Attachment 15

Aspects of Load Management for Electric Utilities

John Tschirhart
Department of Economics
University of Wyoming

June 1980

This research was supported in part by the Public Utility Research Center at the University of Florida. Portions of this paper will appear in the Scottish Journal of Political Economy in October, 1980. I am grateful to Sanford Berg for his suggestions.

As a result of the 1978 Public Utility Regulatory Policies Act (PURPA), state public service commissions in the U.S. are required to consider innovative methods for pricing and supplying electricity. The purpose is to encourage more efficient use of energy. Peak-load or time of use (TOU) pricing is one subject receiving attention, and the theoretical economic literature over the past twenty years on this subject will be finally applied in the U.S. If ongoing studies in many of the 50 states show that the benefits of TOU pricing exceed the costs, then this pricing will be implemented. However, preliminary work has indicated that TOU pricing will probably not be cost effective at the residential level due to metering costs.

Where TOU pricing is not useful, are there other practices that could be implemented instead? One possibility is load management. Essentially, the utility has direct control over the load curve by being able to ration the available supply. This can be accomplished with radio control, ripple control, telephone lines, etc. When demand for power exceeds the available supply, some load is shed to heaters, air conditioners, irrigation pumps, or other nonvital uses.

The question then is: Given that TOU pricing is not economically justified, should load management be practiced? And if so what are the major issues that arise? Clearly, the need for load management comes about because of the fluctuations in demand that rise due to the random nature of demand. As a point of departure, therefore, a review and analysis of the recent economic literature on public utility pricing under stochastic demand would seem appropriate.

The problem of pricing for public utilities has been thoroughly explored under conditions of certainty.[1] Brown and Johnson (1969) (hereinafter referred to as BJ) extended the problem to situations where demand is subject to a random disturbance. They assumed that a producer is directed to maximize expected social welfare and must announce a price and capacity output before actual demand is known. The results of the maximization problem establish that price is equal to short-run marginal cost and that optimal capacity is greater than riskless optimal capacity. This pricing policy prompted Turvey (1970, p. 485) to comment about the problems associated with excess demand and rationing at times of high demand. "There is thus a tradeoff between the sacrifice of consumers' surplus on the one hand and the stringency of rationing on the other hand." Meyer (1975) and Crew and Kleindorfer (1976) extended the BJ model by allowing for this tradeoff. In the former, a chance constraint on reliability is used, while in the latter, rationing costs are used.

There is also the problem of how the producer can set prices to avoid deficits. In the imaginary world of certainty, a budget constraint will ensure adequate revenues.[2] However, conventional budget constraints are meaningless if demand is stochastic. Sherman and Visscher (1978) and Tschirhart (1975) address this problem by using different forms of a stochastic budget constraint showing that a new set of tradeoffs arise.

In all of this work, there is an implicit assumption that demand is independent of the reliability of service. Indeed, when this assumption is relaxed, the controversial BJ solution can no longer be obtained. The analysis presented here shows that for a given output capacity, an increase in price may improve welfare via an improvement in reliability.

Therefore, relaxing the independency assumption extends earlier work by capturing the idea that welfare may explicitly depend on reliability. Reliability is an important part of the load-management problem.

I. Stochastic Demand

Assume that there are n demand periods of equal duration.[3] The producer's problem is to choose a capacity, Z, and a price vector, $p = (p_1,...,p_n)$ where p_i is the price in period i that will maximize expected social welfare. Throughout, subscript i will run over all periods from 1 to n. Demand in each period is subject to random fluctuations so that a unique relationship does not exist between p_i and the quantity demanded given by q_i. The producer cannot choose a particular price and be certain as to what the demand response will be. Instead of the riskless relationship $q_i = q_i(p_i)$, the producer is confronted with $q_i = q_i(p_i, u_i)$, where u_i is a random variable. The producer is cognizant of the riskless demand curves, but knows only the density functions of the random variables given by $f_i(u_i)$.

Leland (1972) suggests that the electric power industry is a good example of a price setter, that is, a firm which sets prices before u is known and then adjusts output to meet demand. However, when demand exceeds capacity, no further output can be produced. The utility cannot very well plan on adjusting capacity to meet demand since capacity adjustments require too much time. Leland, then, was concerned with the short run. In the long run, a public utility is a price-quantity setter; prices and capacity level are chosen before the u_i's are known. This is essentially the situation analyzed in BJ.

BJ examined two specific forms of the demand function, namely the additive and multiplicative forms. Similar results are obtained for both forms. The following analysis will be confined to the latter. Thus,

$$q_i(p_i, u_i) = X_i(p_i)u_i$$

where X_i is the mean demand function for period i. The distribution functions for the u_i are given by $F_i(\gamma) = \int_0^{\gamma} f_i(u_i)\, du_i$ and the expected value of u_i by $\int_0^{\infty} u_i f_i(u_i) du_i = 1$.

The expected value of welfare is

$$E[W] = E[\text{Consumers' Surplus (CS-L)} + \text{Total Revenue (TR)} - \text{Total Cost (TC)}]. \tag{1}$$

The term CS in (1) is the usual consumers' surplus triangle under the demand curve. But when demand exceeds capacity, the entire triangle is not attained. Any portion of the triangle beyond the capcity level is lost, since demand in this region is not satisfied. This lost portion, denoted L, must be subtracted from CS. Figure 1 depicts L for a demand curve where $u_i = \bar{u}_i$. The value of L will be positive when there is excess demand and zero otherwise. Equations (2) and (3) represent the summations over all periods of E[CS] and E[L] respectively.

$$\sum_{i=1}^{n} [\int_0^{\infty} f_i(u_i) \int_{p_i}^{X_i^{-1}(0)} X_i(p_i)u_i dp_i du_i \tag{2}$$

$$\sum_{i=1}^{n} [\int_{Z/X_i(p_i)}^{\infty} f_i(u_i) \int_{p_i}^{X_i^{-1}(Z/u_i)} [X_i(p_i)u_i - Z]dp_i du_i] \tag{3}$$

The upper limit on the second integral in (2) is the intercept of the demand curve with the price axis, and the lower limit is the prevailing price. The lower limit of the first integral in (3) represents the value of u_i such that demand exactly equals capacity (i.e., $X_i(p_i)u_i = Z$ or $u_i = Z/X_i(p_i)$). The upper limit in the second integral of (3) is the price that would clear the market when there is excess demand. These limits are also illustrated in Figure 1.

Implicit in this formulation of L is that consumers are costlessly ranked according to their willingness to pay. When demand exceeds capacity, those consumers with the highest willingness are served first.[4] In reality, ranking consumers is not a costless operation. This suggests that the cost of ranking should be included in the objective function, which is done in the next section.

Total revenue includes revenue when there is excess supply plus revenue when there is excess demand. E[TR] is given by (4).

$$\sum_{i=1}^{n} \left[\int_0^{Z/X_i(p_i)} f_i(u_i)X_i(p_i)u_i p_i du_i + \int_{Z/X_i(p_i)}^{\infty} f_i(u_i)Z\, p_i du_i\right] \tag{4}$$

Total cost is the sum of operating costs and capacity costs. Both are assumed to be constant per unit of output and given by b and β respectively. Equation (5) is the E[TC] which includes operating costs when there is both excess supply and excess demand.

$$\sum_{i=1}^{n} \left[\int_0^{Z/X_i(p_i)} f_i(u_i)X_i(p_i)u_i\, bdu_i + \int_{Z/X_i(p_i)}^{\infty} f_i(u_i)\, Zbdu_i\right] + \beta Z \tag{5}$$

As demonstrated by BJ, when E[W] is maximized over all p_i, the first order necessary conditions yield the prices

$$p_i = b.^{5} \tag{6}$$

There are two fundamental problems with these prices: first, they are likely to result in significant excess demand in periods when u_i is large; and second, there is no possibility that revenue will cover costs. Both of these problems are of great concern to public utilities, and they are discussed in detail below.

The reason that price equals operating cost only in the BJ model can be seen in Figure 2 where four demand curves are drawn for four values of u_i. These values are .5, .71, 1, and 2. Since capacity is chosen <u>ex ante</u> total capacity cost, βZ, is essentially a fixed cost when demand becomes known. For demand curve .5X in Figure 2, $p = b$ is clearly the welfare maximizing price since operating cost is covered and consumers' surplus is at a maximum. This price also applies for any demand up to curve .71X for which demand at price p equals capacity.[6] For demand curve 2X, any price between zero and p''' will yield the same welfare, although the distribution of welfare between revenue and consumers' surplus will vary. Prices near zero for demand curve 2X will require more rationing, but no trouble, rationing is costless in the model. This reasoning also applies for any demand beyond curve .71X. Thus, p is the unique optimum price for all demands up to curve .71X, and p is the optimum but not unique price for all demands beyond .71X. Therefore, p is selected. This argument holds for all positive Z, and crucially relies on the assumptions that those who value the service least are the first to be cut off when rationing is necessary, and that rationing is costless.

II. <u>Accounting for Excess Demand</u>

In the stochastic setting, the following relationships hold:

$$q_i(p_i,u_i) \leq Z \rightarrow q_i(p_i,u_i) = \text{output}$$

$$q_i(p_i,u_i) > Z \rightarrow Z = \text{output}.$$

The latter case gives rise to excess demand. This problem can be avoided under riskless demand by simply adding the constraint $q_i \leq Z$, but this is now impossible. BJ (1970) acknowledged the cost of excess demand, viz. rationing, in a reply to comments by Salkever (1970) and Turvey (1970), but no attempt was made to add this cost to the model. However, suppose a penalty cost of ϕ is assessed for each unit of excess demand; ϕ can be thought of as the cost of rationing, including the cost of ranking consumers (see Crew and Kleindorfer (1976)). To account for this cost, the following terms, one for each period, are subtracted from the right hand side of (1):

$$\phi\int_{Z/X_i(p\ )}^{\infty} f_i(u_i)\ [X_i(p_i)u_i - Z]du_i.$$

The necessary conditions for optimum prices now yield

$$p_i = b + \frac{I_i}{I - I_i} \tag{7}$$

where $I_i = \int_{Z/X_i(p_i)}^{\infty} u_i f_i(u_i)du_i$ is the truncated expected value of u_i when there is excess demand.

Clearly, prices now exceed those given by (6). The value of the integral expressions will depend on the capacity, Z, and $f_i(u_i)$. Accordingly, $0 \leq I_i \leq 1$ because as Z approaches zero, the value of the integrals monotonically approach $E[u_i] = 1$. Thus, the smaller the level of capacity the higher the price; of course the higher price mitigates potential rationing costs. Also, if ϕ increases, indicating higher rationing costs, then prices increase in response.

An alternative approach to the problem of excess demand that also admits higher prices is to introduce chance constraints on demand. This approach is used by Meyer (1975, p. 331) to obtain prices that are "... sufficiently high to meet specified standards of system reliability."

Thomas, et. al. (1972) use a similar approach to allocate uncertain water supplies. The procedure is to add the following constraints to the problem:

$$P\{X_i(p_i)u_i \leq Z\} \geq \varepsilon_i > 0, \tag{8}$$

where P indicates probability. The value of the ε_i reflects the stringency of system reliability requirements: larger ε_i imply more stringent requirements. Let λ_i be the Lagrangian multipliers so that the appended problem is

$$\text{Max } E[W] + \sum_{1}^{n} \lambda_i \left(\int_0^{Z/X_i(p_i)} f_i(u_i)du_i - \varepsilon_i\right)$$

From the necessary conditions for a maximum, prices are

$$\bar{p}_i = b + \lambda_i \frac{\theta_i f(\theta_i)}{X_i(p_i)[1 - I_i]} \tag{9}$$

where $\theta_i = Z/X_i(p_i)$. Since the λ_i are nonnegative, prices equal or exceed short-run marginal cost. As Meyer (1975, p. 334) points out, "... a higher price using risk constraints implies capacity could be reduced over what would have been required to meet the constraints while charging a price of b."

This chance constraint approach to excess demand is representative of current practice in some countries. For instance, in the U.S., electric power generation systems use a "1-day-in-10-year" loss-of-load probability as a reliability target.[7] Crew and Kleindorfer (1976, 1978) use the rationing cost approach and argue that establishing safety margins (i.e., the chance constraint approach) becomes very complex under a diverse technology. Also, the chance constraint approach does not account for the cost of ranking consumers in order of willingness to pay. Of course, the rationing cost approach is not without disadvantages; in particular, there is the problem of measuring rationing costs. If this proves to be a difficult and costly task, using large values of ε_i in the chance constrained approach becomes more attractive. The large ε_i imply that excess demand occurs less frequently, and the simplifying assumption that consumers are ranked properly is less troublesome.[8]

III. Budget Constraints

While prices given by (7) or (9) maintain excess demand at acceptable levels, they do not ensure adequate revenues.[9] This is an important omission if the utility in question is "a tub that must stand on its own bottom." Under risk, prices that will always produce adequate revenues may be nonexistent just as always eliminating excess demand may be impossible. Thus, here again, is a place for some form of risk constraints to replace the standard riskless budget constraints.

Sherman and Visscher (1978) address the revenue problem by utilizing the following constraint:

$$E[TR] = E[TC] \tag{10}$$

Expected welfare is maximized subject to this expected balanced budget constraint, and the optimum prices and capacity are derived. The optimum prices can be written as

$$\frac{p_i - b}{p_i} = \frac{\lambda}{1 + \lambda} \frac{1}{E[\eta_i]} \tag{11}$$

where λ is a Lagrange multiplier and $E[\eta_i]$ is an expected price elasticity. These prices are identical to welfare maximizing Ramsey prices in a nonstochastic framework, except that $E[\eta_i]$ replaces the usual definition of price elasticity.[10] Sherman and Visscher (1978, p. 46) state that these prices and the optimum capacity yield "... a reliability of service that is optimal given the break-even constraint, and thus does not have to be arbitrarily imposed (see Meyer)."

A second way of dealing with the revenue problem is to use a chance constraint. A certain probability distribution for TR and TC is associated with each choice of Z and p. With some probability distributions, the probability that TR equals or exceeds TC, P {TR $\geq$ TC}, will be higher than with other probability distributions. Since the utility is now confronted with a budget constraint, values of p and Z must be selected so that P {TR $\geq$ TC} is as great as possible, or at least as great as some minimum value, say α. Therefore, the budget constraint for the utility is

$$P\ \{TR - TC \geq 0\} \geq \alpha. \tag{12}$$

The form of the pricing rules is shown in the Appendix. A numerical example is provided now that allows for a graphical interpretation of all the pricing rules discussed.

IV. Comparison of Pricing Rules with a Numerical Example

For convenience, a single period is considered and subscripts are dropped. Let demand be given by

$$X(p)u = (30 - p)u$$

and assume that capacity is fixed at Z = 20.[11] Also, let b = 2, β = 8, and the random disturbance is continuous in the interval (0,2). Figure 2 again depicts the situation. The mean demand curve is labelled X and three other demand curves are drawn as previously indicated. The capacity cost is βZ = 160 and is enclosed by lines b, b + β, Z and the price axis. The operating cost will depend on price and u. Four prices are labelled: p = 2, p' = 10, p" = 16 and p"' = 20.

Figure 3 illustrates welfare and profit for the four prices as the disturbance term varies between 0 and 2. In Figure 3B, a simple density function for u is given where the expected value of u is one. Consider first, price $p = 2$ which is equal to short-run marginal cost b in Figure 2. As u varies, the short-run operating cost will be covered, but there will be a constant deficit equal to the long-run capacity cost $\beta Z = 160$. This deficit is represented by the solid horizontal profit curve $a\pi$ in Figure 3A. Welfare for price $p = 2$ is given by the heavy curve awW. This is derived by evaluating welfare for each value of u. For example, when $u = 1$, realized demand at $p = 2$ is $Z = 20$ with an excess demand of 8 (see Figure 2). Welfare is CS(392) - L(32) + TR(40)- TC(200) = 200. The heavy linear segment aw represents values of u for which there is idle supply, while the heavy nonlinear segment wW represents excess demand. The u value at point w on this heavy curve is obtained by noting at w, $X(p)u = Z$ for $p = 2$ (see Footnote 6.)

Next, consider price $p' = b + \beta = 10$. Profit is given by the solid kinked curve $a\pi'$. As u increases from 0 to 1, profit increases from -160 to 0. At $u = 1$, demand is exactly equal to capacity as depicted in Figure 2. Since price equals long-run marginal cost and demand equals capacity, profit is zero. For $u > 1$, total revenue and total cost are the same as for $u = 1$, since no additional demand can be served. Thus, the kink in curve $a\pi'$ occurs where demand equals capacity. From a profit standpoint, price p' obviously dominates price p. Welfare for p' is given by the heavy curve aw'W which is linear to the point w' where demand equals capacity. At point w', $u = 1$ and welfare from Figure 2 is CS(200) - L(0) + TR(200) - TC(200) = 200. From a welfare standpoint, price p obviously dominates price p'. Note that the welfare curves do not level off after demand equals capacity as the profit curves do, since consumers' surplus continues to increase when demand increases beyond capacity.

Individual derivation of the curves for prices p" and p"'are left to the reader. The prime notation on the prices in Figure 2 correspond to the primes used in Figure 3 for the π's and w's. The reliability of service and the probability of covering costs are indicated on the density function. Price p' allows a 50% reliability given by the area to the left of θ', and a 50% probability of covering costs given by the area to the right of t'. For price p", a 79% reliability is obtained, given by the area to the left of θ'', and a 62% probability of covering costs is obtained from the area to the right of t".[12] There is zero probability of covering cost with price p, so there is no corresponding t along the density function. Naturally, values of the θ's on the extreme right and values of the t's on the extreme left are desirable. The reliability constraint, (8), sets the positions of the θ's, while the chance budget constraint (11), sets the positions of the t's.

To illustrate the tradeoffs and properties of the various solution concepts, Figure 4 is constructed using Figure 3. For each price in Figure 3, the profit and welfare curves are weighted by the density function to obtain the expected profit and expected welfare plotted in Figure 4A. Thus, each profit (welfare) curve in Figure 3A corresponds to one point on the expected profit (welfare) curve in Figure 4A. In addition, Figure 4B shows the reliability for each price obtained from the θ values in Figure 3B. For example, consider price p'. To obtain

$E[\Pi]$ from Figure 3A, the kinked segment can be broken into four intervals as follows:

Interval	Average Profit Over Interval	Weight of Interval From Density Function	Expected Profit Over Interval
$0<u<.5$	-120	16 2/3%	-20.0
$.5<u<1.0$	- 40	33 1/3%	-13.3
$1.0<u<1.5$	0	33 1/3%	0
$1.5<u<2$	0	16 2/3%	0
		Expected profit	-33.3

The expected profit of -33.3 is at point d in Figure 4A. Tracing point d down to Figure 4B reveals that the actual reliability of service for p' is $\theta' = 50\%$. Each point on the $E[W]$ and $E[\Pi]$ curve can be derived geometrically in a like fashion. For an alternative derivation of Figure 4A, (1) could be used to determine $E[W]$ and $E[\Pi] = E[TR - TC]$, for each value of price and where $Z = 20$. The density function used is

$$f(u) = \begin{cases} 1/3 & .0 < u < .5 \\ 2/3 & .5 < u < 1.5 \\ 1/3 & 1.5 < u < 2.0 \end{cases}$$

The various pricing solutions discussed in Sections I-III appear in Figure 4, for the fixed capacity $Z = 20$. Expected welfare maximization, the BJ approach, is attained at $p = b$, but expected profit at this price indicates a deficit of -160 and reliability is only 30%. Expected profit maximization is attained at p". The Sherman and Visscher solution is $\hat{p}$ where expected welfare is maximized subject to (10), the expected break-even constraint. Reliability in this case is about 60%. To obtain Meyer's solution where the reliability constraint in (8) is used, a reliability must be specified. Suppose a "one-day-in-ten-year" loss-of-load probability is required so that reliability is virtually 100% (i.e., $\varepsilon_i \approx 1.0$). This requires that price p‴ be levied.

The alternative methodology, eschewing constraints and using rationing costs, can also be illustrated. Suppose rationing costs consist of two parts as suggested by Crew and Kleindorfer: 1) the cost of organizing and administering the program, and 2) the inconvenience of being cut off which will depend on willingness to pay. The former costs must be subtracted from expected profit, and both costs must be subtracted from expected welfare for each possible price below p‴. Prices above p‴ do not require rationing as Figure 2 indicates. Suppose the former cost is b per unit to be rationed, and the latter cost is L. The dashed curves in Figure 4A represent the revised expected welfare and expected profit given these costs. Optimum price is then $\bar{p}$ with a reliability of about 65%.

A comparison of all these prices indicates that p will yield the largest expected welfare, the largest expected deficit, and the lowest reliability. The relationships among the other prices will depend on the particular example used, but clearly tradeoffs among expected welfare, expected profit, and reliability are apparent. In particular, neither a budget constraint nor rationing costs alone ensure high reliability levels, and a stringent reliability constraint alone will not ensure a nonnegative $E[\Pi]$.[13]

V. Demands Dependent on Reliability

In all of the above analysis, demand has been assumed independent of the reliability of service. Figure 4B illustrates that as price varies between zero and p''', reliability varies between 28% and 100%. Yet, in spite of this wide reliability variation, the mean demand curve is stationary. This unrealistic assumption needs to be recognized. A reliable product is certainly of higher quality than an unreliable product and demand should reflect this important point.

Figure 4 brings out this shortcoming very well. Setting a high reliability via the constraints should have some payoff in terms of welfare. Otherwise, utilities would not strive for or be forced to strive for the high reliabilities that prevail in the industry. Yet the solid curve in Figure 4A implies that expected welfare monotonically declines with increased reliabilities between p and p'''. Thus, the models discussed, with the exception of the Crew and Kleindorfer (1976, 1978) model (the dashed curve), do not recognize any welfare payoff to high reliabilities. Crew and Kleindorfer allow a payoff to high reliabilities by adding the rationing cost which includes the inconvenience of being cut off. But the inconvenience should be reflected in the demand function itself, so that greater inconvenience means lower consumers' surplus.[14]

In other words, a price of, say, p = b with a 30% reliability is likely to have feedback effects on X(p), because consumers may seek an alternative supply to ensure a greater reliability. To allow for this dependency, rewrite mean demand as $X(p_i,\rho_i)$, where ρ_i is the reliability of service quoted to the consumers in period i. Also, assume that $\frac{\partial X_i(p_i,\rho_i)}{\partial \rho_i} > 0$, so that higher reliabilities imply greater demand. The expected welfare problem is still given by (1), except that $X_i(p_i,\rho_i)$ replaces $X_i(p_i)$. To ensure that the actual reliabilities are consistent with quoted reliabilities, n constraints given by (13) are needed.

$$P\{X_i(p_i,\rho_i)u_i \le Z\} \ge \rho_i \tag{13}$$

Constraint (13) is very similar to (8). The important difference is that with (13) the optimum reliabilities are chosen, given that demand depends on reliabilities. In (8), reliabilities were arbitrarily imposed and had no effect on demands.

Maximization is performed over all prices, reliabilities, and capacity. The condition for prices is identical to that given in (9) with $X_i(p_i)$ replaced by $X_i(p_i,\rho_i)$. Of course, price will not be the same as that given in (9), unless ε_i just happened to be chosen optimally in (8). The condition for optimum capacity does not change, except again the new demand function must be inserted (See Meyer's equation (13)). However, there are n new conditions corresponding to the n quoted reliabilities. The condition for an optimum ρ_i is given by (14).

$$\frac{\partial X_i}{\partial \rho_i}[p-b][1-I_i] + \int_0^\infty f_i(u_i)\int_{p_i}^{X_i^{-1}(0)} \frac{\partial X_i}{\partial \rho_i} u_i dp_i du_i$$

$$-\int_{Z/X_i}^\infty f_i(u_i)\int_{p_i}^{X_i^{-1}(Z/u_i)} \frac{\partial X_i}{\partial \rho_i} u_i dp_i du_i = \lambda_i\left[\frac{\partial X_i}{\partial \rho_i}\frac{\theta_i f(\theta_i)}{X_i(p_i,\rho_i)} + 1\right] \tag{14}$$

Substituting (9) for the first term and simplifying yields

$$\int_0^{Z/X_i} f_i(u_i)\int_{p_i}^{X_i^{-1}(0)} \frac{\partial X_i}{\partial \rho_i} u_i dp_i du_i + \int_{Z/X_i}^\infty f_i(u_i)\int_{X_i^{-1}(Z/u_i)}^{X_i^{-1}(0)} \frac{\partial X_i}{\partial \rho_i} u_i dp_i du_i = \lambda_i \tag{15}$$

Equation (15) states that the Lagrange multiplier, which is the marginal welfare cost of requiring that actual reliabilities are as great as those quoted, is equal to the marginal welfare benefits from a change in ρ_i. The first term on the left-hand side is the marginal benefit when there is idle capacity, and the second term is the marginal benefit when there is excess demand. The left-hand side of (15) is obviously positive so that $\lambda_i > 0$ and the constraints are binding. For a binding constraint, the actual reliability equals the quoted reliability. The producer does not provide a greater reliability than that which is promised, since reliability is costly.

These results can be understood easily by relating them to the numerical example. Rewrite demand as

$$X(p,\rho)u = \frac{(30-p)\rho}{.8} \quad u$$

The .8 is arbitrary and the implication is that the previous example depicted in Figures 2-4 is now a special case where quoted reliability is 80%, or $\rho = .8$. Obviously, $\frac{\partial X}{\partial \rho} > 0$ as assumed. The expected welfare curve and reliability curve in Figures 4A and 4B are duplicated in Figures 5A and 5B and labelled as 80%. The entire E[W] curve in Figure 5A (or Figure 4A) is derived given that the consumer is quoted an 80% reliability level. However, only one point on this curve is actually compatible with an 80% reliability, because only at one point is the constraint satisfied. The compatible point can be found by locating an actual reliability of 80% on the curve labelled 80% in Figure 5B, (point b) and then tracing this point vertically to the curve labelled 80% in Figure 5A (point a). Point a then is the only pertinent point on the 80% curve in Figure 5A.

Conceptually, the entire procedure used to derive the curves in

Figure 4 and duplicated in Figure 5 must be repeated for all values of ρ between 0% and 100%. This gives rise to a family of curves. In Figure 5, six sets of curves from this family are sketched for ρ values of 30%, 53%, 60%, 70%, 80%, and 100%. The dashed lines connecting Figures 5A and 5B show the compatible points such as point b in Figure 5B and point a in Figure 5A. The resulting curve, cdrae, in Figure 5A illustrates the expected welfare over all prices. Any point on this curve can be thought of as the solution to an expected welfare maximization problem where ρ is fixed at $\bar{\rho}$ and demand is $X(p,\bar{\rho})u$. This would be the equivalent to Meyer's problem where $\bar{\rho} = \varepsilon$ in (8), except that with each different specification of $\bar{\rho}$, mean demand is different.

The linear segment, cd, of curve cdrae in Figure 5A reveals that for quoted reliabilities between 0% and 53%, the optimum price is $p = b = 2$. This is because values in this range (e.g., $\rho = 30\%$ is illustrated) yield demands that are very small, and actual reliability exceeds quoted reliability. The reliability constraint is nonbinding ($\lambda = 0$) and from (9), $p = b$. Essentially, the BJ result is obtained for the reasons cited in Section II. Note that welfare increases with increases in quoted reliabilities between 0% and 53%, although price remains at b. Above 53%, a price of b would cause actual reliabilities to fall short of quoted reliabilities. To avoid this constraint violation, price is raised above b as quoted reliabilities increase. In spite of the increase in price, however, welfare continues to increase up to its maximum of point r where quoted reliability is about 75% and price is slightly less than p''. Thus, for segment dr, the gains in welfare due to higher reliabilities outstrip the losses in welfare due to higher prices. The reverse occurs for segment rae, where welfare declines. For completeness, curve ce illustrates expected profit in Figure 5A.

Figure 5B shows the relationship between actual and quoted reliabilities. Curve gfb represents the actual reliabilities. For quoted reliabilities between 0% and 53%, the actual reliabilities are greater, giving rise to the linear segment gf. As an example, consider the quoted reliability of 30% and the line segment labelled the same. The intersection of this segment and gf is at an actual reliability of about 97%. As quoted reliabilities are raised above 30%, actual reliabilities fall from 97%, until the two merge at 53%. Finally, as quoted reliabilities are raised above 53%, actual reliabilities keep pace via price increases.

The advantage of formulating the problem with ρ as a variable is reflected in curve cdrae, because there is now a tradeoff between prices, reliability and welfare that did not exist previously. As price increases and reliability increases, there may be gains in welfare that were not recognized in previous formulations.

VI Observations

The addition of ρ in the demand function brings the analysis of public utilities with stochastic demand closer to a position of being able to assess accurately the welfare implications of pricing and capacity choices. Reliability is an important aspect of utility service that has not been fully appreciated. While Meyer's method of adding reliability constraints covered an important omission in the BJ analysis, the model was inconsistent in the following sense. High reliability levels were said to be required by regulatory agencies, where these agencies presumably have the consumers' well-being in mind. But the welfare function did not bear out the regulatory decision to require high reliabilities, since welfare decreases with increased reliabilities as seen in Figure 4. Also, maximizing the BJ welfare function constrained by (S) for various values of ε, and then selecting the maximum of the maximum would still yield $p = b$ and most likely a low reliability. Only when reliability is allowed as an argument in the demand function can there be improvements in welfare due to higher reliabilities.

The example used in Section V indicated that optimum reliability was about 75%. Of course, this is entirely dependent on Z, the specific demand function, and on the way in which ρ entered this function. Empirical work on demand estimation with reliability as an argument would be difficult in most countries, since historically, reliability has been on the order of 99% with little variation. The analysis here, however, indicates that there may be welfare gains available through decreases in reliability. Consumers, if given a chance, may opt for lower reliabilities if accompanied by lower prices. A reliability of 75% would undoubtedly be unacceptably low, but a "1-day-in-1-year" (99.73%) or a "1-day-in-5-year" (99.95%) loss-of-load probability may be a better target than "1-day-in-10-year" (99.97%).[15]

Provided that consumers desire less than 100% reliability, rationing is still required.[16] Therefore, rationing costs that include the costs of organizing and administering the program could be included. The result would be that the E[W] curve in Figure 5A would be lower in the same way that rationing costs lowered the E[W] curve in Figure 4A. Note that including willingness to pay in the rationing costs as done in Section IV is unnecessary, since these costs are reflected now in the demand function.

VII Relation to Load Management

The literature cited and examined above is perhaps the closest there is to a theoretical treatment of load management. But it tends to gloss over the rationing method and the method of managing the load. That is, when consumers are rationed, how is this accomplished?

Regardless of the specific technique utilized, load management implies that the consumer will, at times, receive less service than he demands. He is being rationed. The benefits of this are chiefly the cost savings to the utility of not having to satisfy all demands. This may include both short-run fuel cost savings and long-run capacity cost savings. The costs of load management, however, are less clear. There is a revenue loss to the utility, and a loss in welfare to the consumer. If consumer's surplus is used as a measure of welfare, then knowledge of the demand curves is needed to quantify this loss to the consumer. And if consumers are not identical, then interruption of service will create losses that depend on which consumer is interrupted. Unfortunately, price elasticity alone does not provide a guide as to who should be interrupted. For example, Figure 6 illustrates three cases, with two consumers in each case. The price, p^*, is the same for both consumers, and the demand curves are labelled e for relatively elastic and i for relatively inelastic where elasticity is evaluated at p^*. If service is to be interrupted by quantity Δq, then minimizing welfare loss calls for interrupting e for case a, i for case b, and either e or i for case c.

Figure 6 suggests that the slopes of the demand curves are crucial in determining who should be interrupted. To put this in more precise terms, consider the two consumers whose demands are pictured in Figure 7. Consumer i, i=1, 2, faces price p_i and demands and receives quantity q_i'. Marginal cost is given by MC. The total cost of serving the two consumers is $C(q_1' + q_2')$ and total welfare is given by

$$W' = \int_o^{q_1'} p_1(q_1)dp_1 + \int_o^{q_2'} p_2(q_2)dq_2 - C(q_1' + q_2')$$

Now suppose load management is exercised, and a total of $\bar{q}$ units is to be subtracted from q_1' plus q_2'. How should this loss of service be allocated without changing prices? Let q_i'' be the quantity received by consumer i after the curtailment. Welfare becomes

$$W'' = \int_o^{q_1''} p_1(q_1)dq_1 + \int_o^{q_2''} p_2(q_2)dq_2 - C(q_1'' + q_2'')$$

and the welfare loss is W' - W". The problem is to minimize this loss, or

$$\min_{q_1'',q_2''} W'-W'' = \int_{q_1''}^{q_1'} p_1(q_1)dq_1 + \int_{q_2''}^{q_2'} p_2(q_2)dq_2 - C(q_1' + q_2') + C(q_1'' + q_2'')$$

subject to $q_1' - q_1'' + q_2' - q_2'' = \bar{q}$. The solution yields

$$p_1(q_1'') - MC = p_2(q_2'') - MC \tag{16}$$

where $p_i(q_i'')$ is the market clearing price for quantity q_i''. This is not the actual price, since prices do not change. Equation (16) implies that interruptions should be made such that the difference between the market clearing price at the new quantities and marginal cost is equated across consumers. For the simple case of constant marginal cost, Figure 7 shows the service losses for both consumers as $q_i' - q_i''$. The shaded areas are the welfare losses. Both losses are greatest for consumer 2, since his demand curve is flatter in the neighborhood of prices.

The simple example derived the optimum intettuptions for two consumers given exogenously determined starting prices and quantities. In an optimum load management program, the prices and quantities are not exogeneous, but are an integral part of the program. That is, load management should not be instituted on top of an already existing rate structure. Instead, the rate structure and load-management technique must be determined simultaneously.

To illustrate this, a simple stockastic model will be used. The stochastic element enters through a random demand curve. This approach seems necessary in the model, since random fluctuations in demand are really the impetus for load management. Again consider two consumers with peak-period demands as pictured in Figure 8. Off-peaks will be ignored since load management only takes place during peaks. The two curves for each consumer represent demands in the two possible states of the world. Consumer i has demand curve p_i in state one which occurs with probability α, and has demand curve P_i in state two which occurs with probability $1-\alpha$. Both consumers have greater demands in state two which would seem reasonable since both consumers are exposed to the same random elements such as weather conditions. Marginal costs are again constant and given by operating cost, b, and capacity cost, β. The assumption is that prices are set and all demands are satisfied for curves p_1 and p_2 in state one. But in state two, demands increase at these prices, and interruptions are necessary.

Given this information, welfare in state one is

$$W_1 = \int_o^{q_1'} p_1(q_1)dq_1 + \int_o^{q_2'} p_2(q_2)dq_2 - bq_1' - bq_2' - \beta z$$

where q_i' is the quantity consumed by consumer i in state one. In state two, welfare is

$$W_2 = \int_o^{q_1''} P_1(q_1)dq_1 + \int_o^{q_2''} P_2(q_2)dq_2 - bq_1'' - bq_2'' - \beta z$$

where q_i'' is the quantity allocated to consumer i. The problem is to maximize expected welfare, or

$$\max_{q_1'',q_2',q_1'',q_2'',z} \quad \alpha W_1 + (1-\alpha)W_2$$

subject to the following constraints:

$$P_1(q_1'') \geqq p_1(q_1') \qquad (\gamma_1)$$

$$P_2(q_2'') \geqq p_2(q_2'') \qquad (\gamma_2)$$

$$q_1'' + q_2'' \leqq z \qquad (\lambda_1)$$

$$q_1' + q_2' \leqq z \qquad (\lambda_2)$$

The letters in parenthesis to the right of each constraint are the multipliers. The first two constraints require the market clearing price in state two to be at least as great as the actual price which prevails in both states. In other words, in state two each consumer can obtain at most the quantity he demands in state two at the actual price. Of course, if the quantities demanded in state one are equal to capacity, then it is impossible for both consumers to receive all they demand in state two. The constraints would necessarily be satisfied by an inequality for at least one consumer. The third and forth constraints ensure that output does not exceed capacity. Also, if the last constraint is satisfied by equality, load management must be practiced in state two. This is because quantities demanded in state two exceed quantities demanded in state one and if capacity is fully utilized in state one, there must be interruptions in state two.

The conditions for maximum expected welfare (available from author upon request) can be interpreted as follows. If load management is applied to both consumers, then $P_i(q_i'') > p_i(q_i')$ and $\gamma_i=0$ for $i=1,2$. This does not say, however, if q_i' is greater than, less than, or equal to q_i''. Both consumers may be cut back to their original quantities (q_1' and q_2'), or consumer one (two) may be cut back to less than $q_1'(q_2')$ while consumer two (one) gets more than $q_2'(q_1')$. In any case the result for pricing is

$$\alpha p_1(q_1') + (1-\alpha)P_i(q_i'') = b + \beta \qquad (17)$$

for $i=1,2$. A weighted average of the actual price and the market clearing price in state two must be equal to long-run marginal cost. If there is a low probability of load management occuring (α is close to one), then the actual price is slightly less than long-run marginal cost.

The reader who is familiar with Williamson's paper on peak-load pricing will note the similarity between (17) and Williamson's expression for optimum peak and off-peak prices. In his framework, p_i is the off-peak price and P_i is the peak price. The probabilities, α and $(1-\alpha)$, are the portion of the day that the off-peak and peak are in effect. An important distinction is that in the peak-load framework, both p_i and P_i are actually charged while in the load-management framework p_i is the only actual charge. The point is that optimum load management can be carried out via the pricing mechanism alone. When demand exceeds capacity, prices are immediately increased to limit quantity demanded. And ideally, different consumers will face different price hikes. In reality, however, prices cannot be adjusted at a moments notice. There are no digital displays at the site of power consumption that keep the consumer abreast of price changes. And traditionally, stability of electric rates has been considered a virtue. Barring rapid price adjustments, load management is used, but the concept of price is still the key. That is, service is interrupted to the point where the consumer's new quantity is that which would be consumed if a price hike were permissable.

While the above pricing rules maximize welfare, a complication arises on the profit side. In Williamson's peak-load model, revenues will exactly cover costs and profits are zero owing to the constant marginal costs. But with the rule given by (2), costs are not covered. This is because $P_i(q_i'')$, which must exceed $b + \beta$, is not actually charged. Price is always $p_i(q_i')$ which is less than $b + \beta$ causing negative profits. To alleviate the deficit, the following budget constraint could be utilized:

$$\alpha\pi_1 + (1-\alpha)\pi_2 = 0$$

where

$$\pi_1 = p_1(q_1')q_1' + p_2(q_2')q_2' - bq_1' - bq_2' - \beta z$$

Solution of this appended problem would yield an actual price and market clearing price that deviate from the above solution much as Ramsey prices deviate from marginal costs.

A complete theory of load management awaits further work. This paper has sought to point out some of the issues involved. They involve the usual pricing and capacity decisions in addition to questions of reliability, demand response to reliability changes, ordering of consumer interruptions, and budget requirements.

Appendix

In the first part of this Appendix, the derivation of the stochastic Ramsey prices given by (11) is provided; in the second part, the prices alluded to at the end of Section III are derived.

Part 1. Ramsey prices are described by Baumol and Bradford (1970) as optimal departures from marginal cost prices. In a nonstochastic framework they take the form

$$\frac{p_i - b}{p_i} = \frac{\lambda}{1 + \lambda} \frac{1}{\eta_i}$$

where η_i is price elasticity. Sherman and Visscher (1978) derive this price rule in the stochastic framework in their equation (11), given here for the multiplicative demand case as equation (18).

$$\frac{[p_i - b] \int_0^{Z/X_i(p_i)} u_i f(u_i) du_i}{p_i} = \frac{\lambda}{1 + \lambda} \frac{1}{\eta_i} [1 - E[ed_i]/X_i(p_i)] \qquad (18)$$

The integral term in (18) is the truncated expected value of u_i when there is excess supply, and $E(ed_i)$ is the expected value of excess demand, or

$$E[ed_i] = \int_{Z/X_i(p_i)}^{\infty} [X_i(p_i)u_i - Z] \, f_i(u_i) du_i .$$

Since the problem is to maximize expected welfare by setting expected marginal revenues at the optimum levels, it follows that expected elasticities are needed. That is, the monopolist is interested in an elasticity of expected sales instead of the usual elasticity of demand, since sales do not always equal demand in the stochastic case. Let S_i be sales in period i, and define elasticity of expected sales as

$$E[\eta_i] = - \frac{\frac{\partial E[S_i]}{\partial p_i} p_i}{E[S_i]} \qquad (19)$$

where

$$E[S_i] = \int_0^{Z/X_i(p_i)} X_i(p_i)u_i f(u_i)du_i + \int_{Z/X_i(p_i)}^{\infty} Zf(u_i)du_i. \tag{20}$$

The two terms in (20) represent sales when there is excess supply and excess demand. The partial derivative of (20) with respect to p_i is

$$\frac{\partial E[S_i]}{\partial p_i} = X_i'(p_i) \int_0^{Z/X_i(p_i)} u_i f(u_i)du_i \tag{21}$$

Substituting (20) and (21) into (18), utilizing the definition in (19), and noting that $X_i(p_i) = \int_0^{\infty} X_i(p_i)u_i f(u_i)du_i$, yields the stochastic Ramsey rule

$$\frac{p_i - b}{p_i} = \frac{\lambda}{1 + \lambda} \frac{1}{E[\eta_i]} \tag{11}$$

Equation (11) is a generalization of the Ramsey pricing rule that illustrates the robustness of the Ramsey theory. The stochastic definition of elasticity can also be used to simplify the pricing rule for the unregulated monopolist given by Mills' equation (4.4), Meyer's equation (6), or Sherman and Visscher's equation (5). See Tschirhart (1978) for more detail.

<u>Part 2</u>. Budget constraint (12) for a single period can be written as

$$P\{X(p)u[p-b] - \beta Z \geq 0\} \geq \alpha \tag{22}$$

Let $t(p,Z) = \dfrac{\beta Z}{X(p)[p-b]}$ so that (20) becomes

$$P\{u \geq t(p,Z)\} \geq \alpha \tag{23}$$

The t function gives the break-even value of u for a particular p and Z, and several t's are depicted in Figure 3B. Constraint (23) does not eliminate the possibility that $p < b + \beta$, and costs could never be covered for prices in this range. The problem is that (23) does not distinguish between those values of u that result in excess demand and those that do not. A second constraint can be added to ensure that $p \geq b + \beta$; however, a reliability constraint will serve the same purpose. To show this, rewrite (8) as $P\{u > Z/X(p)\} < 1 - \varepsilon$, and combine this with (22) to obtain

$$P\{u \geq \frac{\beta Z}{X(p)[p-b]}\} - P\{u > \frac{Z}{X(p)}\} \geq \alpha + \varepsilon - 1 > 0. \tag{24}$$

The last inequality holds for any reasonable values (close to one) of α and ε.

Thus, (24) implies

$$\frac{\beta Z}{X(p)[p-b]} < \frac{Z}{X(p)}$$

or $p > b + \beta$.

The problem is to maximize E[W] subject to (8) and (23). These constraints and the associated multipliers are given in (8') and (23').

$$\lambda(\int_0^{\theta} f(u)du - \varepsilon) \tag{8'}$$

$$\xi(\int_{t(p,Z)}^{\infty} f(u)du - \alpha) \tag{23'}$$

First-order necessary conditions for a maximum require

$$p-b = \frac{\lambda\theta f(\theta)}{X(p)[1-I]} + \frac{\xi f(t)}{X'(p)[1-I]} \quad \frac{\partial t}{\partial p} \tag{25}$$

The first term on the right-hand side of (25) is the same as that found in (9). The second term is the result of adding the budget constraint. The sum of both terms must exceed β as previously indicated. More detail can be found in Tschirhart (1975).

Footnotes

1. For examples of this literature, see Boiteux (1960), Steiner (1957), Williamson (1966), and Crew and Kleindorfer (1975).

2. See Bailey and White (1974) for an example of a budget constraint in a peak-load model under certainty.

3. The assumption of equal demand periods is for simplicity. The results of this paper can be converted to the more general case of unequal periods in the manner done by Williamson (1966).

4. Visscher (1973) points out that the BJ results are dependent on the type of rationing system used. For example, when consumers with the highest willingness to pay are served first as in BJ, price equals short-run marginal cost. But when consumers with the lowest willingness to pay are served first, price equals long-run marginal cost.

5. To obtain this result, it must be the case that the probability that demand is less than capacity is nonzero. Clearly this is true in the BJ formulation, since the lower bound of u is zero and there is always a chance that demand is less than capacity (except for the uninteresting case where $Z = 0$). However, Crew and Kleindorfer (1976) show that if the range and variance of u are small, the probability that demand is less than capacity may be zero. In this case price is indeterminate.

6. The value $u = .71$ is obtained by noting that $X(p)u = Z$ or $(30-p)u = 20$; therefore, if $p = 2$, then $u = .71$.

7. See Telson (1975) for U.S. standards. Standards for other developed countries are also high (see Webb (1977)).

8. A more reasonable assumption as to how capacity is rationed might be the one suggested by Visscher (1973) where capacity is rationed randomly.

9. Using an additive disturbance term, Nguyen (1978) shows that for very high reliability levels (large ε_i) price will be close to $b + \beta$. But this does not rule out large deficits. See footnote 13.

10. See Part 1 of the Appendix for the derivation of (11) and the relationship to Ramsey prices.

11. Fixing Z allows a diagrammatic presentation of the partial effects from changes in price. The optimum prices discussed are optimum relative to the fixed capacity. Conceptually, to obtain the optimum price for the complete welfare problem, a separate set of diagrams could be constructed for each possible capacity level.

12. These figures are calculated as follows: at θ'', demand equals capacity so $(30 - p'')u = 20$ or $u = 1.43 = \theta''$. To find the area to the left of θ'', note that the area to the left of $u = 1$ is 50% and the area between $u = 1$ and $u = 1.5$ is 33 1/3%. We want 43/50 of this latter area, or (43/50)(33 1/3%) = 28.7%. The total area is then 50% + 28.7% $\approx$ 79%. Point t'' is found by noting that TR = TC so that $X(p)up = X(p)ub + \beta Z$. Therefore,

$(30-p'')up'' = (30-p'')ub + \beta Z$ at $p'' = 16$, $b = 2$ and $Z = 160$, which implies $u = .82 = t''$. The area to the right of t'' is calculated in the same manner as the area to the left of θ''.

13. Nguyen (1978) discusses the case of a stringent reliability constraint under additive stochastic demand. Combining the conditions for optimum price and capacity yields a pricing rule, using the above notation and an additive disturbance term, as follows:

$$p = b + \beta - \int_{Z/X}^{\infty} [X^{-1}(Z - u) - p] \; f(u)du$$

The last term is the expected welfare gain from an increase in capacity, and with a stringent reliability this term is approximately zero. Therefore, price is approximately equal to long-run marginal cost. A multiplicative disturbance term gives similar results. But if price is about $b + \beta = 10$ and reliability is about 100%, then optimum capacity must be about $Z = 40$ (see Figure 2). At these values, calculations yield $E[\Pi] = -160$; therefore, a reliability constraint alone does not eliminate deficits.

14. Carlton (1977, 1978) criticizes consumers' surplus as a measure of welfare benefits when demand is uncertain, because it does not reflect consumers' preferences regarding the interaction between price and probability of obtaining the good. The formulation presented here addresses this criticism.

15. Alessio <u>et</u>. <u>al</u>. (1976) discuss the benefits and costs of generation reliability and they cite several studies which conclude that reliability could be lowered without welfare losses. Webb (1977) discusses optimal reserve capacity levels, and suggests that current standards are arbitrarily imposed and too high. He states that optimum reserves should be established by measuring the associated benefits and costs.

16. The rationing in Section V again uses the assumption that those least willing to pay are cut off first. A problem arises because there is no assurance that each consumer receives a reliability of ρ unless consumers have identical demands. One way around this problem is to quote different reliabilities to each consumer and completely curtail one consumer before going on to the next. This would be an example of interruptible service as described in Tschirhart and Jen (1979).

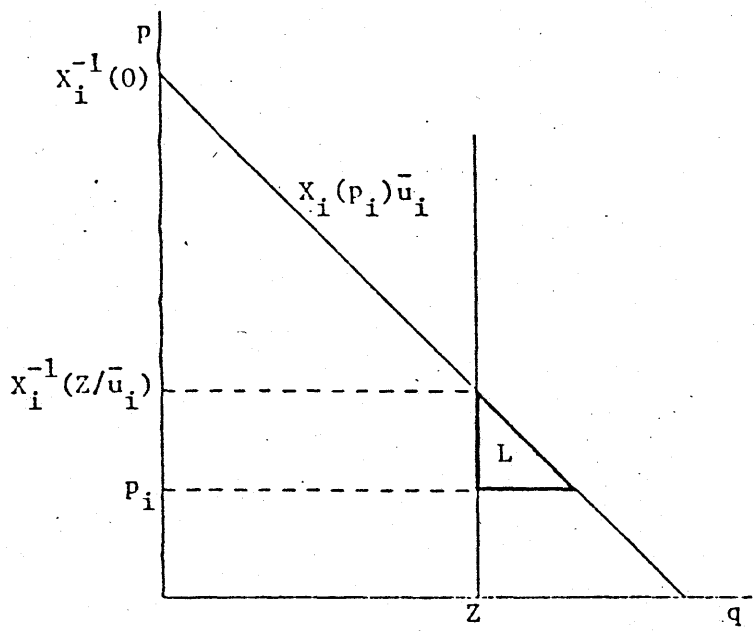


Figure 1

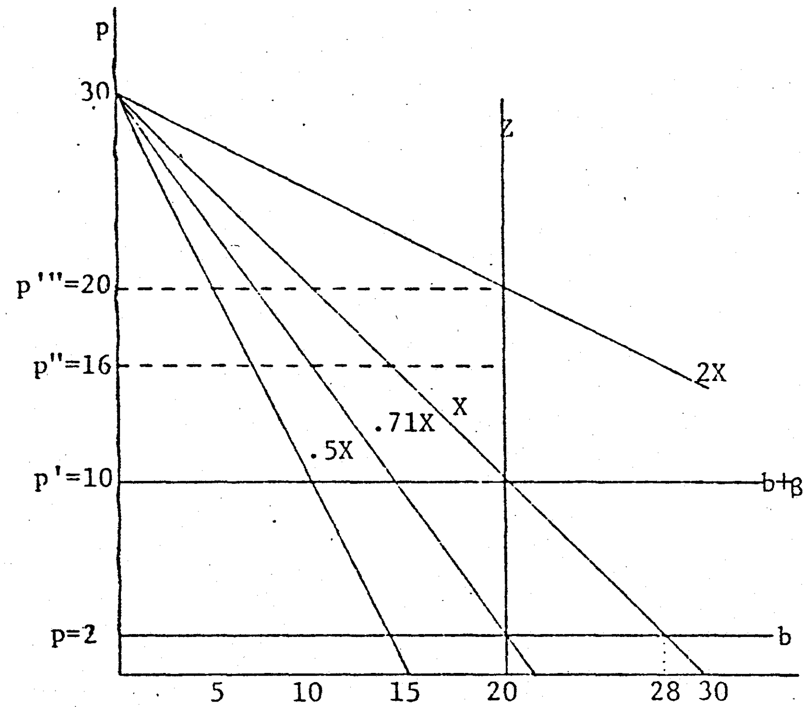
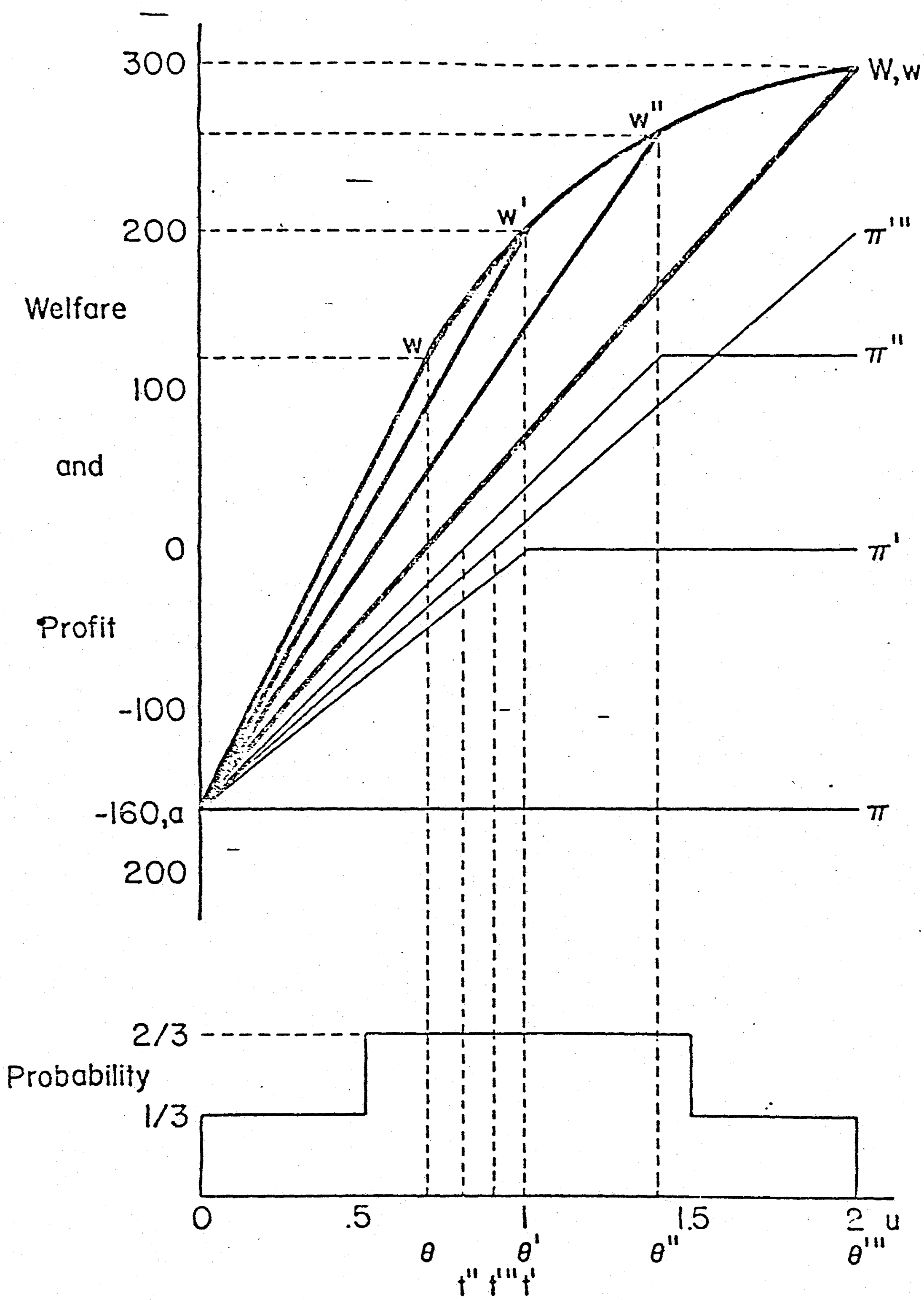


Figure 2

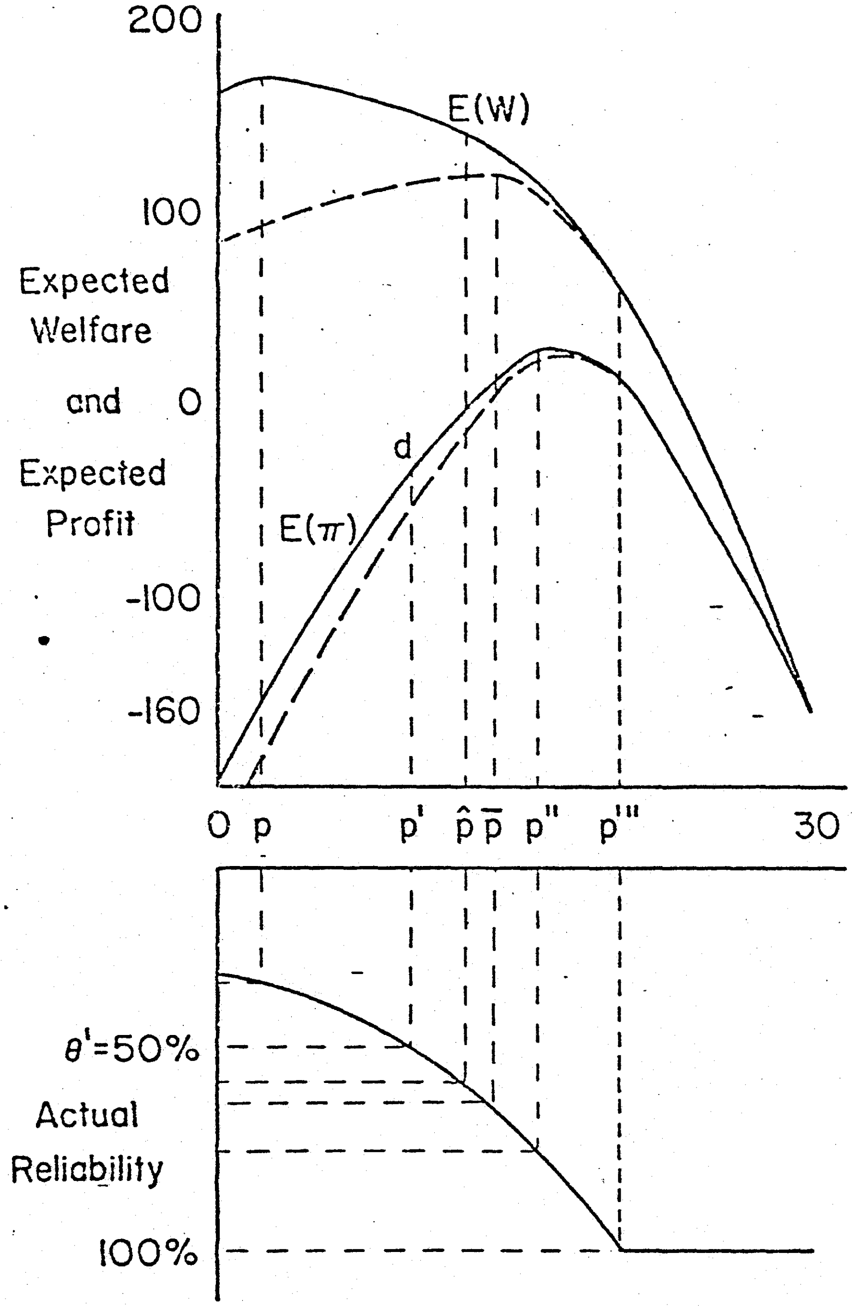


$\theta = 0.71$ $t' = 1.00$

$\theta' = 1.00$ $t'' = 0.82$

$\theta'' = 1.43$ $t''' = 0.89$

$\theta''' = 2.00$



4 A

4 B

Figure 4

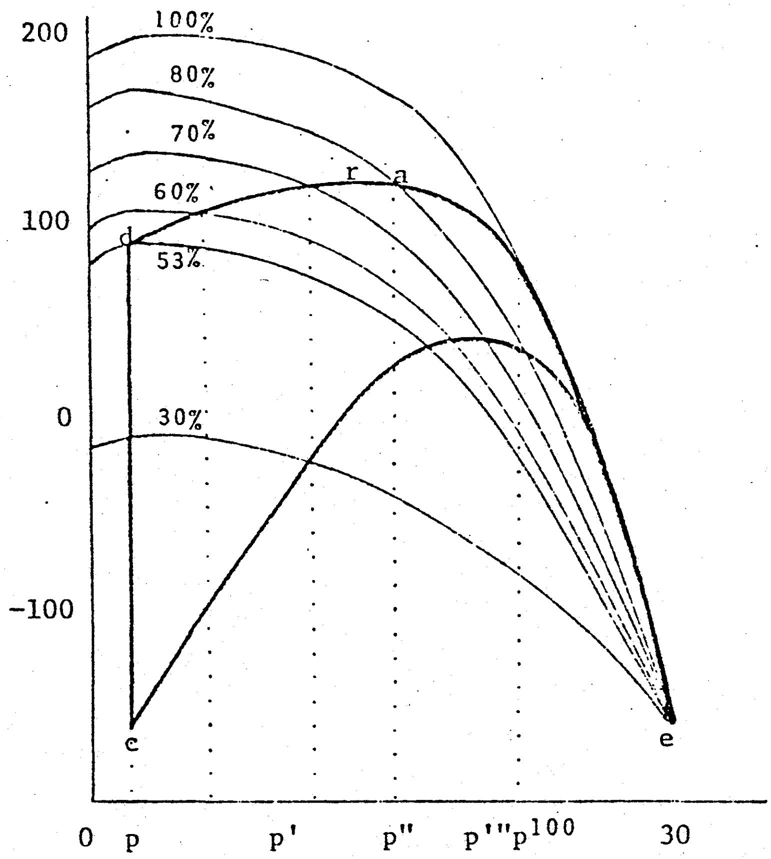


Figure 5A

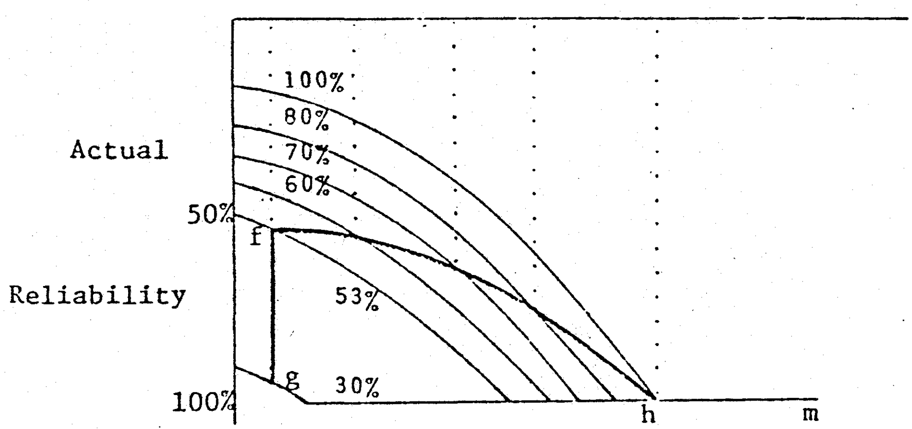


Figure 5B

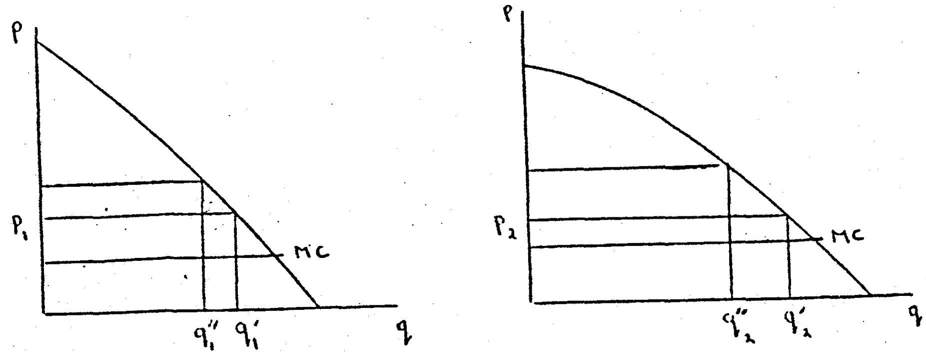


Figure 7

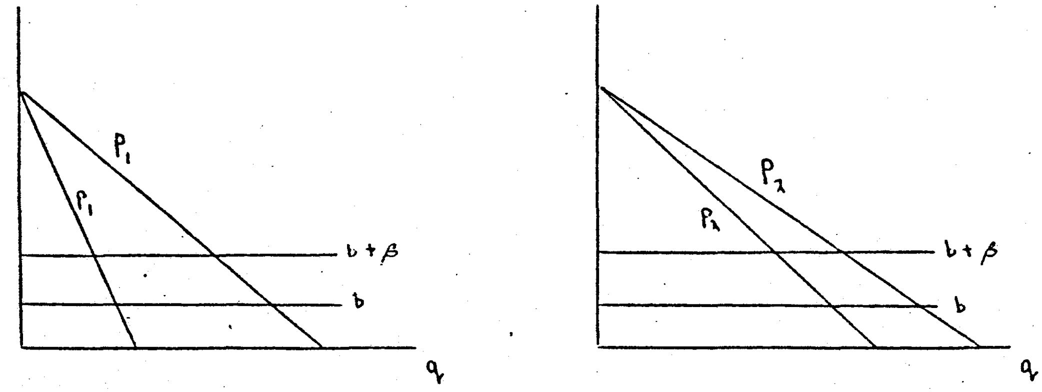


Figure 8