offs that need to be considered in a comparison of flat-rate and measured service in any setting, and in providing an indication of the orders of magnitude of the surplus losses and gains that are involved in these trade-offs.

For his analysis, Train borrowed a previously estimated model of the demand for calls (Train, McFadden, and Ben-Akiva 1987). He obtained cost information from the local phone company. With these data, he simulated the revenues, costs, and consumer surplus that would occur under various usage and access fees. Simulations were performed with a per call charge of zero cents (which represents flat-rate service), 1 cent, 2 cents, and so on up to 12 cents. At each usage charge, the access charge was derived that allows the phone company to break even.[13] Consumer surplus was then calculated.

The consumer surplus generated at each usage fee is given in the third column of table 7.1. A usage fee of zero constitutes flat-rate ser-

13. Because demand for phone lines was assumed to be fixed, the access charge was determined for each usage charge by: (1) determining the net operating profits (revenues from usage minus variable costs of providing usage), (2) subtracting these net operating profits from the portion of fixed costs that must be covered by residential local service, and (3) dividing this difference by the number of phone lines.

Table 7.1
Access/usage fees for local phone service

Usage price (cents per call)	Access fee that allows firm to break even (dollars per month)	Change in consumer surplus (dollars per household per month)	
		Under traditional assumption	With flat-rate bias
0	10.48	0[a]	0[a]
1	9.06	0.73	−1.17
2	7.97	1.25	−3.70
3	7.13	1.56	−5.84
4	6.46	1.77	−8.11
5	5.90	1.93	−10.43
6	5.13	2.02	−12.81
7	5.03	2.05	−15.23
8	4.67	2.07	−17.68
9	4.35	2.08	−20.16
10	4.06	2.07	−22.64
11	3.80	2.05	−25.15
12	3.55	2.02	−27.62

a. Flat-rate service is taken as base for comparison.

vice. The access fee associated with a zero usage fee is $10.48. This situation is taken as the base when comparing consumer surplus under measured rates (that is, under nonzero usage charges). The optimal usage fee (that is, the fee that provides the greatest consumer surplus) is about 9 cents per call. The access charge that allows the firm to cover its costs at this usage fee is $4.35. The cost information provided by the phone company indicates that a call of average dura tion costs about 9 cents to provide. The optimal prices therefore correspond to the Coase result, which states that when access demand is fixed, the optimal tariff consists of pricing usage at marginal cost (i.e., 9 cents) and setting the access fee sufficiently high to cover fixed costs.[14]

Consumer surplus was estimated to increase by an average of $2.08 per household at the 9 cent usage fee compared to flat-rate service. This is a fairly large increase, considering that households pay $10.48

14. The marginal cost of a call is probably lower than 9 cents in actuality. The analysis therefore probably overestimates the surplus benefits to be obtained from measured service.

per month under flat-rate service. It is interesting, however, that over half of these benefits are obtained at a usage price of 2 cents and over 90% are obtained at a price of 5 cents. These sharply declining returns suggest that a politically constrained regulator can implement "compromise prices" (that is, usage charges above zero but below marginal cost) and be assured that a large portion of the potential benefits are being obtained. More precisely, a compromise price that is a given percent of the marginal cost will obtain considerably more than that percent of the potential benefits.

The foregoing results were obtained under the standard economic assumptions regarding consumer behavior (namely, that demand is known and there are no externalities or transaction costs). There are reasons to believe, however, that consumers do not behave in accordance with these assumptions. In fact, the demand analysis of Train, McFadden, and Ben-Akiva (1987), which served as the basis for Train's work, suggests that consumers behave differently. In particular, these researchers found that consumers seem to value flat-rate service over measured service even when the bill that the consumer would receive under the two services, given the number of calls the consumer places, would be the same. This finding requires some elaboration. Suppose a consumer places a certain number of calls during a month. Under flat-rate service, the consumer is billed only the flat monthly fee for service. Suppose rates under measured service are set such that the consumer receives exactly the same bill as under flat-rate service. This measured service necessarily entails a lower fixed monthly fee to offset the usage charge, such that the total bill ends up the same. In this situation, the empirical analysis indicates that consumers strongly prefer the flat-rate service, even though the two cost the same. This phenomenon is called the "flat-rate bias," and has been found in many studies of local phone service.

The existence of this bias is problematical. Standard theory of consumer behavior does not incorporate it. According to this theory, if a consumer makes the same number of calls and pays the same bill under two different tariffs, the consumer's surplus is the same in both cases. (The benefits to the consumer are the same because the same number of calls are made, and the amount the consumer must pay to receive those benefits is the same. Therefore, the surplus, which is the difference between benefits and costs, is also the same.) However, empirically, consumers have been found to prefer flat-rate service over measured service in this situation, indicating that, contrary to theory, their welfare is higher under flat-rate service.

There are many possible explanations for this bias. The bias could reflect inertia. Flat-rate service is traditional, and consumers might prefer it simply because they are accustomed to it. Alternatively, the bias might reflect risk-avoidance. If the consumer does not know its demand exactly, then the consumer bears a risk under measured service that does not exist with flat-rate service. Under flat-rate service, the consumer's bill is fixed independent of how many calls are made. However, under measured service, the consumer does not know beforehand how large the phone bill will be. In fact, the bill will fluctuate from month to month depending on the number of calls made. A risk-avoiding consumer will therefore prefer flat-rate service over measured service even if on average the two cost the same.

The calculation of the impacts of a shift from flat-rate to measured service depends critically on how the bias toward flat-rate service is treated. If the bias is considered to result from inertia, such that it disappears in the long run when consumers eventually adjust to new circumstances, it can perhaps be ignored in the calculation of consumer surplus. This is the approach taken by Train for the results presented above. Alternatively, if the bias is considered to reflect risk avoidance, then it constitutes a legitimate aspect of consumer welfare and should be included in the calculations. A shift from flat-rate to measured service increases the risk that consumers bear, and this risk (in itself) reduces the welfare of consumers.

To investigate the impact of this bias on optimal prices, Train also simulated consumer surplus with the flat-rate bias included in the calculation of consumer surplus. The results are given in the fourth column of table 7.1. As the figures indicate, flat-rate service generates greater surplus than measured service under *any* strictly positive usage price. This implies that, if the bias that is evident in consumer's choices in the real world reflects risk-avoidance or some other factor that truly affects consumer welfare, a switch from flat-rate to measured service is not advisable. This conclusion runs counter to the standard concept that total surplus increases as prices move toward marginal costs. (It also suggests the importance of expanding pricing theory to explicitly account for the possibility of uncertainty in demand and risk avoidance by consumers.)

The overall conclusions of Train's study can be summarized. If the bias toward flat-rate service that is observed in consumers' choices is considered to constitute a real component of consumer surplus, a move from flat-rate to measured service would be inadvisable, because the

evidence indicates that it will decrease consumer surplus, perhaps considerably. However, if the bias is considered not to be a real part of consumer surplus, then measured service offers the potential for significant gains for consumers. The gains rise at a sharply decreasing rate as the usage fee is raised toward marginal cost. As a result, a regulator can obtain a large share of the potential gains without raising price all the way to marginal cost.

7.3 Block Rates

Before describing the welfare implications of block rates, we first examine the behavior of customers under these tariffs.

7.3.1 Declining Block Rates: Outlay Schedule and Budget Constraint

The choice process of the consumer under declining block rates is, in the most basic characteristics, the same as in the standard situation of one price for each good, namely: the consumer maximizes utility subject to a budget constraint. The only difference is that with declining block rates the budget constraint is shaped differently than with one price.

To identify the budget constraint under declining block rates, consider, as a specific example, the tariff for electricity that was used in section 7.1: 7 cents per kWh for up to 1,000 kWhs per month, 5 cents for each kWh between 1,000 and 2,500 per month, and 4 cents for each kWh over 2,500. Panel (a) of figure 7.7 depicts the price schedule for this tariff. The bill that a consumer would receive for any level of consumption is computed directly from this price schedule. The relation between consumption and the bill a consumer would receive is called the "outlay schedule," because it represents the necessary "outlay" by the consumer for each level of consumption. Panel (b) gives the outlay schedule of this tariff. Outlay starts at zero for no consumption of electricity and increases at a rate of 7 cents per kWh up to 1,000 kWh. At this point, outlay is $70. After 1,000, outlay increases at the rate of 5 cents per kWh up to an outlay of $145 for 2,500 kWhs (which consists of $70 for the first 1,000 kWhs and $75 for the next 1,500 kWhs). After 2,500 kWhs, outlay increases at 4 cents per kWh.

The consumer's budget constraint is derived from the outlay sched-

Tariff: 7¢ for 0–1000 kWh's
5¢ for 1000–2500 kWh's
4¢ for 2500+ kWh's

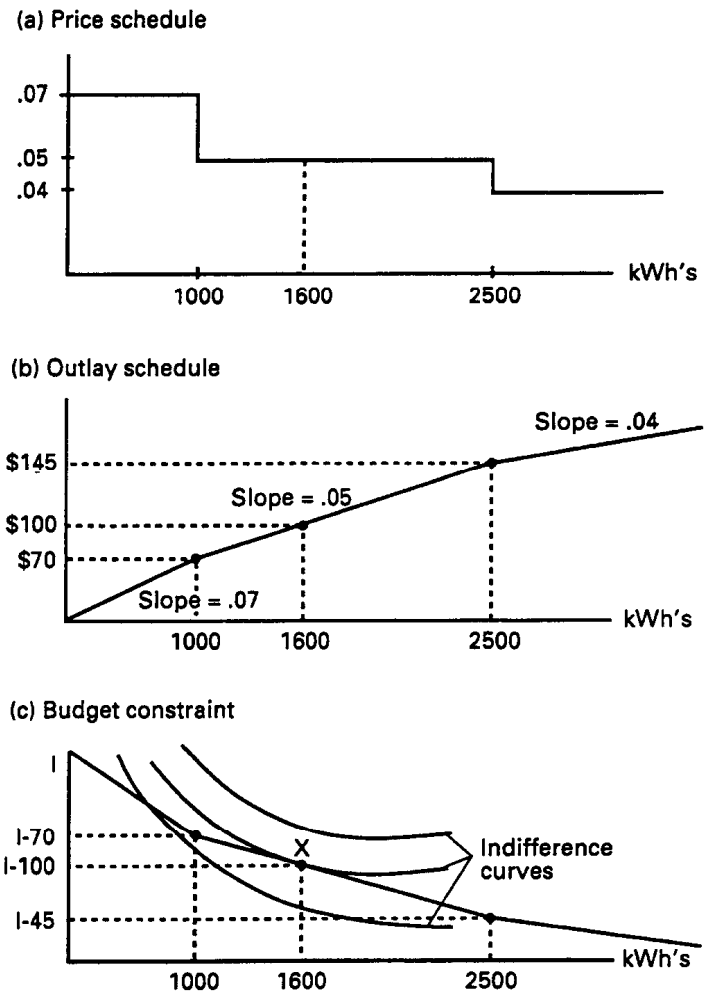

Figure 7.7
Schedule for a declining block tariff

ule. To allow the graph to be shown in two dimensions, suppose the consumer has a choice between two goods: electricity and a composite good called "other goods." The price of other goods is considered to be \$1, by normalization. The budget constraint gives the combinations of electricity and other goods that the consumer can afford, given its income. If the consumer buys no electricity, then it can spend its entire income, denoted I, on other goods. The y-intercept of the budget constraint is therefore I units, because the price of other goods is \$1. For consumption of electricity between zero and 1,000 kWhs, the consumer must forego 0.07 units of other goods to obtain 1 kWh of electricity. The slope of the budget constraint is therefore $-.07$ until electricity consumption reaches 1,000, at which point the consumer obtains only $I-70$ units of other goods. For 1,000 to 2,500 kWhs of consumption, the consumer must forego 0.05 units of other goods to obtain one extra kWh, giving a slope of $-.05$ for the budget constraint in this range. And so on.

The consumer chooses the point on the budget constraint that provides the greatest utility. The consumer's preferences for electricity and other goods are, as usual, represented by the consumer's indifference mapping. The consumer chooses the point on the budget constraint that attains the highest possible indifference curve. Given the mapping shown in panel (c), the consumer chooses point X, which corresponds to 1,600 kWhs of electricity and $I-100$ units of other goods. The consumer's electricity bill in this case is \$100.

The consumer's decision process can also be represented in terms of demand curves. The most straightforward situation is depicted in panel (a) of figure 7.8. The price schedule is obtained directly from the tariff. The consumer's demand curve is derived from the indifference mapping.[15] The consumer purchases additional electricity until

15. To be completely accurate, it is necessary to assume that demand is independent of income, at least within the range of income that is relevant. By definition, the demand curve gives that quantity of electricity that the consumer would choose, given the consumer's income, at any particular level of the price for the good. As such, the demand curve is derived under the concept that there is one price for the good. With block-rate tariffs, there is more than one price. The difference between marginal and inframarginal prices under a block-rate tariff constitutes an income transfer from the consumer to the firm. This fact is demonstrated in the text to follow. It implies that a change in the marginal price changes the consumer's income, which, if demand is income sensitive, changes the demand curve. To maintain one demand curve in the graph, it is therefore necessary to assume that changes in income have no effect on the demand for electricity.

To avoid confusion, note that the change in income that results from a change in the

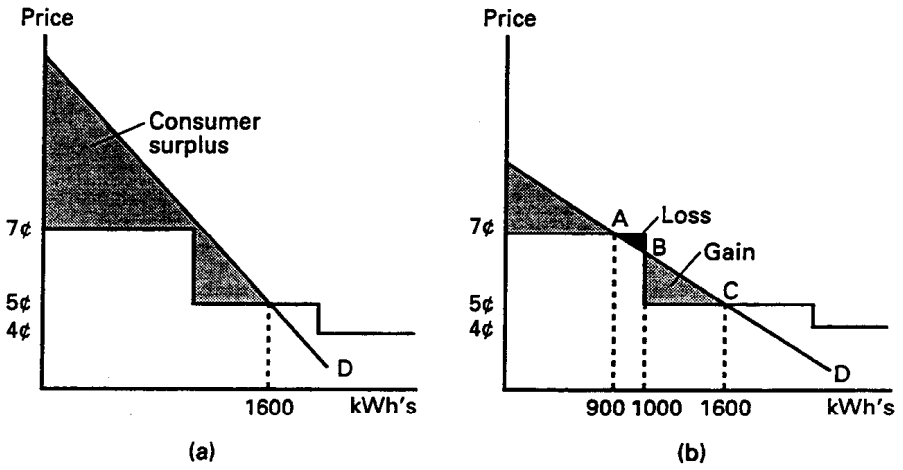

Figure 7.8
Demand curve and chosen consumption under declining block rates

the price for additional consumption, as given by the price schedule, equals the benefits of additional consumption, as given by the demand curve. This event occurs at 1,600 kWhs, consistent with the analysis based on indifference curve in figure 7.7. Consumer surplus is, by definition, the difference between the total benefits the consumer receives and the total amount the consumer pays. Total benefits is the area under the demand curve, and total outlay is the area under the price schedule. Consumer surplus is therefore the shaded area between the demand curve and the price schedule.

Because price declines as consumption rises, it is possible for the demand curve to intersect the price schedule at two or more points. Panel (b) of figure 7.8 is an example. Determining which point the consumer chooses in these situations requires an analysis of the surplus implications of the consumer's moving from one of these intersections to the next. Suppose the consumer is consuming at the first intersection, at 900 kWhs. The surplus at this level of consumption is the shaded area to the left of point *A*. The first question is: will the consumer's surplus increase or decrease by moving from *A* to the second intersection, *B*? The answer is clearly "decrease," because the

---

marginal price under a block-rate tariff occurs in addition to the standard income effect resulting from a price change in the standard one-price situation. The standard income effect represents a change in the buying power of a given nominal income. A change in the marginal price under a block-rate tariff results in a change in the nominal income itself.

value the consumer places on each kWh between *A* and *B*, as given by the demand curve, is less than the price that the consumer would have to pay for them, as given by the price schedule. Moving from *A* to *B* would decrease the consumer's surplus by the amount given in the deeply shaded area between these two points. The next question is: will the consumer's surplus increase or decrease by moving from *A* all the way to the third intersection, *C*? The consumer would lose the amount in the deeply shaded triangle by moving from *A* to *B* on the way to *C*. However, once the consumer reached the threshold for a lower price, additional consumption would be valued at more than this lower price and the consumer would obtain additional surplus. The lightly shaded area is the amount of additional surplus the consumer would obtain by moving to *C* once it had reached *B*. Moving from *A* to *C* would therefore entail a loss of the deeply shaded area and a gain of the lightly shaded area. Because, given the way the demand curve is drawn in this example, the gain exceeds the loss, the consumer would choose to consume at *C* instead of *A* (*B* is of course worse than either *A* or *C*).

Declining block tariffs have the peculiar characteristic that it is possible for the consumer to have two or more utility-maximizing consumption levels under these tariffs. Figure 7.9 illustrates this possibility. The consumer's indifference mapping could be such that the highest attainable indifference curve is tangent to the budget constraint at two points, as shown in panel (a). From a demand perspective, this phenomenon means that the demand curve intersects the price schedule at three points and that the the loss in moving from the first to the second intersection is the same in magnitude as the gain in moving from second to the third intersection. As a result, the first and third intersections provide exactly the same surplus.

It is interesting to note that a consumer with standard preferences would never choose to consume at a threshold of a declining block-rate tariff. Consider point *T* in panel (a) of figure 7.10. This point represents consumption at the threshold at which the price changes from 7 to 5 cents. For any smooth, convex (that is, inward-bending) indifference curves, point *T* could never be the point on the budget constraint that attains the highest indifference curve. Because indifference curves are smooth and the budget constraint is kinked at *T*, the indifference curve through *T* is necessarily below the budget constraint in some area around *T*. From *T*, the consumer could always move to a higher indifference curve by moving away from *T*. In the

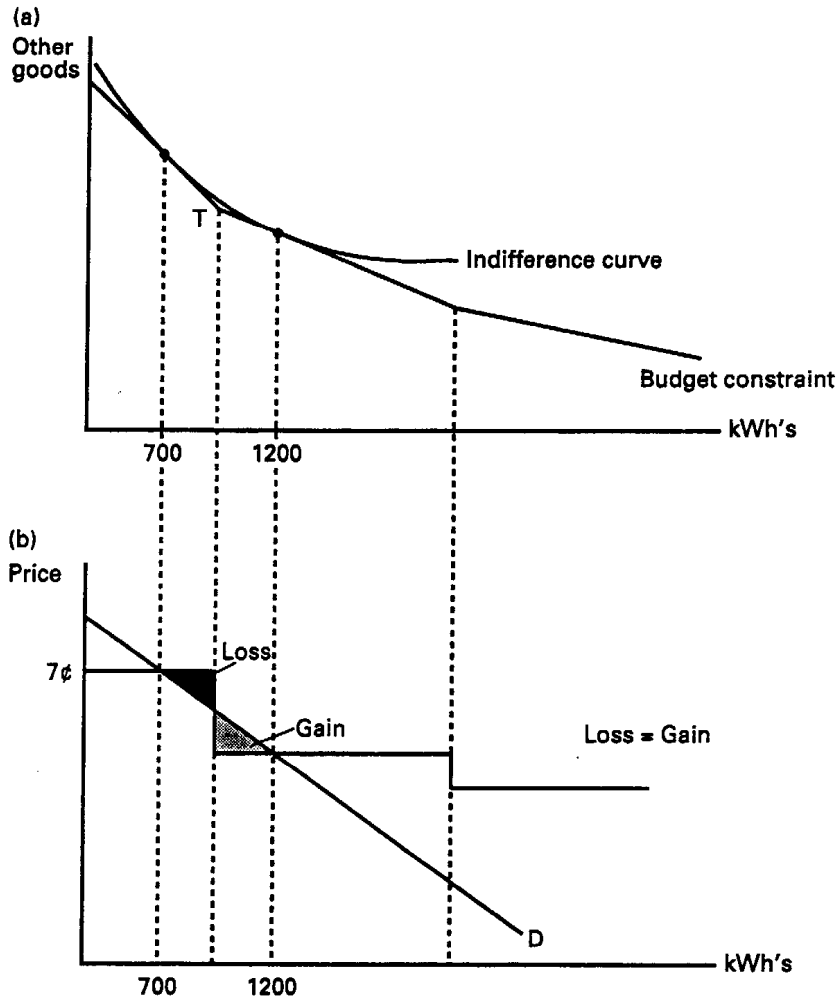

Figure 7.9
Two utility-maximizing consumption levels

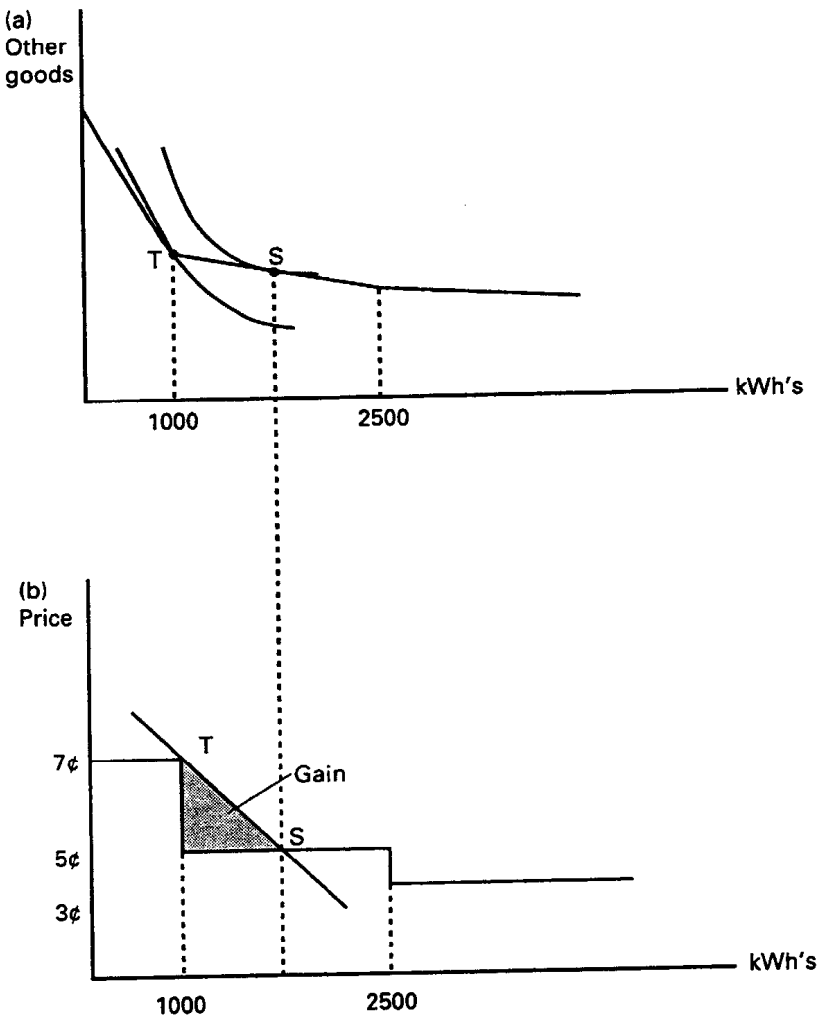

Figure 7.10
Consumer doesn't choose threshold

graph, for example, the consumer would move to S, at which an indifference curve is tangent to the budget constraint.

Panel (b) depicts this situation in terms of the demand curve. If the demand curve intersects the price schedule at a threshold, such as point T, then it necessarily intersects at another point, such as S. Moving from the threshold to another intersection always entails an increase in surplus.

For each consumer, a block-rate tariff can be viewed as a access/usage tariff. That is, an access/usage tariff can be designed that provides the customer the same bill and the same marginal price as the block-rate tariff. Consider, as in figure 7.11, a consumer who consumes in the second block of a declining block tariff. Under the block-rate tariff, this person pays 7 cents for each of the first 1,000 kWhs and 5 cents for each additional kWh consumed in the second block. The total bill for 1,600 kWhs of electricity is \$100 (.07 · 1,000 + .05 · 600). The marginal price the customer faces is 5 cents. An access/usage tariff can be designed under which the consumer would face the same bill and the same marginal price. Let the usage charge be the price in the second block (5 cents) to maintain the same marginal price for the consumer. Then set the access fee at the level necessary for the bill to be the same under the two tariffs. At 1,600 kWhs of consumption, the usage fee itself generates a bill of \$80 (.05 · 1,600). Therefore an access fee of \$20 provides the customer with a total bill of \$100, the same as under the block rates. To summarize: a block-rate tariff with 7 cents for consumption up to 1,000 kWhs and 5 cents for consumption between 1,000 and 2,500 kWhs is "equivalent," for anyone consuming between 1,000 and 2,500 kWhs, to an access/usage tariff with an ac-

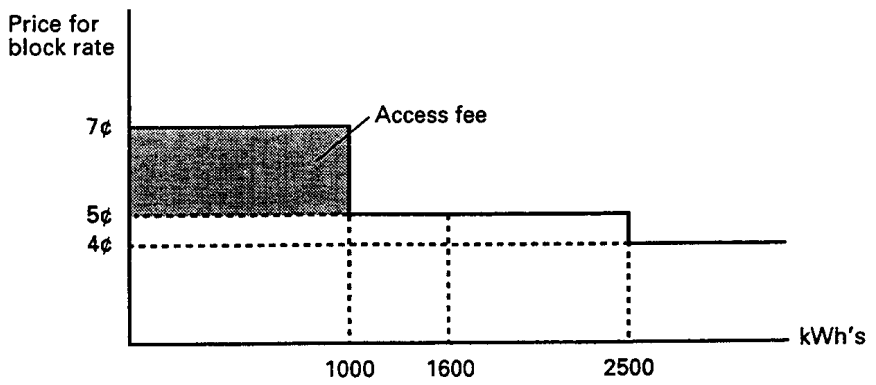

Figure 7.11
Relation of declining block rates to access/usage fees

cess fee of $20 and a usage charge of 5 cents. "Equivalent" in this context means that the customer faces the same bill and the same price at the margin.

The $20 access fee, which makes the two tariffs equivalent, can be visualized in figure 7.11. The first 1,000 kWhs are charged at 7 cents under the block-rate tariff and only 5 cents with the usage fee. The difference of 2 cents for each of the 1,000 kWhs gives the access fee of $20. That is, the access fee that makes the tariffs equivalent for this consumer is the difference between the inframarginal price (i.e., the price in the first block) and marginal price (in the second block) multiplied by the consumption in the first block.

For a consumer whose total consumption is in the third block, the equivalent access/usage tariff is different from that for a consumer in the second block, though it is calculated in the same way. The usage fee is the marginal price for the consumer, which is this case is the price in the third block: 4 cents. The access fee is the amount by which the first block price exceeds the marginal price (seven minus four) times consumption in this block (1,000 kWhs), plus the analogous amount for the second block $((.05-.04) \cdot 1{,}500)$. That is, the access fee that makes the tariffs equivalent is $45.

For a consumer in the first block, the equivalent access/usage tariff consists of a usage fee of 7 cents and no access fee. Note that no one access/usage tariff is equivalent to a given block-rate tariff for all consumers. For those consumers whose consumption falls in the same block under the block-rate tariff, one access/usage tariff is equivalent. But for consumers in different blocks, the equivalent access/usage tariffs are different.

7.3.2 Inverted Block Rates

The choice process of consumers under inverted block rates is represented similarly to that for declining block rates. Figure 7.12 gives the price schedule, outlay schedule, and budget constraint for one such tariff. Opposite from declining blocks, each segment of the outlay schedule has a *greater* slope than the one before, reflecting the fact that the price increases with consumption. Similarly, each subsequent segment of the budget constraint is more steeply sloped. The consumer chooses the point on the budget constraint that attains the highest indifference curve.

Figure 7.13 depicts the choice process in terms of the demand curve. The consumer increases consumption whenever the value of addi-

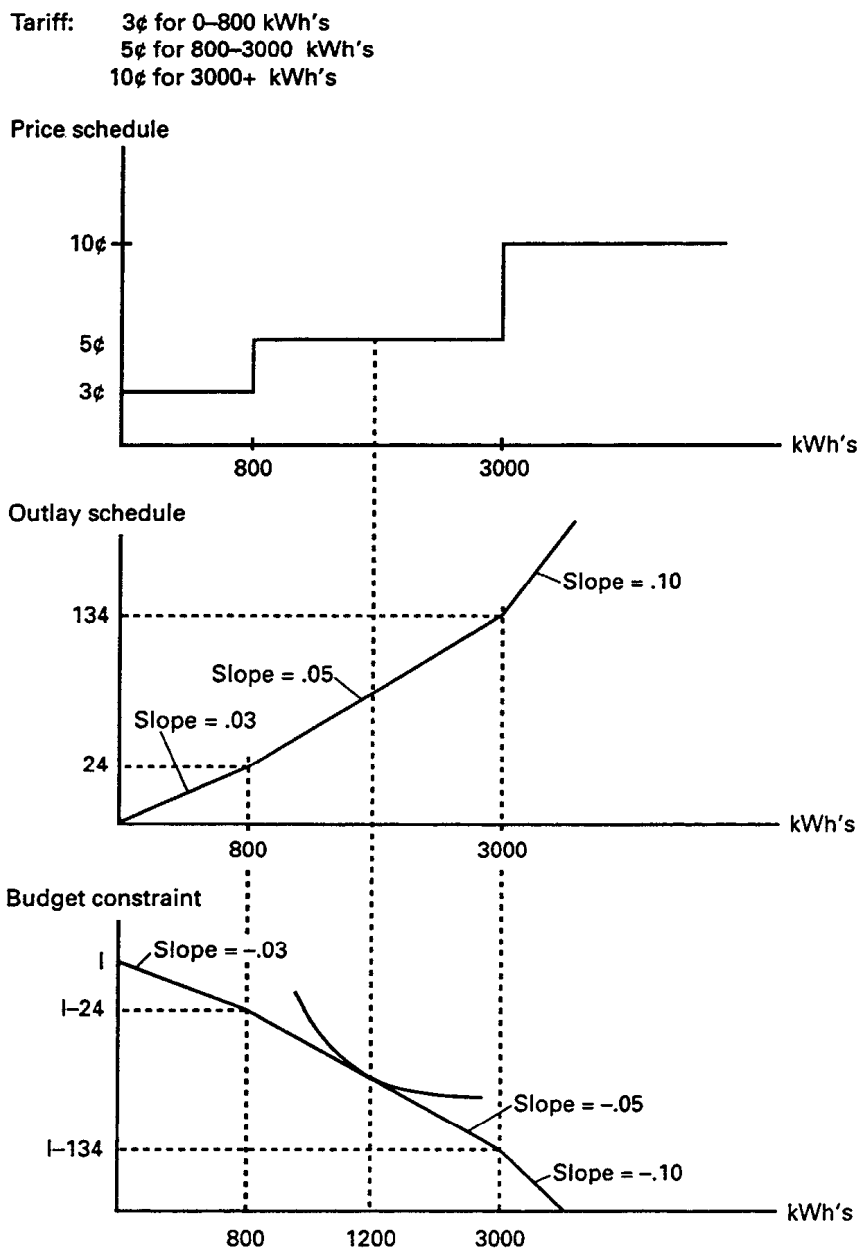

Figure 7.12
Schedules for an inverted block tariff

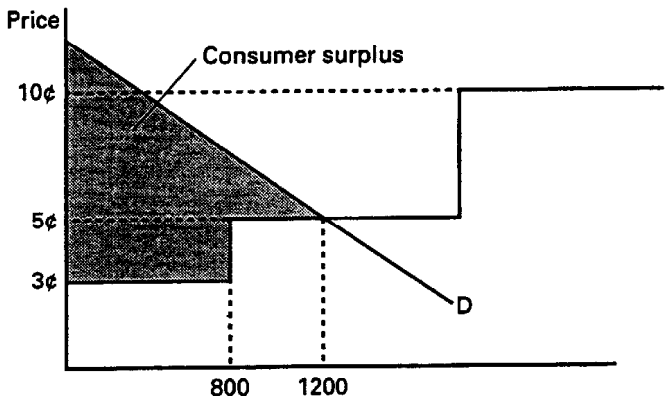

Figure 7.13
Demand curve and chosen consumption

tional consumption, as given by the demand curve, exceeds the price of additional consumption, as given by the price schedule. Consumer surplus is the area below the demand curve and above the price schedule.

While the choice process is essentially the same as for declining block rates, there are some important differences in implications. First, it is not possible under inverted block rates, unlike declining blocks, for utility to be maximized at two different points when indifference curves are convex. The budget constraint bends outward under inverted block rates (reflecting the fact that price becomes successively higher). It is not possible for an inward-bending indifference curve to be tangent to the outward-bending budget constraint at two different points. Visual inspection of panel (c) of figure 7.12 will verify this. The same conclusion can be drawn from the demand curve. Because the demand curve is downward sloping and the price schedule is upward sloping, they can intersect only once. Consequently, the situation depicted in figure 7.9 of multiple intersections under declining blocks cannot occur with inverted blocks.

Second, many consumers will choose to consume at a threshold under inverted blocks, unlike the situation with declining blocks, under which no consumer would choose a threshold. Figure 7.14 illustrates this fact. A consumer with demand curve D_1 will choose to consume 800 kWhs of electricity, which is the first threshold. A consumer with higher demand, at D_2, would also choose this threshold level of consumption. In a standard situation of one price for each

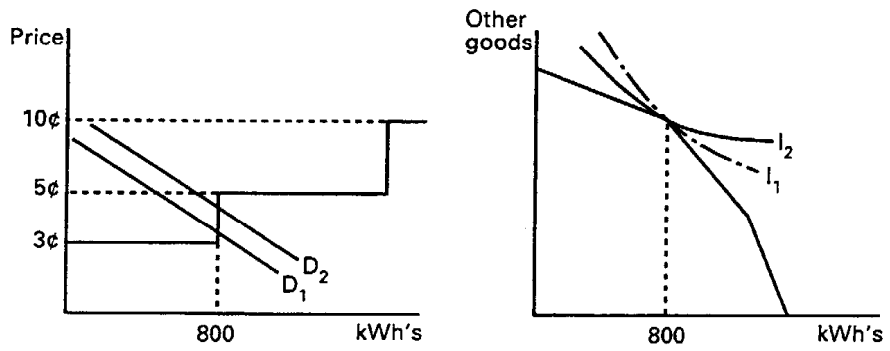

Figure 7.14
Consumer often chooses threshold: "collection point"

good, a consumer with greater demand will consume more of the good. However, the kink in the price schedule for inverted blocks creates a region in which demand can vary and yet consumption remains at the threshold level. As a result, in a population with heterogeneous demand, more consumers (in theory at least) consume at the threshold than at any single other level. In recognition of this phenomenon, the thresholds in inverted block tariffs are called "collection points," because consumers with different demands "collect" there.

The phenomenon can also be seen in terms of the budget constraint and indifference mapping. For any one consumer, the highest indifference curve can often be reached at a kink in the budget constraint, which corresponds to a threshold in the price schedule. Two consumers with different preferences, as represented by indifference curves I_1 and I_2, will both choose the threshold level of consumption. This result is contrary to the standard analysis with one price for each good,in which differences in preferences lead to different levels of consumption.

The third difference between inverted and declining block tariffs relates to their equivalence to access/usage tariffs. In some sense, access/usage tariffs cannot be constructed that are equivalent to inverted block tariffs. Because inframarginal prices are less than marginal prices under an inverted block tariff, the access fee would have to be negative for an access/usage tariff to be equivalent to an inverted block tariff. That is, the access fee would have to be a credit on the consumer's bill offsetting the charges for usage. If a negative access fee is considered infeasible or contrary to the definition of an access/usage

tariff, then the equivalence that exists between declining block and access/usage tariffs does not carry over to inverted block tariffs.

7.3.3 Optimal Block-Rate Tariffs

In this section we first describe the optimal prices and thresholds for a block-rate tariff with a given number of blocks. We then show that increasing the number of blocks in a tariff has the potential to raise surplus in most situations. Because a uniform price (one price for all levels of consumption) is essentially a block-rate tariff with only one block, this result implies that surplus can usually be improved by moving from a uniform price to a block-rate tariff with two or more blocks. *Throughout this discussion, the optimal tariff is considered to be the tariff that provides the greatest consumer surplus while allowing the firm to break even.* For a natural monopolist, the optimal tariff defined in this way does not usually attain the first-best outcome because some price must exceed marginal cost for the firm to break even. We will show, however, that under some circumstances, even the first-best outcome can be achieved with block-rate tariffs.

1. The Optimal Two-Block Tariff

A two-block tariff consists of two prices (one for each block) and the threshold that identifies the consumption levels at which the price changes. Our task is to determine the optimal threshold and the optimal price in each block.

Given a threshold, the optimal prices for the two blocks are essentially Ramsey prices. Consumption up to the threshold and consumption beyond the threshold can be considered two different goods with interrelated demand. The inverse elasticity rule of Ramsey pricing states that price is raised more for the good with the lower elasticity of demand. One would usually expect the elasticity of demand for consumption in the first block to be lower than that in the second block for the following reason. For customers that consume in the second block, the first-block price is an inframarginal price that does not affect their consumption (at least as long as the first-block price is not raised too high). These customers demand the threshold amount of consumption in the first block for a wide range of first-block prices. The only response occurs when the price in the first block is raised so much that the consumer reduces consumption from the second to the first block. For example, the customer in figure 7.15 would consume

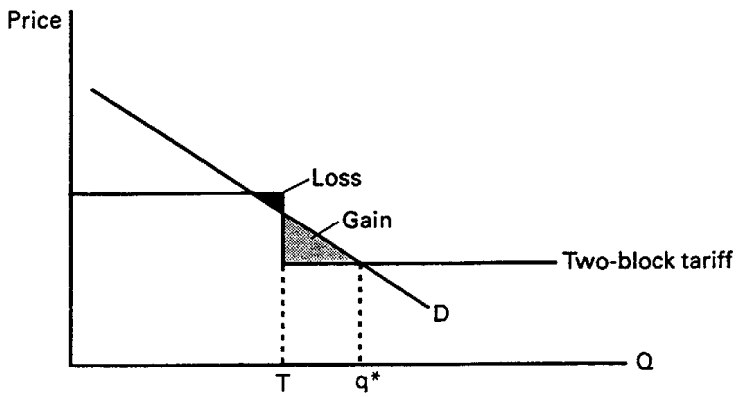

Figure 7.15
Customer in second block of two-block tariff

q^* in total and T (the threshold amount) in the first block until the first-block price was raised so high that the area labeled "loss" increased to a size that exceeded the area labeled "gain." As long as customers continue to consume in the second block, their price elasticity for consumption in the first block is zero. The total (market) elasticity in the first block would therefore be comparatively small because it incorporates the zero elasticities of customers in the second block.

If elasticity is lower in the first block than the second, the inverse elasticity rule suggests that the optimal price in the first block is higher than that in the second block. With the optimal price being higher in the first block, declining rather than inverted blocks are optimal.[16]

At least one price must exceed marginal cost for the natural monopoly to break even. If demand is price responsive in both blocks, then, according to the Ramsey rule, both prices exceed marginal cost (though by different amounts). If, however, demand in the first block is fixed independent of price (at least within the relevant range of price), optimality is attained by setting the price in the second block to marginal

16. It is possible, in unusual circumstances, for the optimal block rates to be increasing rather than declining. The reason is analogous to that described in footnote 8 for an access fee below marginal cost. If the "net" elasticity (that is, the same minus cross-elasticity) is greater in the first block than the second, the Ramsey rule suggests that the optimal price in the first block is lower than that in the second.

cost and raising the price in the first block sufficiently that revenues cover total costs. In this case, first-best optimality is attained.[17]

The discussion so far has taken the threshold as given. The optimal threshold depends on a variety of factors, including the shape and distribution of individual demand curves. A general characterization of the optimal threshold is difficult and cannot, from a practical perspective, be implemented because information on individual demand curves is essentially unobtainable. However, the trade-off that occurs in choosing the threshold is meaningful, and knowing the factors that enter this trade-off can assist in setting the threshold even when the information needed to determine the exact optimum is unavailable.

For a declining block tariff with two blocks, consider the possibility of reducing the size of the first block, that is, reducing the threshold. Two countervailing impacts occur. First, more customers consume in the second block due to the reduction in the size of the first block per se. For example, a customer who consumes 90 units when the threshold is 100 would necessarily consume in the second block if the threshold is reduced below 90. With the price in the second block being closer to marginal cost than in the first block, these customers face a marginal price that is closer to marginal cost. A surplus gain occurs for these customers.

Second, a surplus loss occurs because of the requirement that the firm break even. Because the first-block price exceeds that in the second block, the firm loses revenues if the size of the first block is reduced but the price in each block remains the same.[18] To maintain zero profits, one or both of the prices must be raised. A loss is incurred because at least some customers face marginal prices that are further from marginal cost than with the larger threshold.

There is a trade-off, therefore, in the impacts on surplus from reducing the threshold: the gain from having more customers face the second-block price that is closer to marginal cost, versus the loss due

17. If an access/usage tariff is considered a two-block tariff with a threshold at the first unit of consumption, the optimal pricing rule for two-block tariffs are the same as those presented in section 7.2. In particular, when demand in the first block is fixed (that is, when access demand is fixed), the Coase result applies.

18. Revenues could conceivably increase due to a sufficiently large price response by those customers whose marginal consumption changes from the first to second block. Whenever such an event occurs, prices and the threshold can be adjusted such that consumers and the firm are all better off. At optimality, therefore, a shortening of the first block without a change in prices necessarily reduces revenues.

to one or both of the prices being raised further above marginal cost. If the gain exceeds the loss, the threshold should be reduced. At the optimal threshold, the gain equals the loss. Of course, the same impacts, though in the opposite directions, occur when the threshold is increased. The optimal threshold therefore occurs when the gains equal the losses for changes in the threshold in either direction.

2. Generalization to Three or More Blocks

If the number of blocks in a tariff is specified, the issues regarding optimal prices and thresholds for tariffs with three or more blocks are conceptually the same as with two-block tariffs. Given the thresholds, the Ramsey rule determines the optimal prices (taking into account the cross-elasticities among demand in different blocks). In choosing each threshold, the same trade-off arises as for two-block tariffs. However, the optimal thresholds are even more difficult to identify. With multiple thresholds, the trade-offs for the various thresholds are interrelated because the gains and losses from changing any one threshold depend on the positions of the other thresholds.

So far, we have taken the number of blocks in a tariff as given and examined issues related to the optimal prices and thresholds for those blocks. The question naturally arises: what is the optimal number of blocks? This is the topic of the next two sections.

7.3.4 Pareto-Dominating Block-Rate Tariffs

1. Background Concepts

Results regarding the implications of block-rate tariffs rest on the important distinction between Pareto improvements and surplus improvements. To facilitate this discussion, we first review the basic concepts of surplus and Pareto.

Total surplus is the sum over all consumers of each individual consumer's surplus, plus the profit of the firm (producer's surplus). A change that increases this sum is said to be surplus-improving. Optimal prices are those that provide the greatest possible surplus. A movement, therefore, from nonoptimal to optimal prices constitutes a surplus improvement. This improvement is the rationale for switching to optimal prices.

A change that increases total surplus can hurt some parties. In fact, most movements from nonoptimal prices to optimal prices hurt some parties. For example, a person who has low demand would be hurt

by a switch from uniform pricing (i.e., one price) to the optimal two-block rates, because the optimal price in the first block would exceed the uniform price. A surplus improvement denotes simply that total surplus increases, not that each individual's surplus increases. Consequently, some individuals' surplus can decrease and there would still be an overall increase in surplus as long as other individuals' surplus rises sufficiently that their net sum increases.

By definition, a Pareto improvement is a change that benefits at least one party without hurting any other parties. Because at least one party gains and none loses, a Pareto improvement necessarily increases total surplus and hence is a surplus improvement. However, a surplus improvement need not constitute a Pareto improvement because, as stated above, some parties can be hurt by a change that increases total surplus.

Surplus-improving changes that are not Pareto improving are justified on theoretical grounds by the fact that the parties that benefit from the change could compensate the parties that are hurt such that all parties are better off. That is, any surplus improvement can, with compensations, be made into a Pareto improvement. From a practical perspective, however, the issue is more difficult. Although in theory one can consider lump-sum transfers that do not affect consumption, it is rare that a mechanism for compensation can be found that does not entail some distortion of prices and hence consumption. Furthermore, even if a compensation mechanism were possible, it is difficult to calculate the appropriate amount of transfer to and from each party. Each injured party has an incentive, of course, to overestimate the required compensation; and each party that benefits from a change has an incentive to pretend that the benefit is small or nonexistent. Because of these problems, compensations to accompany a surplus-improving change hardly ever occur in the real world.

Without compensations, it is difficult to justify the ethics of a surplus improvement that is not Pareto improving: how can one person's "happiness" be traded against another person's "pain?" Furthermore, in regulatory as well as other political settings, surplus-improving changes that are not Pareto improving can often be blocked by the parties that would be hurt. These might be the reasons that optimal pricing, which maximizes surplus without regard to which parties gain and lose, is observed in regulated settings with far less frequency than many economists would consider advisable.

Given the ethical and practical advantages of Pareto-improving changes, it is useful to identify tariffs that offer Pareto improvements over current tariffs. Such tariffs are called Pareto-dominating tariffs. One of the appeals of block-rate tariffs is that they can be designed to be Pareto dominating in certain situations. That is, in many circumstances, block-rate tariffs can be designed that benefit some parties without hurting others relative to an existing tariff. In fact, block-rate tariffs can be designed that benefit *all* parties, hurting none.

A block-rate tariff that provides Pareto improvements also, by definition, increases total surplus. However, the optimal block-rate tariff (that is, the tariff that generates the greatest total surplus without regard to who gains or loses) generally provides greater total surplus than any Pareto-improving block-rate tariff. Essentially, Pareto dominance is a constraint on surplus maximization when compensations are not possible: a tariff that maximizes total surplus under the constraint that no party be hurt relative to the current situation cannot provide greater surplus than the tariff that maximizes surplus without this constraint. There is, therefore, a "price," in terms of total surplus, to be paid for Pareto dominance. Whether the ethical and practical advantages of Pareto dominance are worth this price is an issue that regulators must consider.

2. *Design of a Pareto Dominating Tariff*

Willig (1978) and Panzar (1977) have demonstrated the following result.

Result 1: Given an N *block tariff with prices that exceed marginal cost, an* N+1 *block tariff can be designed that Pareto dominates the* N *block tariff.*

Consider first what this result means. A uniform price is essentially a one-block tariff. The result states that if a uniform price is currently being charged and this price exceeds marginal cost (as required for a natural monopolist that is breaking even), it is possible to design a tariff with two blocks under which at least one party would be better off and no parties would be hurt. Similarly, given a two-block tariff with prices above marginal cost, a three-block tariff can be designed that benefits at least one customer and hurts no one. And so on for more and more blocks.

Willig and Panzar each demonstrated this result in a remarkably simple and intuitive fashion. We give the proof for a one-block tariff (uniform price), showing that a Pareto-dominating two-block tariff can be designed that Pareto dominates the uniform price. The generalization to tariffs with more blocks is straightforward.

Under a uniform price, each customer consumes the quantity of the good indicated by its demand curve evaluated at this price and obtains surplus from doing so. To design a Pareto-dominating two-part tariff, focus on the customer with the greatest demand. That is, identify the customer who, at the uniform price, consumes more of the good than any other customer. Label this customer L, the uniform price P_1, and the consumption of this highest-demand customer under the uniform price Q_1. Figure 7.16 depicts the demand curve of this customer, along with the firm's marginal cost curve (which for convenience is assumed to be constant).

Construct a two-block tariff as follows. Set the price in the first block to be the same as the original, uniform price, and make this price applicable for consumption up to the amount consumed by the largest-demand customer. That is, the first block in the tariff consists of price P_1 for consumption up to Q_1 units. For the second block, set price at any level between the price in the first block and marginal cost. Label this price P_2. A tariff constructed in this way is depicted as the heavy line in figure 7.16.

This two-block tariff Pareto dominates the uniform price. First note that no customer is hurt by this tariff relative to the uniform price. Because no customer consumes more than Q_1 under the uniform price, and this same price is applied under the two-block tariff to all consumption up to Q_1, each customer can consume the same amount under the two-block tariff as under the uniform price without a change in its bill or, hence, its surplus. Customers might change their con-

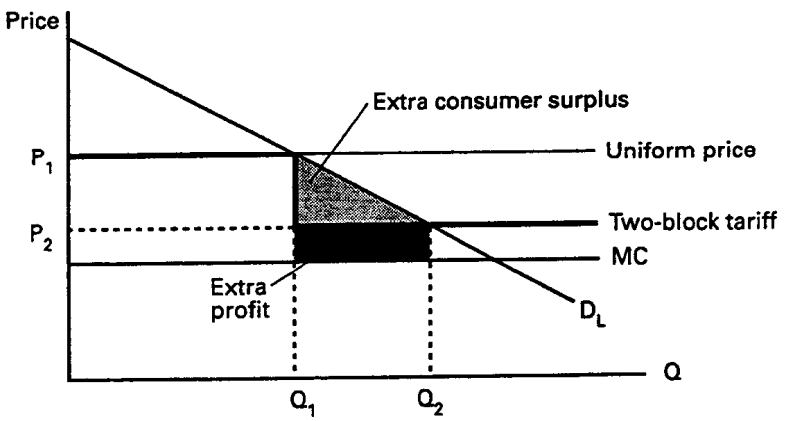

Figure 7.16
Pareto-dominating two-block tariff: customer with largest demand will increase consumption

sumption under the new tariff, but they will do so only if it benefits them. Each consumer therefore obtains either the same or greater surplus under the two-block tariff as under the uniform price.

To demonstrate Pareto dominance, at least one party must actually benefit (rather than staying the same). In fact, we can show that at least one customer *and* the firm benefit.

Consider the customer with the largest demand under the uniform price. This customer consumes Q_1 under the uniform price. Under the two-block tariff, this customer can continue to consume Q_1 with no change in outlay because the original price, P_1, applies for consumption up to Q_1. However, additional consumption can now be purchased by the customer at a lower price: at P_2 instead of P_1. The consumer will respond to the lower price by increasing consumption (given a downward-sloping demand curve), moving from Q_1 to Q_2 in figure 7.16. This additional consumption at the lower price provides the customer with extra surplus, represented by the lightly shaded area in the figure. The largest customer therefore benefits from the tariff.

Consider now the effect on the firm of this customer's increased consumption. The firm incurs the same costs and obtains the same revenues for the first Q_1 units of consumption as it did under the uniform price, because the price for this quantity of consumption is the same under the two-block tariff. For consumption beyond Q_1, the price, while lower, is still above marginal cost. Hence, any additional consumption generates additional profit for the firm. The extra profits generated by the increased consumption of the largest-demand customer are $(Q_2 - Q_1)(P_2 - MC)$, which is the darkly shaded area in the figure.

The largest-demand customer benefits from the two-block tariff, and the firm benefits from this customer's response. What about all the other customers? Clearly, as stated above, no customer is hurt by the two-block tariff because each customer can continue to consume the same quantity as under the uniform price and pay the same amount. It is possible, however, that some of these customers will choose to increase their consumption. For example, consider a customer with a demand curve as depicted in figure 7.17. Under the uniform price, the customer would purchase q_1 units, which is less than that of the largest-demand customer (that is, less than Q_1). Under the two-block tariff, the customer would increase consumption to q_2 because its sur-

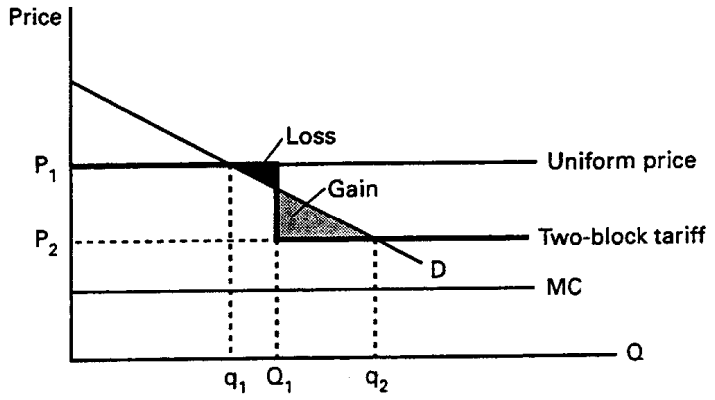

Figure 7.17
Pareto-dominating two-block tariff: a customer with lower demand might increase consumption

plus would increase by doing so. The change in surplus consists of the loss incurred from increasing consumption from q_1 to Q_1 (because the price exceeds the consumer's willingness to pay as indicated by the demand curve) and the gain obtained from increasing consumption from Q_1 to q_2. Because, given the way the demand curve is drawn, the gain exceeds the loss, the customer will choose to increase consumption from q_1 to q_2 to obtain this net gain in surplus.

Any customer that chooses to increase consumption under the two-block tariff necessarily gains from doing so (otherwise it would remain at its original level of consumption). Furthermore, because price exceeds marginal cost in both blocks of the two-block tariff, any increase in consumption generates additional profits for the firm.

In short: the two-block tariff that we constructed increases the surplus of the largest customer and the profits of the firm. It does not hurt any other customers and might, depending on the shapes of the individual demand curves, benefit some of them.

Actually, all customers can benefit from an appropriate two-block tariff. Because the firm earns extra profit under the two-block tariff described above, it is possible to lower the price in the first block and still have the firm as well off as under the uniform price. That is, the extra profit, or some portion of it, can be refunded to customers though a lower first-block price. If all the profit is refunded through this first-block price reduction (as would occur if the firm were required to earn no more than zero profit), then the change from a uniform price to

the two-block tariff would benefit all consumers without hurting the firm. If a portion, but not all, of the extra profit is refunded, then all parties, including the firm, benefit from the switch to the two-block tariff. Stated in the strongest terms: given a uniform price that exceeds marginal cost, a two-block tariff can be designed that is better for all customers and the firm than the uniform price.

The same concepts can be applied to show that, given a two-block tariff with prices above marginal cost, a Pareto-dominating three-block tariff can be designed; and so on for more blocks. Figure 7.18 depicts a two-block tariff along with the demand curve for the customer whose consumption under the two-block tariff exceeds that of all other customers. Label the consumption of this largest-demand customer as Q_2. A Pareto-dominating three-block tariff is constructed as follows: the first block is the same as the first block of the two-block tariff; the second block has the same price as the second block of the two-block tariff, but instead of extending indefinitely, this block stops at Q_2; the third block starts at Q_2 and has a price that is closer to marginal cost than the price in the second block. Because prices are the same for consumption up to Q_2, no customer is hurt by this three-block tariff. The largest customer increases consumption under this new tariff, because the price of additional consumption for this customer is lower. This customer gains surplus and the firm earns extra profit, as designated by the shaded areas in the figure. Other customers might also increase consumption; if they do so, their surplus increases and the firm earns more profit. If the firm's extra profit is refunded through

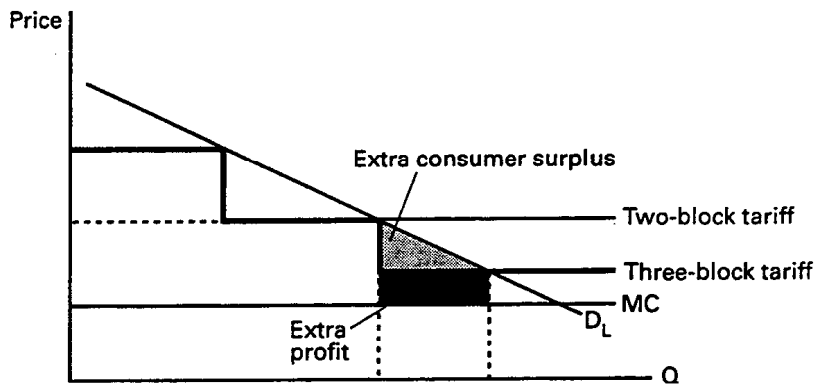

Figure 7.18
Pareto-dominating three-block tariff: the customer with largest demand will increase consumption

lower rates in the first and second blocks, all customers benefit by the change in tariff.

7.3.5 The Optimal Number of Blocks

Section 7.3.3 described, for a tariff with a *given* number of blocks, the optimal price in each block and the optimal thresholds between blocks. Result 1 in section 7.3.4 allows us to compare surplus under optimal tariffs with *different* numbers of blocks. As before, the term "optimal" is used throughout this section to refer to the tariff that provides the greatest surplus while allowing the firm to break even. In some cases, the optimal tariff attains first-best optimality (e.g., when demand in the first block is fixed). Otherwise, the optimal tariff attains second-best optimality.

Result 2: The optimal N + 1 *block tariff provides strictly greater surplus than the optimal* N *block tariff whenever prices exceed (do not equal) marginal cost under the optimal* N *block tariff.*

This result is one of the primary motivations for implementing block-rate tariffs. Consider first what it says. A uniform price (that is, one price for the good independent of consumption) is a one-block tariff; in a natural monopoly situation, the optimal uniform price is average cost, which exceeds marginal cost. The result states that greater surplus is obtained with the optimal two-block tariff than with this optimal uniform price. Similarly, optimal prices in a two-block tariff exceed marginal cost whenever demand in the first block is not fixed. The result states that in this situation, the optimal three-block tariff attains greater consumer surplus, while maintaining zero profits for the firm, than the optimal two-block tariff. And similarly for four-block tariffs compared to three-block, and so on.

Because of the difficulty of characterizing the optimal N block tariff, a heuristic proof of this result (that is, a proof that determines the optimal N block and $N+1$ block tariffs and calculates surplus under each) has not been developed. Instead, the result is seen as a corollary to result 1. Consider first the optimal two-block tariff compared to the optimal uniform price (that is, one-block tariff). The optimal uniform price is average cost, which exceeds marginal cost if the firm is a natural monopolist. Because price exceeds marginal cost, result 1 states that a two-block tariff can be designed that Pareto dominates this optimal uniform price. This Pareto-dominating two-block tariff neces-

sarily increases surplus. The optimal two-block tariff provides at least as much surplus as the Pareto-dominating tariff.[19] Consequently, the optimal two-block tariff provides strictly greater surplus than the optimal uniform price.

The same arguments apply in comparing the optimal two- and three-block tariffs. Under the optimal two-block tariff, both prices exceed marginal cost unless demand in the first block is fixed. When first-block demand is fixed, the optimal two-block tariff achieves the first-best outcome: price in the second block is set to marginal cost and the first-block price is set sufficiently high for the firm to break even. This means that, unless the first-best outcome is achieved by the two-block tariff, a three-block tariff can be designed that Pareto dominates the optimal two-block tariff. The optimal three-block tariff attains at least as much surplus as the Pareto-dominating three-block tariff and therefore provides greater surplus than the optimal two-block tariff. And so on for tariffs with more blocks. The basic finding is that, until the first-best outcome is achieved, surplus can be improved by adding more blocks to a block-rate tariff and adjusting the prices and thresholds accordingly.

7.3.6 Equity Considerations

The discussion of optimal and Pareto-dominating tariffs highlights the surplus advantages of having prices *decline* as consumption increases. The optimal block-rate tariff consists of declining blocks, and Pareto-dominating tariffs are constructed by adding a final block with

19. In fact, it is possible, and perhaps even likely, that the optimal two-block tariff provides considerably greater surplus than the Pareto-dominating two-block tariff. Given the way the Pareto-dominating tariff was designed, the only customer that is guaranteed to benefit is the customer with the largest demand under the uniform price. This surplus improvement might be very small, whereas the optimal two-block tariff (which is not constrained to be Pareto dominating) could provide a substantial surplus improvement. It is also important to note that other tariffs might exist that are Pareto-dominating and provide greater surplus than the tariff we constructed to guarantee Pareto dominance. In constructing our Pareto-dominating tariff, our goal was to prove that at least one Pareto-dominating tariff exists. Other Pareto-dominating tariffs might also exist, and these might, and probably will, provide greater surplus. Among the set of tariffs that Pareto dominate a given tariff, the tariff that provides the greatest total surplus is called the optimal Pareto-dominating tariff. An interesting area of research is to compare total surplus under the optimal tariff with that under the optimal Pareto-dominating tariff. The difference is, in a sense, the "price" of Pareto dominance, that is, the surplus that is potentially foregone in return for changing tariffs in a way that hurts no one.

a lower rate. From an equity perspective, however, *inverted* block rates might be preferable to declining blocks in many situations. Under declining block rates, customers with lower demand face relatively higher prices. Insofar as customers have low demand because their income is low, declining rates force lower-income customers to face higher prices. Inverted block rates, which are lower for small levels of consumption and increase for greater consumption, might be preferable from an equity perspective in cases like this.

Consumption does not always increase with income. For example, low-income housing units might be built with electric heating, rather than gas or oil heating, because electric heaters are less expensive to install. The households who occupy these units would demand more electricity in their attempts to stay warm than a higher-income household whose home has gas or oil heat. In cases like this, declining block rates can be advantageous from an equity as well as total surplus perspective.