10

Marginal- and Average-Cost Pricing

In a pure and simple static world of perfect competition, where production units purchase or rent all their inputs in competitive markets and each sells a single homogeneous product competitively, production takes place at a point of constant returns to scale where the marginal cost and average cost of the product are equal to each other and to its price. If in addition there are no neighborhood effects of externalities operating outside the market, the result will be Pareto-efficient, meaning that there is no feasible alternative arrangement that would be better for someone and no worse for anyone.

Difficulties with the Concept of Average Cost

As soon as production takes place with durable capital facilities that must be adapted to the needs of an individual firm there may no longer be an effective market for these facilities and a cost of their use during any particular period must be determined by other means. In the rather extreme case of the "one-horse-shay" asset that in a static environment yields a stream of identical services over a known lifetime, a constant periodic rental cost can be derived by the use of a "sinking fund" method of depreciation in which the rent is the sum of an increasing depreciation charge and a decreasing interest charge on the net value. But where the value of the service varies over time, whether because of physical deterioration, an increasing cost of maintenance needed to keep the item in "as new" condition, or shifts in demand, this would in principle cause depreciation charges to vary; in practice this is done in one of a number of arbitrary ways by

The New Palgrave, Eatwell *et al.*, Eds., Vol. III (Macmillan, 1987), pp. 311–18.

using "straight-line" or various forms of "accelerated" depreciation. If these charges are used as a basis for pricing, where competition is imperfect enough to give some leeway, the results can be correspondingly arbitrary.

More serious problems arise in the increasingly widespread cases of joint production of several distinguishable products or services. Where competitive markets exist, the market conditions dictate the allocation of joint costs among the various products, as when a meat-packing establishment produces steaks, hides, glue, and offals. There is no way in which one can determine a meaningful average cost of hides by considering only the production process. Where the products, though economically widely different, are physically similar, it is tempting to cut the Gordian knot and average over the entire output, often at the cost of serious impairment of economic efficiency. Even when elaborate rationales are concocted by cost accountants, unless demand conditions as well as production conditions are taken into account the results are essentially arbitrary.

One can do a little better with marginal cost, at least if one is seeking a short-run marginal social cost (hereafter SRMSC), which is the concept that would be relevant for efficiency-promoting pricing decisions. Unless a consumer is presented with a price that correctly represents the marginal social cost associated with the various alternatives open to him, he is likely to make inefficient decisions.

The Importance of Emphasizing the Short Run

One often finds in the literature proposals to use a "long-run marginal cost" as a basis for setting rates. The trouble is that in an operation producing a multitude of products with interrelated costs it is not possible even to define in any precise way what could be meant by a "long-run marginal cost," any more than one could define a relevant long-run marginal cost for the hides and steak that are derived from the same carcass in the face of fluctuations over time in relative demand.

The attempt to use a long-run concept seems to be motivated in part by the notion that in some sense the long-run concept is more inclusive in that it allows for variation in capital investment and would include a return on such investment, whereas short-run marginal costs would fail to cover the costs of capital investment. In the single-product steady-state case, however, which is the only case for which the long-run marginal cost can be clearly defined, if the investment in plant is at the optimal level, i.e., the level which will result in the given output being produced at the lowest total cost, short- and long-run average-cost curves will be tangent to each other at the given output, and short- and long-term marginal costs will be equal. Short-run marginal-cost prices will therefore cover just as much of the total cost as will prices based on "long-run marginal cost." If

short-run marginal cost is below the long-run marginal cost, this would indicate that the installed plant is larger than optimum, and conversely if plant is below optimum size, short-run marginal cost will be above long-run marginal cost.

Flexible Versus Stable Prices

A long-run approach is sometimes advocated on the ground that it results in more stable prices. Price rigidity, however, exacts a high toll in terms of reduced efficiency. It is sometimes argued that stable prices are required for intelligent planning for installations that commit the investor to the use of a given volume of service. There is nothing in a SRMSC pricing policy, however, that precludes providing the consumer with estimates of the probable course of prices in the longer term, or even entering into long-term contracts to purchase specified quantities of service. If they are not to interfere with efficiency, however, such contracts should allow for the possibility of purchasing additional amounts at the eventual going rates, or of selling back some of the contracted-for output if this should prove profitable for the consumer.

Lack of flexibility in pricing has, indeed, been a major source of inefficiency in the use of utility services, whether arising as a result of the cumbersomeness of the regulatory procedures in privately owned utilities, or of bureaucratic inertia in publicly owned ones. At times it has even appeared that it takes longer to carry out the bureaucratic procedures involved in altering a price than to install additional capacity, whereas in terms of the underlying capabilities prices can and should be altered on shorter notice than the time taken to adjust fixed capital installations.

Optimal Decision-making Sequence

The efficient pattern of decision making consists of first establishing a pricing policy to be followed in the future (as distinct from the application of that policy to produce a specific set of prices), then planning adjustments to fixed capital installations according to a cost–benefit analysis based on predicted demand patterns and predicted application of the pricing policy, subject to whatever financial constraints may be applicable, and then eventually determining prices on a day-to-day or month-to-month basis in terms of conditions as they actually develop.

Too often a rigid adherence to inappropriate financial constraints results in a pattern of pricing over time that leads to gross inefficiency in the utilization of facilities that are added in large increments. In the setting of tolls on bridges, for

example, a high fixed toll is often imposed from the start in an attempt to minimize early shortfalls of revenues below interest and amortization charges. When the indebtedness incurred to finance the facility is finally paid off, tolls are often eliminated, sometimes just at the time that they should be increased in order to check the growth of traffic and congestion and defer the necessity for the construction of additional facilities

The Forward-Looking Character of Marginal Cost

Since changes in present usage cannot affect costs incurred or irrevocably committed to in the past, it is only present and future costs that are of concern in the determination of marginal cost. Past recorded costs are relevant only as predictors of what current and future costs will turn out to be. The marginal cost of ten gallons of gasoline pumped into a car is not determined by what the service station paid for that gasoline, but by the cost expected to be incurred to replace that gasoline at the next delivery. The substantial time-lag that often exists between a change in price at the raw material level and its reflection at the retail level is one of the pervasive failings that contributes to the inefficiency of the economic system.

Another more important case in which future impacts are of vital importance in the calculation of marginal cost is where congestion accumulates a backlog of demand that has to be worked off over a period of time. A particularly striking case of this occurs when traffic regularly accumulates in a queue during rush hours at a bottleneck such as a toll bridge. The consequence of adding a car to the traffic stream is that there will be one more car waiting in the queue from the time the car joins the queue until the queue is eventually worked off, assuming that the flow through the bottleneck will be unaffected by the lengthening of the queue.

The marginal cost of a vehicle trip will be measured in terms of a number of vehicle hours of delay equal to the interval from the time the car would have arrived at the choke point if there had been no delay, to the time the queue is finally worked off. This is not measured by the length of the queue at the moment, but will be determined by the subsequent arrival of traffic over an extended period. A car arriving at the queue after it began to accumulate at 7:30 may get through the bottleneck at 8:00, after being delayed by only fifteen minutes, but if the bottleneck will not be worked off until 10:00 the marginal cost will be 2¼ vehicle hours of which only ¼ hour is borne by the added car itself. The remaining two hours, if evaluated at $5 per vehicle hour, would indicate that under these conditions the toll that would represent this externality would be $10. Marginal cost cannot be determined exclusively from conditions at the moment, but may

well depend, often to an important extent, on predictions as to what the impact of current consumption will be on conditions some distance into the future.

Marginal Cost of Heterogeneous Sets of Uses

It will often happen, for various reasons, that the same price will have to be applied to a nonhomogeneous set of uses. To set such a price properly, the marginal costs of the various uses within the set covered must be combined in some way to get a marginal cost relevant to this decision. It would be wrong, however, merely to average the marginal costs of all the uses for which this price is to be charged. Rather, the decision as to whether a decrease in a given price is desirable must consider the cost of the increments or decrements in the various outputs that will be bought as a result of the price change. In averaging the marginal costs of the various usage categories, the weighting will have to be in proportion to the responsiveness of each usage category to the change in price.

For example, if a price is to be set for electricity consumption on summer weekday afternoons, in a system where air-conditioning is an important load, consumption and marginal cost may be higher on hot days than on warm days, but it may be considered too difficult to differentiate in price between the two categories of days. An increase in the price for this entire set of periods may induce some customers to adjust the thermostat setting. But during hot days the equipment may work full tilt without reducing the temperature to the thermostat setting, whereas on warm days there will be a reduction in power consumption. The marginal cost relevant to the setting of the common price would then be determined predominantly by the lower marginal cost of the warm-day consumption, and relatively little, if at all, by the higher marginal cost hot-day consumption.

Anticipatory Marginal Cost

In many cases a customer will make his effective decision to consume an item some time in advance, and it will be the expected price as perceived by him at that time that determines his decision. If, as in services subject to reservation, a firm price must be quoted the time the reservation is made, it is the expected marginal cost as of that moment that should govern the price charged. In the case of a service where the demand is highly variable and to a considerable extent unpredictable, such an expected marginal cost would be an average of marginal costs that might arise under alternative possible developments, possibly ranging from a very low value, if there turns out to be unused capacity, to the possibly

quite high value if another latecomer must be turned away. The respective probabilities of these outcomes, as estimated at a given time, will vary with the proportion of the total supply already sold, the time remaining to the delivery of the service, and the pricing policy to be followed in the interim.

At one extreme, for long-haul airline reservations where the unit of sale is large, one might find it worth while to have a fairly elaborate pricing scheme in which the price quoted would vary according to the proportion of seats on a given flight already sold and the time remaining to departure, in simulation of what an ideal speculators' market might produce, the price at any time being an estimate of the price which, if maintained thereafter, would result in all the remaining seats being just sold out at departure time. This would correspond to marginal cost in that the sale of a seat at any given time would slightly raise the price during the remaining period to decrease demand by one unit, at a price that would be expected to be on the average equal to the price at which the seat was sold, indicating that the price was equal to the value of the seat to the alternative passenger.

Quality–Volume Interrelationships

In principle, in the absence of barriers to entry, competition would induce the supply of just sufficient seats on the various routes to cause revenues produced by such pricing to just cover costs. Even this, however, would be optimal only on those routes where traffic is so heavy that even with planes of a size producing the lowest cost per seat, further increases in service frequency would be of negligible value. On most routes there will remain economies of scale in that either providing more seats at the same frequency of service with larger planes would reduce costs per seat, or providing more seats with the same size of planes would provide an increased frequency of service that would be of value to others than the additional riders. In the latter case the marginal cost of providing for the additional passengers would be calculated by deducting the increase in the value of the service by reason of increased frequency from the cost of providing the added seats.

If it were possible to adjust plane size and frequency in a continuous fashion, then if the situation is optimal the two marginal costs would be equal. In practice both plane size and service frequency can be varied only in discrete jumps, so that this relation would be only approximate. Optimal price would be above a downward marginal cost calculated on the basis of a reduction in service, and below an upward marginal cost calculated on the basis of an increase in service. The decreases and increases might involve a combination of frequency and plane size changes. To preserve the formulation that price should equal marginal cost it

may be useful to define marginal cost in such cases as consisting of the range between these upward and downward values rather than as a single point.

In practice, between the existence of economies of scale and the imperfect cross-elasticity of demand between flights at various times and with different amenities, removal of regulation tends to result in an emphasis on nonprice competition, attempts to subdivide the market by various devices and restrictions to permit discriminatory pricing, and a bunching of service schedules at salient times and places that provides a lower overall level of convenience than would be possible were the given number of seat miles distributed more efficiently.

Where the unit of sale is small it may not be worth while to incur the transaction costs of varying price in strict conformity with SRMSC. One could, in theory, apply the same principle to the sale of newspapers at a given outlet. The price of a newspaper would vary according to the number of unsold papers remaining and the time of day. This would result in less disappointment of customers having an urgent desire for a paper late in the day and encountering a sold-out condition, and fewer unsold papers returned. But unless some ingenious device can be found for executing such a program at low transaction costs, it probably would not be considered worth while, even by the most sanguine advocate of marginal cost pricing.

Wear and Tear, Depreciation, and Marginal Cost

Even in the absence of lumpiness or technological change, existing methods of charging for capital use often fail to give a proper evaluation of marginal cost. This is especially true where the useful life of a unit of equipment is determined more by amount of use than by lapse of time. In the extreme case of equipment that must be retired at the end of a given number of miles or hours of active service, or after the production of so many kwh of energy, and which, in one-horse-shay fashion, gives a uniform quality of service over its lifetime without requiring increasing levels of maintenance, the marginal cost of use at a given time will be the consequent advancing of the time of retirement of the equipment. The marginal cost of using the newest units will be the lowest, and will advance over time at a rate equal to the rate of interest as the equipment ages and the advancement of replacement consequent upon use becomes less and less remote.

In a service subject to daily and weekly peaks, the newest equipment will be allocated to the heaviest service, operating during both peak and off-peak hours. Equipment will be relegated to less and less intense service as it ages. The marginal cost of service at a particular moment will be that for the oldest unit that has to be pressed into service at that instant. The rental charge for the use of the unit will vary gradually over the entire range of demands, rather than dropping off

to zero whenever the full complement of equipment is not required. At the other end, in this extreme case, the service provided would not necessarily be held constant by price variation over an extended peak period: under the conditions postulated it would be possible to provide for needle peaks by planning for the stretching out over time of the final service units of the oldest equipment. In this way the required peak capacity can be provided at a cost much lower than that which would be calculated by loading all the capital charges for the added equipment on this brief period of use.

Another way of looking at the matter is to appeal to the proposition that perfect competition under conditions of perfect foresight will produce optimal results. To this end one can suppose a situation in which vehicles are rented by the hour from a large number of lessors operating in a competitive market. For simplicity, initially, one can assume all vehicles to be of the one-horse-shay variety, being equivalent to bundles of hours of active service, with the quality of service being independent of age up to a final "bubble-burst" collapse. Also, for simplicity, assume a steady state in which vehicles are scrapped and replaced at a constant rate over time, so that at any given moment vehicles are evenly distributed by age.

A common market rental price for all vehicles at any given time of the week will emerge, being higher as the number of vehicles in service at the time is greater. During any given week, each renter will have a reservation price for his vehicle, such that he will rent his vehicle during those hours for which the market rental is above this reservation price and never when the market rate is lower. This reservation price will increase over time for any given vehicle at the market rate of interest, since a renter will rent his vehicle if and only if the net present value of the rental discounted back to the time of purchase exceeds some fixed amount. The owner would not want to rent his vehicle for a net present value less than he could have got by selling one of this stock of service units at some other time at or just below his reservation price. New buses will have the lowest reservation price and will be assigned to the schedules calling for the most hours of service per week, while old buses will be held idle during slack hours and used only for peak service. As each bus ages it will be assigned to less and less heavy service along the load-duration curve.

This pattern of usage can be regarded as resulting from a desire to recover the capital tied up in the usage units of each bus as rapidly as possible. It is related to the practice in electric utilities of using the newest units for peak service, in that case motivated in part by the tendency for the newer units to be more efficient in thermal terms. To be sure, occasionally new units are designed specifically for peaking service, with a correspondingly low capital cost, though this is a relatively recent phenomenon related to a slowing-down of secular increases in potential thermal efficiency.

In any case, where wear-and-tear is a factor, one cannot properly allocate

depreciation charges primarily to peak service, however defined, nor should they be spread evenly over all service, much less spread evenly over hours of the week so that vehicle hours in off-peak periods would get higher charges than during the peak. Rather the depreciation charge per vehicle hour will vary gradually and in a positive direction with the intensity of use of the equipment at any given hour.

The analysis becomes a little complicated when equipment life is dependent on mileage or loading or intensity of use as well as hours of active service, so that different rentals would properly be chargeable according to the nature of the service for which the unit is being rented. Also further analysis is required if equipment is laid up between runs at isolated terminals rather than at a central depot where a market could be postulated, or if the fleet contains vehicles varying in size or other characteristics. It would even be theoretically appropriate to charge different fares for the same trip at the same time if made on vehicles with different origins or destinations. (In Hong Kong, indeed, the practice is to charge a flat fare on each route, but to differentiate the fare fairly elaborately as among routes. On segments where routes converge, this has the unfortunate result of unduly concentrating riding on buses with the lower fares, even where the higher fare buses have empty seats and are making stops in any case for other passengers.)

Costs of major overhauls that are performed at relatively long intervals would also complicate the picture. There are also problems associated with gradual or sudden changes in overall demand levels, or special events that can be anticipated sufficiently to present an opportunity for reacting in terms of a change in price. The picture can be further complicated if, as was discussed above, there are changes in available technologies or other changes in quality or cost. But the same method of analysis in terms of a hypothetical competitive market can be used to obtain appropriate results.

For the sake of simplicity the above analysis has been couched mainly in terms of a bus service, but the analysis is applicable wherever the useful life of equipment is in part a function of the intensity with which it is used.

Responsive Pricing

In some cases, notably in telephone and electric power services, the technical possibility exists for conveying information as to the current price to customers at the instant of consumption, and for customers to respond to such information in a worthwhile manner at modest cost. In the case of telephone service the information as to the level of charges for local calls can be substituted for the dial tone, with information on rates for long distance calls provided to users who wait for it before dialing the final digits. If the charge exceeds what the customer is willing to pay the call can be aborted with little occupancy of equipment or inconvenience to

the user. Prices can be varied from moment to moment in accordance with marginal cost, as estimated from the degree of busyness of the relevant sets of equipment.

In the case of electric power, the costs of providing for a variation of the price according to the conditions of the moment would be somewhat greater. But if the facilities take the form of remote meter reading, either by carrier current over the power lines or by a separate communications channel, much of the cost would be covered by the avoidance of costs involved in manual meter reading. A signal of rate changes can then be provided to the customer as a byproduct of the signal required to initiate a new rate period. The customer can then respond either manually or by installing automatic equipment which will adjust the operation of such items as air-conditioning and refrigeration compressors, water heaters and the like, according to the level of rates in a manner determined by the customer himself. Retrofitting of existing meters by attaching a pulse-generating device such as a mirror and photo-electric cell to the rotor shaft of the existing meter and feeding the pulses to electronic counters and registers should be possible at relatively low cost.

Such responsive pricing would be especially valuable in dealing with emergencies, providing greater assurance of the maintenance of essential services than is possible with existing techniques, and making it possible to reduce substantially the cost of providing reserve capacity. In the case of floods, conflagrations, breakdowns in transit, or other emergencies that under present conditions tend to result in the overloading of telephone facilities and difficulties in completing calls of a vital nature, rates can be charged that are high enough to inhibit a sufficient number of less important calls so that the ability of the system to handle vital calls promptly is preserved. This is difficult to do with present techniques, for while it is relatively easy to give priority to calls originating at such points as police stations, hospitals, and the like, most emergency calls are calls to rather than for these points and it is much more difficult to distinguish such calls close to the point of origin. And there are always a certain number of vital calls not distinguishable in terms of either origin or destination.

Again, in the case of unscheduled power cuts, it would be possible to cause an almost instantaneous shedding of substantial water-heating and refrigeration loads, followed, in the case of an extended cut, by partial shedding of elevator, transit and batch process loads for which it is more inconvenient to respond quite so promptly, after which a sufficient refrigeration load can be picked up as needed to avoid food spoilage. Many of the serious consequences of major power blackouts could have been avoided had such a system been in place at the time. Reserve capacity might well be cut back to provision for scheduled maintenance, leaving the load-shedding capability of responsive pricing to function as a reserve. In many cases the speed of response possible with responsive pricing would be

faster than the reaction time within which reserve capacity can pick up load, leading to better voltage regulation and a higher quality of service to customers remaining on the line. And if, in spite of everything, areas must be cut off completely, responsive pricing would also be of considerable help in facilitating a smooth recovery from an outage: instead of having a whole army of motors trying to start up at once upon the restoration of power, with consequent load surges, voltage fluctuations, and malfunction of equipment, load could be picked up smoothly and gradually as the price is lowered from the inhibiting level.

Preserving Incentives with Escrow Funds

With privately owned utilities the regulatory process is too slow to permit prices established directly by regulation to be constantly adjusted to changing current conditions, unless indeed the regulators were to assume a large part of what are normally the responsibilities of management. The problem thus arises of how to allow the prices to be paid by customers to be varied by the utility management without giving rise to incentives for behavior contrary to the public interest. Even if a formula could be devised that would require the utility to adjust prices to track short-run marginal cost, if the utility were allowed to keep the revenues thus generated without restriction, this would set up undesirable incentives for the utility to skimp on the provision of capacity in order to drive up the marginal cost, price, revenues, and profits.

A resolution of this dilemma can be achieved by separating the revenue to be retained by the utility from the amounts to be paid by customers. We can have the "responsive" prices paid by customers vary according to short-run marginal social cost, while the revenues to be retained by the utility are determined by a "standard" price schedule fixed by regulation in the normal manner, the difference being paid into or out of an escrow fund. Failure of the utility to expand capacity adequately would drive marginal cost up, and with it the responsive price, causing revenues to flow into the escrow fund, but the only way the utility could draw on these funds would be to expand capacity sufficiently to drive marginal cost down, causing the responsive rate to fall below the standard rate on the average, entitling the utility to make up the difference from the escrow fund as long as it lasts. Excessive expansion would result in the escrow fund being exhausted, with a corresponding constraint on the revenues obtainable by the utility from the unaugmented low responsive rates.

The setting of the responsive rates would have to be to a large extent at the discretion of the operating utility, though the regulatory commission could monitor the process and even attempt to establish guidelines according to which the responsive price should be set. The utility would normally have no incentive to

set the responsive rate below marginal cost, since this would merely increase sales and hence costs by more than any possible long-run increase in revenues to the utility. To be sure, in the short run it might be able to draw on the escrow fund to the extent of the excess, if any, of the standard rate over the responsive rate, but since from a long-run perspective there will normally be other more advantageous ways of drawing on this fund this will not be attractive.

When marginal cost is below the standard price, which would tend to be the usual situation, the utility would in general have an incentive to set the price between the marginal cost and the standard price, since each additional sale produced by the lower price will yield an immediate net revenue equal to the difference between marginal cost and the standard price, offset only by the drawing down of the escrow fund by the difference between the responsive and the standard price. When marginal cost is above the standard price, which with a properly designed standard rate schedule with time-of-day variation should happen relatively rarely, the utility would have an incentive to set the price at least at the marginal-cost level, since to set it lower would tend to increase output at a cost in excess of anything the utility could ever recover. How much higher than marginal cost the price might be set would in theory be limited by the condition that the price could not be high enough to curtail demand sufficiently to drive marginal cost below the standard price. If the standard price has an adequate time-of-day variation, this constraint, loose as it may seem, may be sufficient. Additional guidelines could of course be imposed by the regulatory commission for those rare occasions where this constraint might seem insufficient to keep prices within bounds.

Actual Steps Toward Responsive Pricing

Some actual practices of utility companies are steps in the direction of responsive pricing. Contracts for "interruptible" power provide for load shedding at the discretion of the utility subject to some overall limits. As these are fairly long-term contracts that usually require *ad hoc* communication between the utility and the customer, their applicability is limited and there is no assurance that the necessary shedding will be done in the most economical manner. Many customers are reluctant to submit to load shedding that is not under their control at least to some extent, and that might be imposed under awkward circumstances. Where reserves are ample and interruption is highly unlikely, such contracts have been challenged as being a form of concealed discriminatory concession. On the other hand customers entering into such contracts in the expectation of not being interrupted may feel aggrieved if interruption actually takes place.

Another experimental provision applied by a company with a heavy summer

air-conditioning load is for a special surcharge to be applied to the usage of larger customers on days when the temperature at some standard location exceeds a critical level. And another company bases its demand charge on the individual customer's demand recorded at the time that turns out to have been the monthly system peak load, supplying the customers with information as to moment-to-moment variations in the system load. This leads to interesting game-playing on the part of customers as they attempt to keep their own consumption down at times that look as though they might become the monthly peak, with the result that this action may itself shift the peak to another time.

Economies of Scale, Subsidy and Second-Best Pricing

Where there are economies of scale, prices set at marginal cost will fail to cover total costs, thus requiring a subsidy. One reason for wanting to avoid such a subsidy is that if an agency is considered eligible for a subsidy much of the pressure on management to operate efficiently will be lost and management effort will be diverted from controlling costs to pleading for an enhancement of the subsidy. This effect can be minimized by establishing the base for the subsidy in a manner as little susceptible as possible to untoward pressure from management. But it is unlikely that this can be as effective in preserving incentives for cost containment as a requirement that the operation be financially self-sustaining. To achieve this, prices must be raised above marginal cost, and in a multi-product operation the question arises as to how these margins should vary from one price to another within the agency.

Another objection to subsidy is that it raises hard questions of who should bear the burden of the subsidy. More fundamentally the taxes imposed to provide the subsidy will often have distorting effects of their own, and minimizing the overall distortion would again require prices to be raised above marginal cost. One can, indeed, regard these excesses of price over marginal cost as excise taxes comparable to other excise taxes that might be levied to raise a specified amount of revenue.

The answer given to the problem of how to allocate excise taxes and other margins of price above marginal cost so as to minimize the overall loss of economic efficiency given by Frank Ramsey in 1927 can be expressed for the case of independent demands as the inverse elasticity rule, which says that the margin of price over marginal cost as a percentage of the price shall be inversely proportional to the elasticity of demand. A more general formulation is one that states that prices shall be such that consumption of the various services would be decreased by a uniform percentage from that which would have been consumed if

price had been set at marginal cost and demand had been a linear extrapolation from the neighborhood of the "second-best" point.

A more transparent formulation, devised by Bernard Sobin in work for the US Postal Service, is the requirement of a uniform "leakage ratio," leakage being the difference between the net revenue actually derivable from a small increment in a particular price and the hypothetical revenue that would have been obtained had there been no change in consumption as a result of this increment. Leakage is the algebraic sum of the products of the changes in consumption of the various related products induced by the small change in a given price, and the respective margins between their prices and marginal costs. Leakage is a measure of the loss of efficiency resulting from the change in the particular price, and the leakage ratio is the ratio of this loss of efficiency to the hypothetical gain in gross revenue if there had been no change in consumption. If one leakage ratio should be greater than another, the same net revenue could be obtained at greater economic efficiency by getting more revenue from the price with the smaller leakage ratio and less from the other. The second-best solution accordingly requires that all leakage ratios be equal.

This analysis can be extended to the case where the agency is being subsidized by taxes which involve an adverse impact on the economy, in terms of marginal distorting effects, compliance costs, and collection costs, which can be expressed as the "marginal cost of public funds" (MCPF). For a net decrease in the subsidy derived by increasing a price, which can be considered to be equivalent to imposing a tax equal to the difference between the marginal cost and the price, $MCPF = LR/(1 - LR)$, where LR is the leakage ratio. A second-best optimum is then one where the MCPF's are equalized over both external and internal taxes.

Special Sources of Subsidy: Land Rents and Congestion Charges

In the case of goods and services with economies of scale that are provided primarily to consumers within a particular urbanized area, methods of financing may be available that involve no marginal cost of public funds or even result in an enhancement of efficiency. The existence of large cities, indeed, is to a predominant extent due to the availability in the city of goods and services produced under conditions of economies of scale: if there were no economies of scale, activity could be scattered about the landscape in hamlets, with great reduction in the high transportation costs involved in movement about a large city. If prices of these services are reduced to marginal cost, the increased attractiveness of the city as a consequence would tend to drive up land rents within the city, and it appears quite appropriate that a levy on such rents should be used to finance the required subsidies. And while there are practical and conceptual difficulties in defining

exactly how land rents or land values should be specified for purposes of levying a tax, it is generally considered that a tax on land values, properly defined, has negligible adverse impacts on the efficient allocation of resources.

Indeed, there is a theorem of spatial economics which states that in a system of perfect competition among cities, the availability in the city of services and products subject to economies of scale, priced at their respective marginal social costs, will generate land rents just sufficient to supply the subsidies required to permit prices to be lowered to marginal cost. Among the more important of these services are utility services such as electric power, telephone, cable communications, water supply, mail collection and delivery, sewers and waste disposal, and local transit. It is not clear just how broad the conditions are under which this theorem would hold, and there are difficulties in capturing all land rents for subsidy purposes, but steps in this direction are clearly desirable.

On a more intuitive level, one can note that a person who occupies or uses land that is provided with services such as the availability of transit, electricity, telephone, mail delivery, and the like will be requiring that these services be carried past his property to serve others whether or not he himself uses them. The user of tennis courts located conveniently in a built-up area should no more be excused from contributing to the costs of carrying these services past the courts, even though no direct use is made of electric power, telephone, mail, or other services, than he should expect his auto dealer to cut the price of an automobile by the cost of the headlights and windshield wipers merely because he asserts that he will never drive at night or in bad weather. Tennis players will indeed pay a rent enhanced by the presence of these services and the consequent greater demand for the land for other purposes, but the rent will go to the landlord, not to the purveyor of the services, and the price of the services to those who use them will be too high for efficiency, unless indeed they are subsidized by other taxes that have their own distorting effects.

It is a corollary of this theorem that it would be to the advantage of the landlords in the area, *faute de mieux*, to agree collectively to pay a tax based on their land values, in order to subsidize the various utility services to enable the prices to be set closer to marginal social cost. They could expect in the long run that this action would increase their rents by as much or more than the taxes. To be sure, they might do better by getting someone else to pull their chestnuts out of the fire, but they can do this only at considerable damage to the overall efficiency of the economy of the city, to say nothing of the inequity of such a parasitic relationship.

In addition to land rents in the conventional sense, there is the land used for city streets for the use of which no adequate rental is generally charged. Charging on the basis of SRMSC for the use of congested city streets would in most cases yield a revenue far in excess of the cost of maintaining such facilities, which could

appropriately be used for the subsidy of other urban facilities. Properly adjusted, such charges would increase efficiency by bringing home to the users the costs that their use directly imposes on others.

Formerly it would have been considered impractical to attempt to charge for the use of city streets according to the amount of congestion caused: the collection of tolls by manual methods at a multitude of points within the city might well create more congestion than it averted. Advances in technology have, however, made it possible to do this at minimal interference with traffic flow and at modest cost. One method, proposed as long ago as 1959 and recently carried to the point, is to require all vehicles using the congested facilities to be equipped with electronic response units which will permit individual vehicles to be identified as they pass scanning stations suitably distributed within and around the congested areas so that the records thus generated can be processed by computer and appropriate bills sent to the registered owners at convenient intervals. If properly done, this would greatly improve traffic conditions so that the net cost of the revenue to the road users would be far less than the amount collected as revenue. A pilot installation has recently been tested in Hong Kong with satisfactory results, but full implementation appears to have been deferred, because of the political situation associated with the impending transfer of sovereignty.

Indeed, one can define "hypercongestion" as a condition where so many cars are attempting to move in a given area that fewer vehicle miles of travel are being accomplished than could be if fewer vehicles were in the area but could move more rapidly; for example if 1,000 vehicles in an area move at 8 mph and produce 8,000 vehicle miles of travel per hour, reducing the number of cars in the area at a given time to 800 might raise speed to 11 mph producing 8,800 vehicle miles of travel per hour. By restricting the flow of traffic in the period leading up to the hypercongestion period, road pricing could prevent hypercongestion from occurring, except possibly sporadically, and in any case so improve conditions that more movement would be accomplished during the peak period at faster speeds. The improvement during peak periods might even be such that total movement throughout the day would be increased, and where conditions are now severe users could find that they are better off than before, even inclusive of the payment of the congestion charge.

If there are bridges, tunnels, or other special facilities for which a toll is already being charged, and which regularly back up a queue during the morning rush hour, substantial revenues can be obtained at no overall net cost to the users by adding a surcharge to the toll during the period where queueing regularly threatens, rising gradually from zero to a maximum and down again in such a way that by gradual adjustment regular queueing is substantially eliminated. The toll surcharge will then be taking the place of the queue in influencing decisions as to when to travel, and in general those who plan their trips in terms of time of arrival

at their destination will be able to leave as many minutes later as they formerly wasted in the queue, pass the bottleneck at the same time as before, and arrive at their destinations at the same time as before. The extra toll will be roughly the equivalent of the value of the extra time enjoyed at the origin point, and the revenue will in effect be obtained at no net burden on the users. In practice the results may be even better than this as a result of the added encouragement to car-pooling, the reduction of obstruction to cross-traffic, and the expediting of emergency or other trips where the delay had been a particularly serious matter.

Gains in the evening may be not quite so dramatic. The situation is not symmetrical, as typically the timing of the trip will be determined in terms of time of departure, which is separated by the queue from the time at the bottleneck. On the other hand the risk of conditions approaching gridlock is greater, since the accumulation of queues inside circumferential bottlenecks is more likely to create congestion, and there is less of a physical barrier to the simultaneous emergence of large quantities of traffic from parking lots into the downtown streets than there is in the morning to the convergence on the congested area of traffic arriving from the outside.

Congestion charges should be imposed, at least notionally, without exception on all forms of traffic. Such charges would be a necessary element in the cost–benefit analysis by which decisions are made as to the level and pattern of bus service to be provided, even though they would not be directly relevant to the determination of the price structure to be applied to that service.

Paradoxes in the Behavior of Marginal Social Cost

A strict calculation of marginal social cost in particular circumstances may produce what may appear to be quite paradoxical results. For example, in many circumstances it will be optimal, and even essential, to maintain at least a minimum frequency of service in off-peak hours with buses of a standard size, resulting in there being practically always a large number of empty seats in each bus. Under these circumstances the cost of carrying additional passengers is predominantly the cost of boarding and alighting, including the time of the driver and the other passengers on the bus who are delayed in the process. This cost will be relatively higher if the bus is half full than if it is nearly empty. The result is likely to be that the cost of a trip from a point near one end of the run to a point near the other end, at both of which points the bus is likely to be lightly loaded, may be smaller than for a shorter trip between points near the middle of the run where the bus is likely to be more heavily loaded. This is not a trivial matter: if it were there would be no sense to the refusal of express buses with empty seats to pick up local passengers. It is highly unlikely, however, that fares based on such a

seemingly perverse behavior of cost would meet with popular approval. Indeed, the original US interstate commerce legislation contained prohibitions against higher rates being charged shorter hauls than for any longer hauls within which they might be included.

Another paradoxical example can occur in mixed hydro-thermal electric power systems: an increase in fuel prices could result in the marginal cost of power at particular times being reduced rather than increased. If hydro-dams are spilling water at certain seasons of the year, increased fuel costs may make it economical to increase the installed generating capacity to make use of the spilling water, even for a briefer period of time over the year than was previously worth while. If during the wet season installed hydro-generating capacity is more than sufficient to meet trough demand, marginal cost during such periods will be substantially zero, or at most limited to a small element of wear and tear on equipment pressed into service. Installing more turbo-generators would expand the period during which this low marginal cost is effective, so that while increased fuel costs cause marginal cost to rise during the peak, the result could also be to lower marginal cost in these intervals into which the period of exclusive hydro-supply expands.

In the case of long-distance telephone service, the drastic reductions in the cost of bulk line-haul transmission have created a situation where distance, especially beyond the range where separate wire transmission is economical, is relatively unimportant as a cost factor, and where satellite transmission is involved, ground distance is indeed irrelevant. What remains important is the number of successive circuits, with their associated termination and switching equipment, involved in the making of a call. Thus a call between two small communities over a moderate distance, for which the volume of calling is insufficient to warrant the provision of a separate circuit, will generally cost substantially more than a call between important centres over a much longer distance, since the latter will involve only a single long-haul circuit, while the former will require patching through two or more long-haul circuits.

Another anomaly occurs when an innovation promising substantial reductions in costs appears on the horizon, such as has happened repeatedly in telecommunications. Any further installation of the old technology in the interim before the new technology is actually available will involve an investment which will have its capital value diminished over a brief period to that determined by its competition with the new technology. High depreciation or obsolescence charges are in order, and the prospect of the new lower costs results in higher current prices which would serve to hold back current demand and lessen the amount of old technology required to be installed.

Marginal-cost pricing is thus not a matter of merely lowering the general level of prices with the aid of a subsidy; with or without subsidy it calls for drastic

restructuring of pricing practices, with opportunities for very substantial improvements in efficiency at critical points.

References

Beckwith, B.P., *Marginal Cost Price-Output Control* (New York: Columbia University Press, 1955).

Mitchell, M., Manning, G., and Acton, J.P., *Peak-Load Pricing* (Cambridge, Mass.: Ballinger, 1978).

Nelson, J.R. (Ed.) *Marginal Cost Pricing in Practice* (Englewood Cliffs: Prentice-Hall, 1964).

Ramsey, F., "A Contribution to the Theory of Taxation," *Economic Journal*, 37 (March, 1927), pp. 47–61.

Vickrey, W., "Optimization of Traffic and Facilities," *Journal of Transport Economics and Policy*, 1(2) (May, 1967), pp. 1–14.

"Congestion Theory and Transport Investment," *American Economic Review*, 59(2) (May, 1969), pp. 251–60.

"The City as a Firm," in *The Economics of Public Services*, M.S. Feldstein and R.F. Inman, Eds., Proceedings of a conference held by the International Economic Association, Turin, Italy (London: Macmillan; New York: Wiley) pp. 334–43.

"Responsive Pricing of Public Utility Services," *Bell Journal of Economics and Management Science* 2(1) (Spring, 1971), pp. 337–46.