6 Surplus Subsidy Schemes

6.1 Introduction

Most of the regulatory mechanisms examined so far attempt to achieve the *second-best* outcome.[1] If the regulator is able to subsidize the firm, the first-best outcome is feasible. By subsidizing the firm for the losses it incurs at marginal-cost prices, the firm can remain solvent and continue to produce at these prices indefinitely. The question then becomes: what regulatory mechanisms can the regulator use to attain the first-best outcome, given that it is able to subsidize the firm?

Were the regulator able to know the firm's cost and demand curves, the task of regulation would be simple. The regulator would require the firm to set its prices at marginal cost and then subsidize the firm for the difference between revenues and the minimum cost of producing the output demanded. Unfortunately, the regulator seldom knows the cost and demand curves of the firm. Without knowing costs, the regulator cannot determine marginal costs and, just as important, does not know the minimum cost of producing a given output. If the regulator relied on the firm to report its costs, the firm would clearly have an incentive to misreport. By reporting a marginal cost that is higher than actual, the firm could keep prices above their first-best level. And by reporting higher-than-actual total costs, the firm could increase its subsidy and earn additional profits at any price level. Even if the regulator were able to audit the firm costlessly and accurately, the first-best outcome would still not be attained. Without knowing the firm's cost curves, the regulator would have to subsidize the firm on the basis of *incurred* costs, rather than minimum costs of

1. The one exception is price discrimination, which, as shown in chapter 2, attains first-best optimality.

production. The firm would have no incentive to produce efficiently, because it would earn zero profits whether it was efficient or wasteful. Total costs and marginal costs (and hence prices) would inevitably be above their first-best levels.

Several regulatory mechanisms have been proposed to induce the firm to price at marginal cost and produce efficiently without the regulator knowing the costs of the firm. Each of these procedures involves the regulator subsidizing the firm on the basis of the consumer surplus that the firm's pricing decisions generate. By letting the firm benefit (through the subsidy) whenever it acts in a way that benefits consumers (that is, increases consumer surplus), the regulator is able to induce the firm to act in accordance with social goals.

Three mechanisms are described in this chapter. Loeb and Magat (1979), who seem to have been the first to suggest this type of regulation, introduce the important concept that transferring consumer surplus to the firm induces the firm to behave optimally. Sappington and Sibley (1988) propose a mechanism that transfers to the firm only the period-to-period *change* in consumer surplus. This procedure, in addition to inducing first-best prices and efficient production, provides the firm with only zero profits in equilibrium, rather than positive profits as under Loeb and Magat's procedure.[2] Both the L-M and S-S procedures require that the regulator have information about the firm's demand curve, at least in the vicinity of optimal prices. Finsinger and Vogelsang (1981, 1982, 1985) have developed a procedure that can be implemented without this information, that is, without knowledge of *either* the demand or cost curves. This procedure uses an *approximation* to the period-to-period change in consumer surplus. Because the subsidy is based on an approximation rather than the actual change in consumer surplus, equilibrium is attained more slowly than under the other procedures. However, once attained, equilibrium consists of the first-best outcome.[3]

---

2. The L-M procedure can also result in zero profits if there are many potential producers and the regulator can auction the right to be the monopoly among these potential producers. See section 6.2.

3. Chronologically, Sappington and Sibley's analysis developed from issues raised by the earlier work of Finsinger and Vogelsang. Finsinger and Vogelsang suggested the use of the period-to-period change in surplus and developed an approximation to this change that does not require information on demand curves. Sappington and Sibley, responding to the fact that the F-V mechanism can take many periods to attain equilibrium, proposed a means of using information on the demand curve to attain equilibrium more quickly. It is pedagogically useful to present the procedures in the reverse

6.2 Loeb and Magat

Loeb and Magat have proposed a mechanism that induces the firm to charge the optimal price and produce efficiently even when the regulator does not know the firm's costs. Under this mechanism, the regulator allows the firm to chose its price without constraint. Given the firm's price, the regulator subsidizes the firm by the amount of consumer surplus that is generated at that price. To calculate this quantity, the regulator must possess information on the firm's demand curve, but not its costs. (Recall that consumer surplus is the area under the demand curve and above the price: the shaded area in figure 6.1 for price P_0. The size of this area depends on the demand curve but not on costs.)

The firm's total profits with this subsidy are the sum of the producer's surplus (that is, its profits without the subsidy) and consumer surplus. The mechanism therefore gives all surplus to the firm. Because all surplus accrues to the firm, it chooses the price that provides the greatest total surplus, which by definition is the first-best outcome. Stated alternatively: because total surplus is greatest when price is set at marginal cost, and because the firm's profit with the subsidy

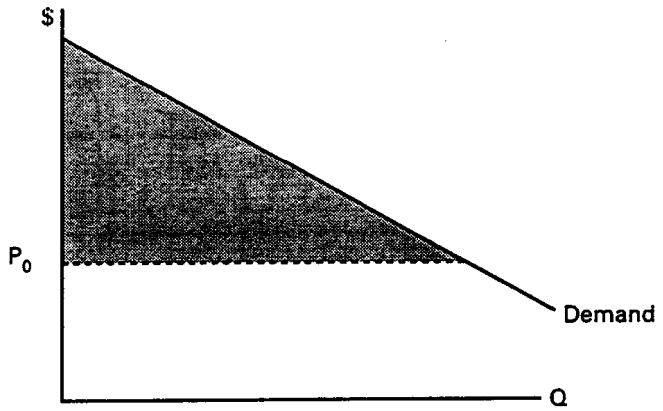

Figure 6.1
Subsidy under L-M: consumer surplus

order. The L-M procedure shows the value of transferring consumer surplus to the firm. The S-S procedure illustrates that the same effect can be obtained by transferring the period-to-period change in surplus. Then, the F-V procedure shows that the period-to-period change in surplus can be approximated when the regulator does not know the demand curve, with the only loss being the speed at which the first-best outcome is attained.

consists of the total surplus, the firm maximizes its own profits by pricing at marginal cost.

The result is the same with many products as with one: the multiproduct firm maximizes its profit (which consists under the subsidy of total surplus) by setting all prices equal to their marginal costs. Furthermore, the firm produces efficiently. Any reduction in costs translates into an increase in surplus, which the firm keeps; the firm therefore makes the most profit by producing at the least possible cost.

The L-M procedure is an extreme, and thereby illuminating, example of the general principle that optimality is attained by creating consistency between the goals of the firm and the goals of the regulator. The goal of the regulator is to maximize total surplus, and the goal of the firm is to maximize profits. The L-M procedure makes these two goals consistent by giving all surplus to the firm, such that the firm's profits *are* the total surplus. The firm, in maximizing its own profits, maximizes total surplus.

This way of attaining consistency results in an outcome that, while efficient, might not be considered equitable. The firm obtains all the surplus, and consumers obtain none.[4] Loeb and Magat suggest two

4. If the regulator considers it inequitable for the firm to obtain all the surplus, the regulator's goal is apparently more complex than simply maximizing total surplus. Suppose the regulator's goal is to maximize the *weighted* sum of consumer and producer surplus rather than the simple sum, with the weights representing the relative importance the regulator places on surplus for the two parties. Baron and Myerson (1982), Sappington (1983), and Laffont and Tirole (1986) derive optimal pricing and subsidy policies under this more general goal, along with the assumption, as in L-M, that the regulator knows demand but not costs. When consumer and producer surplus are weighted equally (that is, the regulator maximizes total surplus), the L-M mechanism is of course optimal: the firm is subsidized by the amount of consumer surplus, sets price equal to marginal cost, and minimizes costs. However, when the producer's profit is weighted less than consumer surplus, these studies indicate that the regulator's goal is better met by having a smaller subsidy and a price above marginal cost. The reason for these results hinges on the fact that the subsidy is a transfer from consumers to the producer. This transfer, in itself, reduces the regulator's welfare measure when the producer (who receives the transfer) is weighted less than consumers (who provide the transfer). The purpose of a subsidy is to induce the firm to price closer to marginal cost, which generates additional surplus. A subsidy is justified if the surplus gained from pricing nearer marginal cost is greater than the loss incurred by the transfer from consumers to producers. There comes a point, however, as the subsidy is raised and prices move closer to marginal cost, that an additional subsidy generates a greater loss due to the transfer than a gain due to pricing nearer marginal cost. That is, the regulator's welfare measure is higher by not subsidizing as much and allowing the firm to price somewhat above marginal cost. (Of course, when consumer and producer

methods for correcting this inequity. First, the monopoly could be auctioned. That is, different firms could bid for the right to be the monopoly producer of the good, with the regulator choosing the firm with the highest bid. Supposedly, each firm would be willing to bid up to the maximum surplus that can be obtained in the market, because, under the terms of the regulation, the firm knows that it will be able to earn that much if it wins the auction and becomes the monopoly. The highest bid will therefore be essentially equal to the total surplus that is attainable in the market.[5] When the winning firm becomes the monopolist, it will earn a profit, including subsidy, that is the same as the amount it paid to become the monopolist. On net, the firm will earn zero profits (profits including subsidy and minus auction bid), and all surplus will accrue to consumers.

Second, the regulator can subsidize the firm by a *portion* of consumer surplus rather than the entire amount. Suppose, for example, that the regulator knows that the firm will not choose a price over P_a. The regulator can then subsidize the firm for the portion of consumer surplus between P_a and the price the firm actually charges. This subsidy is the shaded area in figure 6.2, where P_b is the price that the firm charges. Stated alternatively, the subsidy is the consumer surplus at the firm's chosen price P_b minus the consumer surplus at P_a.

---

surplus are weighted equally, the transfer of surplus from consumers to the producer has no effect, in itself, on the welfare measure. As a result, there is no loss from raising the subsidy as high as necessary to induce the firm to price exactly at marginal cost.)
It is important to note that the optimality results obtained by these studies apply to situations in which regulation occurs entirely in one period. When the regulator can use information in one period to determine prices and/or subsidy in the next, other mechanisms can be utilized that bring prices to marginal cost in equilibrium with a subsidy that is the minimum necessary to keep the firm in business. (The S-S and F-V mechanisms, described in the following sections, are examples.) Total surplus is maximized under these mechanisms, because price equals marginal cost. And since profits are the minimum feasible (zero, through subsidy), consumer surplus is also as great as possible. With total surplus and consumer surplus both as high as possible, the regulator's welfare measure (in equilibrium) is necessarily as high as possible no matter what weights the regulator places on consumer and producer surplus (provided of course that producer surplus is not weighted more than consumer surplus). Stated succinctly: with multiperiod regulation, the first-best outcome with unequal weighting of consumer and producer surplus is the same as the first-best outcome with the weights being equal. In either case, prices are set at marginal cost and the firm is subsidized sufficiently to just break even.

5. The winning bid will equal the entire surplus if each producer faces the same costs or if there are many bidders, each of whom has an independent assessment of costs. Riordan and Sappington (1989) derive optimal methods for awarding the monopoly franchise under more general conditions.

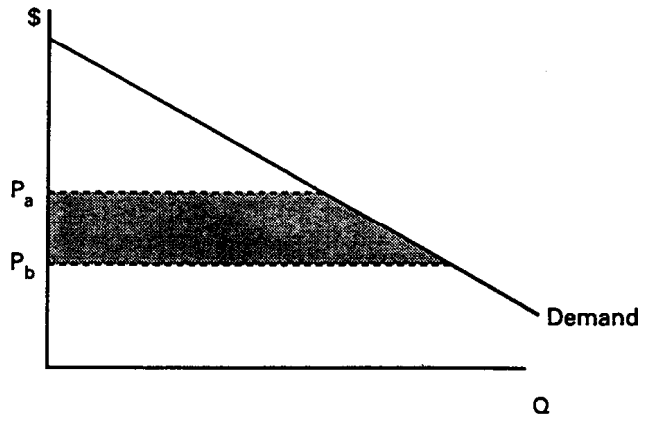

Figure 6.2
Subsidy as portion of consumer surplus

Subtracting a fixed amount from the firm's profits does not change the firm's *relative* profits at each outcome: the firm's profits including subsidy are still highest when total surplus is highest. The firm will again choose the surplus-maximizing (that is, first-best) price and will produce with cost-minimizing inputs. And yet, because the firm obtains only a portion of consumer surplus, some surplus is retained by consumers.

The difficulty with this approach is that the regulator does not generally know how high to set P_a. P_a must be sufficiently above the first-best price to provide enough subsidy to the firm for it to break even. Otherwise, the firm will choose to stop production rather than produce at a loss. However, because the regulator does not know the firm's costs, it does not know how high P_a must be. The regulator, in making sure that P_a is sufficiently high, could easily establish a subsidy that is far larger than needed to induce the firm to behave optimally. This difficulty is the motivation for the S-S and F-V procedures described below.

6.3 Sappington and Sibley: The Incremental Subsidy Surplus (ISS) Scheme

Sappington and Sibley have introduced a multiperiod regulation mechanism in which the regulator uses information on the firm's prices, revenues, and expenditures in one period to determine the subsidy the firm obtains in the next period. The mechanism is based on the

concept that the firm need not receive the *entire* surplus in order to choose first-best outcomes. Rather, in each period, the firm can be allocated the improvement, or gain, in surplus that its actions in that period generate. Under this subsidy, the firm will, in each period, choose to provide the greatest possible *improvement* in surplus. This period-to-period improvement leads over time (and in fact very quickly) to surplus being as great as possible.

Suppose the firm in period t is charging P^t and selling the quantity Q^t demanded at that price. The consumer surplus generated at this price is CS^t, which is the area under the demand curve and above P^t. The firm expends E^t producing the output. Expenditures E^t are perhaps higher than the minimum cost of producing the output due to inefficiency. (We show below that the firm will not waste in equilibrium, such that E^t actually does equal minimum cost in equilibrium.) The firm earns profit from its operation, called operating profit, of $O^t = P^tQ^t - E^t$, that is, revenue minus expenditures. This operating profit includes any waste the firm incurs in its operations and excludes any subsidy it receives. By definition, total surplus is the sum of consumer surplus and operating profit.

The regulator allows the firm to choose its price and expenditures in each period. The regulator subsidizes the firm on the basis of the *extra* consumer surplus it generates each period. In particular, the regulator provides the following subsidy in period t:

$$S^t = (CS^t - CS^{t-1}) - O^{t-1}.$$

The first term is the change in consumer surplus from the previous period to the current period; that is, it is the improvement in consumer surplus that the firm generates in the current period. Visually, this quantity is the area $ABCE$ in figure 6.3 for a firm that charges P^0 in one period and P^1 in the next. The second term is the firm's operating profit in the previous period. Taken together, the subsidy is therefore the improvement in consumer surplus minus the previous period's operating profit.

The reason this subsidy is effective becomes clear when the subsidy is added to the firm's operating profit in each period. Under this subsidy, the firm's *total* profit in each period is its operating profits plus the subsidy

$$\pi^t = O^t + S^t.$$

Substituting in the formula for the subsidy, total profit is

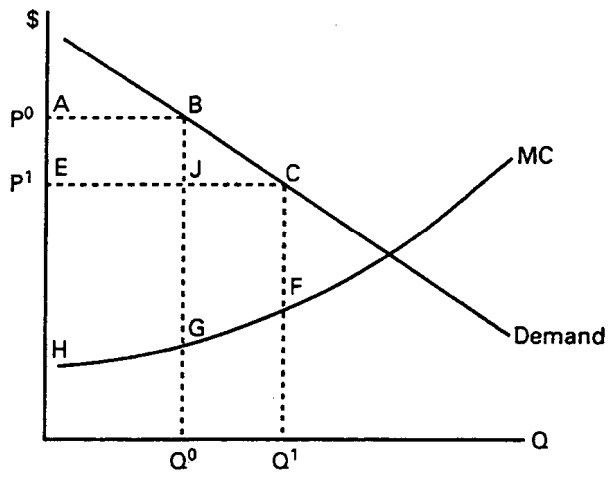

Figure 6.3
Effects of price reduction

$$\pi^t = O^t + (CS^t - CS^{t-1}) - O^{t-1}.$$

Rearranging:

$$\pi^t = (O^t - O^{t-1}) + (CS^t - CS^{t-1}).$$

That is, the firm's total profit in each period is the change in operating profit and consumer surplus from the last period. Because total surplus is the sum of operating profit (producer surplus) and consumer surplus, the firm's total profit in each period under this subsidy is the change in total surplus since the last period. Visually, and ignoring the possibility of waste, the firm's total profit in figure 6.3 is area *BCFG*, the increase in total surplus.[6]

Given that the firm's profit is the change in total surplus, the firm maximizes its profit by generating the greatest possible improvement in surplus in each period. No matter where the firm starts out, the greatest improvement in surplus is attained by the firm moving to the first-best outcome, namely, to marginal-cost prices with no waste. In

6. When price is reduced from P^0 to P^1, consumer surplus increases by *ABCE*. Operating profit in period zero is area *ABGH* (the difference between price and marginal cost for each unit sold). The firm's subsidy in period one is therefore area *ABCE* minus area *ABGH*. Operating profit in period one is area *ECFH*. The firm's total profit in period one is its operating profit in that period plus its subsidy: *ECFH* + (ABCE − ABGH), which is area *BCFG*. Note that area *BCFG* is the increase in total surplus that results from the price reduction from P^0 to P^1: it is the difference between the value of each unit to consumers (as denoted by the demand curve) and the marginal cost of each unit, summed over all the extra units sold.

fact, because future profit is discounted relative to current profit, the firm will want to obtain the largest improvement in surplus *as soon as possible*. Therefore, the firm will move to the first-best outcome in the first period after this subsidy mechanism has been established.

This fact can be demonstrated visually. Recall that if the firm lowered its price in period one from P^0 to P^1, then its profit, including subsidy, in period one is area *BCFG* in figure 6.3. The firm will therefore choose the price in period one (that is, will choose P^1) in such a way as to make area *BCFG* as large as possible. As P^1 is lowered, points *C* and *F* move out, such that area *BCFG* increases in size. The area is as large as possible when P^1 is lowered to the level shown in figure 6.4, namely to marginal cost. That is, the firm maximizes its profit in period one by setting its price in period one at the first-best level.

Under the S-S scheme, the firm's profit changes over time in a particular way. Before regulation is imposed, the firm earns some profits, which can be denoted π^0. The firm does not choose the first-best outcome prior to regulation; in fact, this is the reason to impose regulation. Two types of losses occur relative to the first-best outcome. First, the firm might waste. Second, a "deadweight loss" occurs because price is above marginal cost. This deadweight loss is the difference between the surplus attained at marginal cost prices and the surplus attained at the prices the firm charges, independent of any waste. Visually, it is the shaded area in figure 6.5 given that the firm is pricing at P^0.

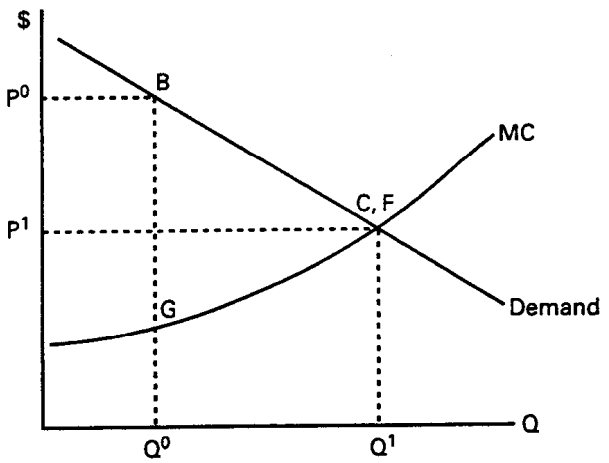

Figure 6.4
Period 1 price change for firm under S-S regulation

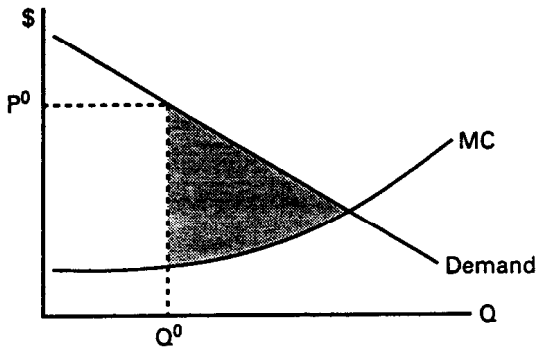

Figure 6.5
Deadweight loss due to pricing above marginal cost

The subsidy scheme is imposed in period one. The firm's total profit in period one is equal to the improvement in surplus that the firm generates in period one. The firm moves to the first-best outcome in period one to generate the maximum possible improvement in surplus (which, under the scheme, translates into maximum profit for the firm). The firm's profit in this period is therefore equal to the deadweight loss and waste that the firm incurred prior to regulation.

In period two the firm remains at the first-best outcome, because no further improvement in surplus is possible.[7] Because consumer surplus does not increase, the subsidy to the firm is $S^2 = -O^1$. That is, the firm is subsidized by the amount of operating profits of the firm in period one. Because the firm was at the first-best outcome in period one with prices equal to marginal cost, the firm, being a natural monopoly, incurred negative operating profit. The subsidy in period one is therefore equal to the operating loss that the firm incurs at the first-best outcome. The firm's total profit in period two is the sum of its operating profit and its subsidy: $\pi^2 = O^2 + S^2 = O^2 - O^1$, which equals zero since the firm operates at the first-best outcome in both periods one and two such that $O^1 = O^2$.[8]

7. The firm will not move away from the first-best outcome after it has reached it because doing so would result in the firm's total profits in that period being negative. The firm would be able to make up the loss in the following period by moving back to the first-best outcome (being subsidized for the improvement); however, because the firm discounts future profit relative to current profit, it will not choose to incur a loss in one period that is just made up in the next period.

8. The firm's profit in period two can be derived more directly. With the subsidy, total profit in any period is the additional surplus the firm generates in that period. Since the firm remains at the first-best outcome in period two, surplus does not change and total profit, including subsidy, is therefore zero.

All subsequent periods are the same: the firm stays at the first-best outcome, charging marginal cost prices and not wasting, and receives a subsidy that allows it to just break even.

In short, in the first period of regulation, the firm moves to the first-best outcome. Its profit in this period, including subsidy, is equal to the amount of waste and the deadweight loss that the firm incurred prior to regulation. In the second and subsequent periods, the firm stays at the first-best outcome and its profit, including subsidy, is zero.

6.4 Finsinger and Vogelsang: An Approximate Incremental Surplus Subsidy (AISS) Scheme

To implement the subsidy scheme proposed by Sappington and Sibley, the regulator needs to know the shape of the demand curve, at least in the region between the firm's price prior to regulation and the optimal price.[9] Finsinger and Vogelsang have proposed a mechanism that is conceptually similar to that of Sappington and Sibley, but does not utilize information on the demand curve. Rather than providing a subsidy to the firm on the basis of the exact improvement in consumer surplus, F-V subsidizes the firm on the basis of an approximation to this improvement. The approximation is calculated on the basis of information that the regulator can observe directly, namely, the prices and quantities sold in each period.

Under the F-V scheme, the firm receives a subsidy each period equal to

$$S^t = Q^{t-1}(P^{t-1} - P^t) - O^{t-1}.$$

The first term is the approximate improvement in consumer surplus. It is area *ABJE* in figure 6.3: the change in price multiplied by the

9. Sappington and Sibley point out that their mechanism can be used even when the regulator does not know the demand curve exactly, provided that the regulator possesses the same information as the firm regarding the unknown demand curve. For example, demand might vary and its realization in each period be unknown to both the firm and the regulator prior to the setting of prices. If the regulator and firm both know the distribution of demand, the S-S procedure will induce the firm to price at expected marginal cost (that is, marginal cost of the output demanded at that price, averaged over all possible levels of demand). Similarly, demand might be fixed though unknown. If the regulator and firm have the same a priori concepts about the probability that demand is at a certain level, the procedure will induce the firm to choose prices that are optimal given the regulator's concepts of demand. However, if the regulator and firm are not symmetrically informed about demand, the procedure does not necessarily lead to the first-best outcome.

quantity sold in the previous period. The exact improvement in consumer surplus is area *ABCE*, which enters in the subsidy under the S-S procedure. The F-V scheme differs from the S-S scheme in that the subsidy to the firm for a given price change is less by the amount *BCJ*. The second term is the operating profit in the previous period, which, as in S-S, is subtracted from the improvement in consumer surplus.

The total profit of the firm in each period is the operating profit in that period plus the subsidy:

$$\begin{aligned}\pi^t &= O^t + S^t \\ &= O^t + Q^{t-1}(P^{t-1} - P^t) - O^{t-1} \\ &= (O^t - O^{t-1}) + Q^{t-1}(P^{t-1} - P^t).\end{aligned}$$

Total profit is therefore the change in operating profit since the last period plus the approximate improvement in consumer surplus. That is, total profit in each period is the approximate improvement in total surplus for the period.

It is useful to visualize this approximate improvement in total surplus. In figure 6.3, the *exact* increase in total surplus is area *BCFG*. Under the S-S procedure, this area is the total profit (operating profit plus subsidy) that the firm would obtain in the period. Under the F-V procedure, the firm's total profit in each period is not equal to this exact improvement in surplus, but rather to an approximation. The firm's subsidy is less than the true change in consumer surplus by the amount *BCJ*. Total profit under F-V regulation is therefore area *JCFG*, which is less than the exact improvement in surplus by area *BCJ*.

Consider now the behavior of the firm under the F-V scheme. Suppose first that the firm does not discount future profits relative to current profits. Then it would change prices each period to obtain the largest possible sum of profits over time. The firm can obtain essentially all the deadweight loss as profits if it takes many very small price reductions over time. Figure 6.6 depicts this fact. The original deadweight loss (that is, the deadweight loss at price P^0) is *AJG* in either graph. This is the amount of total profit that the firm would receive under the S-S scheme, under which the firm receives exactly all the improvement in surplus. Under the F-V scheme, suppose the firm took two steps, reducing price from P^0 to P^1 in the first period and then from P^1 to P^2 in the second period. Its total profit over both periods would be the shaded area in panel (a): *CBFG* in the first period plus *FIE* in the second period. Because of the approximation in

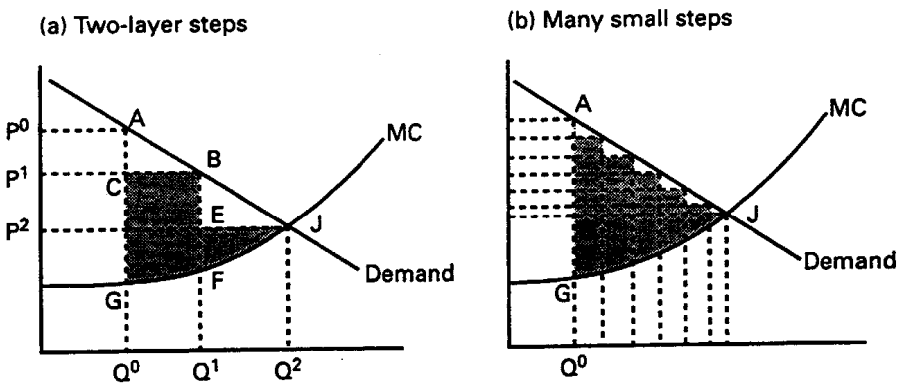

Figure 6.6
Total profit over time under F-V scheme

the F-V scheme, the firm would lose the areas *ABC* and *BJE*. However, suppose the firm took small price reductions over many periods. Its total profit over time in this case would be the shaded area in panel (b). As this graph suggests, the firm can obtain essentially all the surplus improvement, just as in the S-S scheme, if it takes sufficiently many small steps.

If the firm does not discount future profits, the F-V scheme is the same as the S-S scheme in that the firm obtains (essentially) all the improvement in surplus that it generates. The firm therefore maximizes the total improvement in surplus, moving eventually, as under the S-S scheme, to the first-best outcome. The only difference is that under F-V regulation the firm chooses to move very slowly to the first-best outcome, taking very small steps along the way.

In actuality, the firm discounts future relative to current profit. The firm therefore does not simply try to maximize the simple sum of profit over time; rather, it maximizes the sum of *discounted* profit over time, with future profit discounted more than current profit. As a result, the firm does not necessarily choose to take many small steps, because doing so means that much of its profit would be deferred far into the future. The firm does better by making larger price reductions early, incurring some loss due to the approximation, but gaining by receiving the profits earlier. The firm can therefore be expected to move toward the first-best outcome more quickly than would occur if the firm did not discount future profits.

The speed of the movement to the first-best prices is still not as great as under the S-S procedure: the firm will generally not move to the first-best outcome in one step. The reason is most easily discern-

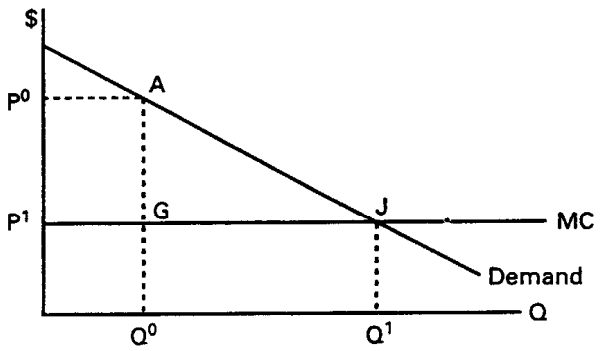

Figure 6.7
Constant marginal cost and one step to optimality

to decrease price to marginal cost in the first period after regulation is imposed, its profits in that period (and all future periods) would be zero. The move generates extra surplus in the amount of area *AJG*. However, the area *AJG* is also the amount that the approximation under the F-V scheme falls short of the true surplus gain: it is the part of the surplus gain that the firm does not receive because of the approximation. The firm therefore receives nothing in the first period. And because the firm stays at the first-best outcome once it reaches it, profit is zero in each subsequent period. Clearly, the firm would be better off, whatever its discount rate, by taking more than one period to reach the first-best outcome, because its profits would be greater than zero in each period along the way.

The conclusions can now be summarized. The F-V scheme can be implemented without the regulator having information on the demand curve of the firm. In equilibrium the firm chooses the first-best outcome, setting the price of each good at its marginal cost and producing at minimum cost. However, the firm may take many periods before it reaches the first-best outcome. This slowness is the "cost" the regulator incurs, relative to the S-S scheme, for not knowing the firm's demand curve.

The issue of speed can be fairly important. In an environment in which costs or demand shift over time, the outcome that constitutes first-best changes. For example, the price at which the demand curve intersects the marginal cost curve changes when either curve shifts. Under F-V regulation, the firm might move toward the first-best outcome more slowly than the first-best outcome itself moves, such that optimality is never achieved. In a changing environment, therefore, the speed of the S-S mechanism becomes more attractive.