

THE UNIVERSITY OF CALGARY

Priority Pricing

by

Denelle Peacey

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTERS OF ARTS

DEPARTMENT OF ECONOMICS

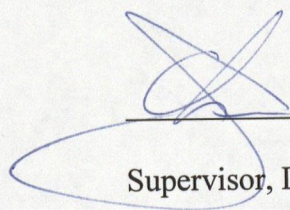
CALGARY, ALBERTA

MAY, 1995

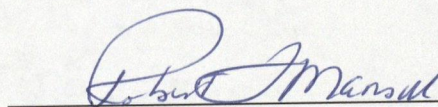
© Denelle Peacey 1995

THE UNIVERSITY OF CALGARY
FACULTY OF GRADUATE STUDIES

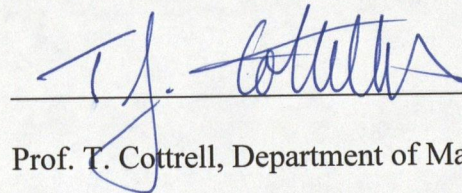
The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Priority Pricing" submitted by Denelle Peacey in partial fulfillment of the requirements for the degree Masters of Arts.



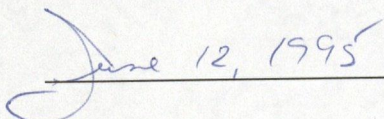
Supervisor, Dr. J. R. Church, Department of Economics



Dr. R. Mansell, Department of Economics



Prof. T. Cottrell, Department of Management



Date

ABSTRACT

Production cannot always be expanded to meet customers' demands. The marginal cost of increasing supply may increase steeply as capacity is exhausted, so that some levels of demand are too large to be fully served in the short-run. When demand or supply is stochastic, the importance of this short-run constraint on capacity must be recognized. Priority pricing enables customers to obtain lower service rates in exchange for greater probabilities of interruption. The voluntary selection of higher interruption levels reveals that their valuation of service is relatively low. Thus, it is more efficient to interrupt these customers than others purchasing higher service options; these contracts enable the firm to substitute low priority demands for expensive capacity additions. As compared with other pricing schemes, priority pricing allows for gains in efficiency over uniform pricing through the closer matching of service options with customers' needs.

ACKNOWLEDGMENTS

I would like to thank Dr. T. M. Horbulyk for his invaluable suggestions during the beginning of this project, and Garth Renne for his proof-reading of the final draft. My thanks is also extended to Dr. Mansell and Prof. Cottrell for their careful proof-reading and insightful comments. I am most grateful to Dr. J. R. Church for his guidance and for the many, many hours spend reading over drafts and checking for mathematical errors.

I am particularly indebted to Paul Mortenson for his support and encouragement during this last year, and much thanks goes out to my family, friends and colleagues for their understanding.

Above all, my thanks goes out to Dan Doll for his unfailing patience, support and loan of his computer, a computer which he feared he may never possess again.

Funding for this thesis was in part provided by Environment Canada's Research Program on Economic Instruments for Achieving Environmental Objectives, the receipt of which is gratefully acknowledged.

DEDICATION

This thesis is dedicated to Dr. J. R. Church, for his unwavering enthusiasm, support and inspiration. His unfailing encouragement made this project possible.

TABLE OF CONTENTS

Approval Page	ii
Abstract	iii
Acknowledgment	iv
Dedication	iii
Table of Contents	v
List of Tables	vii
List of Figures	viii
 CHAPTER ONE: INTRODUCTION.....	 1
1.1 The Problem.....	1
1.2 Background	1
1.3 Outline	6
 CHAPTER TWO: PEAK-LOAD PRICING.....	 7
2.1 Optimal Prices.....	8
2.2 Optimal Capacity.....	11
2.3 A Numerical Example.....	14
 CHAPTER THREE: MODELING STOCHASTIC DEMAND.....	 17
3.1 Introduction	17
3.2 The Brown and Johnson Model	21
3.2.1 Form of Rationing	26
3.2.2 Revenue Constraints.....	29
3.2.3 Rationing Costs	30
3.2.4 Reliability	32
3.3 A Numerical Example.....	36
 CHAPTER FOUR: INTERRUPTIBLE PRICING.....	 43
4.1 Introduction	43
4.2 Tschirhart and Jen	47
4.3 A Numerical Example.....	54
 CHAPTER FIVE: A MODEL OF PRIORITY PRICING.....	 62
5.1 Introduction	62
5.2 Consumer Specifications	65
5.3 Supply Specifications	70
5.3.1 Probabilities of Service.....	71
5.3.1.1 Firm Service.....	71
5.3.1.2 Interruptible Service.....	73
5.4 Social Welfare Maximization.....	76
5.5 Results	78
5.5.1 Pricing.....	78
5.6 Capacity Choice: A Comparison with Uniform Pricing.....	82

5.6.1 Uniform Pricing	83
5.6.2 Priority Pricing	85
5.7 Conclusion	90
CHAPTER SIX: Priority Pricing: Further Discussion.....	92
6.1 Introduction	92
6.2 Woo and Toyama.....	93
6.3 Extensions to Priority Pricing.....	96
6.3.1 Optimal Prices: Two Part Tariffs.....	98
6.3.2 Optimal Capacity	99
6.4 Spot Markets	100
6.5 Insurance	102
6.6 Market Organization.....	104
CHAPTER SEVEN: CONCLUSION	107
BIBLIOGRAPHY	110

LIST OF TABLES

3.1	Brown and Johnson and Extensions: A Numerical Summary	42
4.1	Price/Reliability Elasticities and Priority Orderings	53
4.2	Numerical Summary of Interruptible Pricing and Extensions.....	60
5.1	Numerical Summary of Priority Pricing.....	89

CHAPTER ONE

INTRODUCTION

1.1 The Problem

The demand for many products tends to be variable. For instance, the demand for power may peak briefly and suddenly due to changes in temperature. For many products supply is also subject to fluctuations caused by seasonal cycles or equipment failure. Often supply is nonstorable, so these fluctuations cannot be smoothed with inventories built up in slack periods to satisfy demand in peak periods. In addition, capacity expansions are expensive or otherwise the firm would simply build capacity to satisfy maximum possible demand. Once capacity is chosen, it is set or "sunk" and production cannot be expanded to meet customers' needs because demand varies more quickly than capacity can be adjusted. Thus, during periods of shortage available supply must be rationed among all customers. This rationing imposes costs on customers who are not served and on those who are forced to queue for service. For a social optimum these costs must be traded-off against the cost of capacity expansion. The issue of interest is then how to set price and capacity optimally to maximize total social welfare when there are demand fluctuations.

1.2 Background

There are two main possibilities with respect to demand fluctuations: (1) the timing and extent of fluctuations are known with certainty, and (2) the fluctuations are stochastic or unknown. Although the distribution of stochastic fluctuations is known, it is not known when the fluctuations will occur and how large they will be. Whether demand fluctuations are certain or stochastic will affect the optimal pricing and capacity decisions.

In the peak load pricing literature both peak, or high, and the off-peak, or low, demand is known with certainty and it is known when each level of demand will be realized. For example, it is known that the demand for electricity will peak at 6:00 PM and that demand at 3:00 AM will be significantly lower. In this case, higher prices are charged during peak periods than in off-peak periods in order to ration supply efficiently. Those customers with the highest willingness to pay for service will pay peak prices for service and those with a lower willingness to pay will go unserved. Thus peak-load pricing rations supply in periods of shortages to those customers who value it the most. In off-peak

periods the price is lowered to utilize excess capacity. The differentiation of prices between periods lessens what can be thought of as a congestion externality. Congestion is a familiar phenomenon, especially with roadways. As more and more travelers enter the roadway the length of travel time increases for all travelers. Price differentials help to decrease this congestion since the peak price will dampen the demand in the peak period.

Another question is how to set the optimal capacity. Should there be a large investment in capacity so that demand in the peak period is satisfied while capacity sits idle in the off-peak period, or should there be a smaller investment in capacity such that off-peak demand is satisfied but many consumers are left unserved during peak periods? In the peak-load model the optimal capacity is set in between these two extremes. The optimal capacity level in this model is where the marginal valuation of service, averaged across both peak and off-peak customers, is equal to the marginal cost of capacity expansion. Once the optimal capacity is installed the optimal prices are set such that they clear the market and leave as little capacity idle as possible.

Will these optimal capacity and pricing rules change when demand or supply is stochastic? Both prices and capacity are set *ex ante*, and so they cannot be adjusted to meet demand *ex post*. If demand *ex post* was known with certainty, price and capacity could be set optimally, but *ex ante* there can only be an expectation of demand. While most of the literature discussed here addresses the problem of stochastic demand, stochastic supply poses similar difficulties in the choice of the optimal price and capacity. For consumers it does not matter how congestion was caused. Whether congestion was caused by many travelers wanting to leave the city or by the closure of a roadway, the result is still that travelers will suffer congestion costs.

Given that either demand or supply is stochastic, what is the optimal pricing rule? If price is set in anticipation of off-peak demand and peak demand is realized instead there will be severe excess demand. Conversely, if price is set in anticipation of peak demand and off-peak demand is realized a significant portion of capacity will sit idle. If there is only one period but demand is uncertain, how can prices be used to ration supply?

Fixed or uniform prices lead to idle capacity when demand is low and to the random or arbitrary allocation of available supply during shortages. This random assignment of supply is inefficient because some customers value service more than others. For example, it may not matter as much to the owner of a small store as to a hospital if the supply of electricity is interrupted during a generation shortage. However, while some customers may value service more than others, all customers are served with the same probability of

service under uniform pricing. A more efficient allocation would ration supply during shortages in the order of customers' preferences or valuations so that those customers who value service most highly are served first. The differences in prices in the peak-load model allowed for more efficient rationing because those customers who value service the most in the peak periods pay the peak price and receive service.

If all consumers could be costlessly and efficiently rationed in periods of excess demand, the price mechanism would not be needed to ration supply. The long run uniform price which maximizes total welfare is then the marginal operating cost since this price will both utilize capacity in periods of low demand and maximize consumer surplus in periods of high demand. While this welfare-maximizing price increases the probability that some customers go unserved in periods of excess demand, this surplus loss is mitigated by the gain in social welfare by serving all customers who demand service at a price above the short run opportunity cost of production when demand is low.

If customers can be perfectly rationed during supply shortfalls, then each customer is served with a reliability that is related to their willingness to pay for service. The customer with the highest willingness to pay will always be served with the highest reliability and the customer with the lowest willingness to pay will always be served with the lowest reliability. But if customers cannot be perfectly rationed, how will supply be allocated?

The lesson learnt from the peak-load literature is that supply can be more efficiently allocated by rationing through prices than it can with random rationing. During peak demand periods the good has a higher "quality" in that total demand is greater during these periods and some customers are willing to pay more in order to receive the good. Those customers who value the good relatively more than others will pay higher prices to receive the good during such periods. Non-uniform pricing of interruptible service can be introduced through the pricing of reliabilities of service so that customers pay higher prices for higher reliabilities of service. The price paid is then an index of the reliability or quality of service, rather than the quantity. Each customer's demand is ranked by their willingness to pay for service, essentially establishing a queue. Customers are interrupted during supply shortfalls in the order of this queue until all customers are served or supply is exhausted. Interruptible pricing allows efficiency gains over uniform pricing since service options are designed to more closely match the needs of different customers. For example, if some customers prefer to receive electricity with a reliability of 100% but are subject to random outages due to uniform pricing, when it is offered these customers will gladly pay

higher prices for a higher level of reliability. In addition, because some customers are served at a lower level of reliability, the reliability of higher priority customers can be satisfied at less than the cost of additional capacity. Thus lower priority demand substitutes for capacity expansions while still maintaining high reliability of service for higher ranked demand.

Preferences or tastes for service reliability are often private information however, known only to customers. While the firm may know the distribution of preferences among all customers, it cannot know the individual preference of each customer. It cannot simply ask customers for their preferences. Since customers value reliability of service, depending on the price and reliability offered they might have an incentive to overstate or disguise their preferences if they think that this will increase their probability of service. For example, if residential customers of telephone utilities receive lower rates than do businesses for essentially the same quality of service,¹ then businesses will want to masquerade as residential customers in order to receive this lower price. Low priced, high quality commodities can only be offered if customers cannot hide their preferences and if resale of the commodity among customers is not possible. Unless there is perfect information this will not be possible and higher prices will always be charged for higher reliabilities.

Priority pricing accommodates asymmetric information through self-selection by customers of service options. Whenever customer preferences are diverse, the offering of a service menu with several options promotes efficiency gains because it enables customers to adapt their purchases to their preferences. Customers will self-select among the offered options to choose the option which is best for them. Such self-selection is different from demand management in that customers are not targeted for service options. All customers are instead presented with a menu of service options from which they make their selection. Customers' voluntary selections of lower reliabilities reveals that their valuation of service is relatively low, and so it is more efficient to interrupt these customers than others electing high reliability. In addition, this self-selection will reveal the valuation of additional capacity. Priority pricing allows for substantial efficiency gains over uniform pricing through the closer matching of preferences with service options but the self-selection of service options lessens the informational requirement needed to construct the price menu. As we will see in the model developed later in this paper, the implementation of priority pricing results in a more efficient rationing of supply and a greater level of social welfare over uniform pricing and random rationing.

¹This is due primarily to price discrimination.

Models of priority pricing can be designed such that customers self-select from the offered menu of service options to either maximize their consumer surplus or minimize their expenditures. In Chapter 5 we will explicitly develop a simple model of priority pricing which uses this second approach. Two service classes are offered in this model, firm and interruptible, where firm customers have a higher probability of service than interruptible customers. Customers will choose between these two classes to match their preferences. The marginal customer is the customer who is indifferent between the two classes, and it is this marginal customer who determines the probability of service for all customers. Total consumer surplus in this model is defined as the difference between total service expenditures under uniform pricing and under priority pricing, given that customers must maintain their original level of utility if they are interrupted. The results of the model clearly show that there are welfare gains in offering priority service. Furthermore, we construct parallel models to demonstrate the effect on capacity choice when interruptible service is offered.

The model of priority pricing developed in Chapter 5 can be compared and contrasted against other models of interruptible and priority service in the literature. While the model of priority pricing developed here offers a menu of only two service classes, it is possible to extend it through the offering of a continuum of classes. Given a complete continuum of service classes, every customer will be offered the price and reliability which exactly matches their preferences. While this continuum maximizes total surplus through perfect differentiation of service, the majority of the welfare gains under priority service can be realized with only a few classes of service.

In general, the implementation of priority pricing lessens the need to expand capacity. Priority pricing allows the substitution of low-priority demands having relatively low value to customers for expensive additions to capacity that would otherwise be required to sustain reliable service. The contracts for low-priority service essentially frees up capacity to meet shortfalls, thereby protecting the higher reliability expected from high-priority contracts.

In the peak-load model, the optimal capacity is set where the consumers' willingness to pay for capacity, averaged over all periods, covers the cost per period of capacity expansion. The optimal capacity is more difficult to find when there is asymmetric information. However, the implementation of priority pricing will reveal the marginal valuation of capacity, thereby allowing the optimal capacity to be selected even in a world of uncertainty. Because the price schedule is designed to conform to the distribution of

customer preferences, the choice of a price option by a customer will reveal not only the preference of that customer, but their valuation of a capacity expansion as well. If many customers elect to choose a high priority tariff, this will reveal that valuation of service may be such that capacity should be expanded. Ultimately, capacity will optimally be built at the level where the marginal contribution to social welfare of an additional unit of capacity justifies the marginal cost of capacity.

In a world of uncertainty one method of addressing multiple futures is to hold spot markets in all contingencies. Spot markets are efficient because customers will bid for the available supply, thereby rationing supply to those willing to pay the highest price. However, spot pricing is impractical in many markets due to the transaction costs involved, the infrastructure needed, and the ability of customers to react to rapid price changes.

Priority prices are the expectations of spot prices for comparable service. Thus, priority pricing can be viewed as forward contracts for supply. Given observable demand or supply shocks, the selection of a priority service or contract yields a probability of service for all possible states of the world. However, one limitation of the priority pricing model is that it is assumed that customers are risk-neutral. If customers are instead risk-averse, supplementary insurance should be offered to compensate consumers for interruption for full efficiency.

1.3 Outline

In the analysis below we first trace the origins of priority pricing from peak load pricing as summarized briefly in Chapter 2 to the introduction of stochastic demand described in Chapter 3. The model discussed in this chapter assumes perfectly efficient rationing and uniform prices. Interruptible pricing for service reliabilities is outlined in Chapter 4, while a simple model of priority service is detailed in Chapter 5. In Chapter 6 this model is then compared against other models of priority pricing and further extensions of priority pricing are also discussed. The results of the analysis are then briefly summarized in Chapter 7 and avenues for implementation and further research are discussed.

CHAPTER TWO

PEAK-LOAD PRICING

2.1 Introduction

The demand for many products is not uniform but, rather, is stochastic or cyclical. Cyclical demand moves through periods of peak, or high, and off-peak or low demand and is typical of facilities such as roadways; we are all familiar with the phenomenon of rush hour. It is also typical of public utilities such as water, telephone service, electricity and natural gas. The demand for electricity tends to be higher at 6:00 PM as consumers arrive home from work than at 3:00 AM when most consumers are asleep. Congestion results from constrained capacity. Once installed, capacity is set or "sunk" and cannot be adjusted to match fluctuations in demand. Often, such commodities are non-storable, either due to technological limitations or to high storage costs, and so production cannot be smoothed by the accumulation of inventories during periods of low demand. For such commodities production is initiated when it is demanded.

This naturally brings us to the question of optimal pricing and capacity. On the one hand, if capacity is very expensive it may be difficult to justify the installation of enough capacity to meet peak demand in the peak period if this capacity then sits idle in the off-peak period. On the other hand, it is also important to reduce the amount of congestion due to insufficient capacity. Thus, the costs of excess capacity must be traded off against the cost of congestion, and when there is congestion, the commodity must somehow be rationed. One way to ration supply is through price. Raising the price in the peak period reduces the amount of excess demand and also ensures that the commodity is allocated to those customers who value it the most since those customers will be willing to pay the higher price to receive the good. In off-peak periods the price will be lower in order to use as much capacity as possible.

Notice that with peak-load pricing both the timing and the magnitude of peak and off-peak demand are known with certainty. Knowing this, the optimal prices for each period can be set *ex ante* for the demand realized *ex post*.

In the peak-load model we briefly summarize here it is assumed for simplicity that there are only two periods of equal length;² the peak or high demand period, and the off-

²The model can be generalized, however, to any number of periods of any length.

peak or low demand period. It is assumed that demand in each period is independent of prices and demand in the other period.³ ⁴ We further assume a uniform technology with a Leontief production process so that constant marginal operating and capacity costs are incurred for each unit of production supplied.

2.2 Optimal Prices

Peak-load pricing is a form of price discrimination or non-uniform pricing; a different price is charged to the peak consumers than to the off-peak consumers. In general, prices are adjusted in each period to just clear the market of any excess demand or supply while covering variable costs. In peak periods the price is raised to clear the market of any excess demand. The increase in price causes fewer customers to demand the commodity and so there will be less consumer surplus lost due to rationing. In the off-peak period the price is lowered to clear the market of any excess supply. Since capacity is sunk social welfare can be increased by using as much of the available capacity as possible while still covering operating costs.

In Figure 2.1 below capacity is installed at Q^* . Assuming that there are only two periods, peak demand is denoted by D_p and off-peak demand is D_o . The short run marginal cost, c , is the marginal operating or variable cost. This is the cost of producing one more unit of supply. The long run marginal cost is $(c + \beta)$, where the marginal capacity cost per period is β . Assuming that capacity is perfectly divisible and not "lumpy", the marginal capacity cost is the cost of increasing capacity by one unit. The long run marginal cost is the cost of production across both periods: the marginal operating cost, plus the cost of capacity. Both the marginal operating and capacity costs are assumed to be constant.⁵ The willingness to pay for service above the marginal operating cost, c ,

³We are assuming that the cross-elasticities of demand between periods are zero. If the cross-elasticities of demand between periods were non-zero, the peak-load pricing structure would shift some consumer demand from the peak to the off-peak period, making a smaller investment in capacity more viable. For a discussion of this and other issues related to peak-load pricing, see Berg and Tschirhart (1988), Chapter 5.

⁴ This is an unrealistic assumption if peak and off-peak services are close substitutes. With interdependent demand the change in marginal revenue from a change in peak price would consist of a change in total revenue in both the peak and the off-peak periods. With interdependent demand the optimization conditions remain in essence the same although the individual expressions are somewhat more complex. See Burness and Patrick (1991).

⁵Marginal costs are assumed to be constant such that any two customers purchasing the same quantity impose the same costs on the firm. Constant costs is a simplifying but not a necessary assumption. Economies of scale will, however, affect the costs of capacity since production costs will fall as production

averaged across both demands is D_{apo} . This is the average of the vertical summation of the two demand curves above the operating cost. For the demand curves as drawn the optimal peak price is P_p , and the optimal off-peak price is P_o .

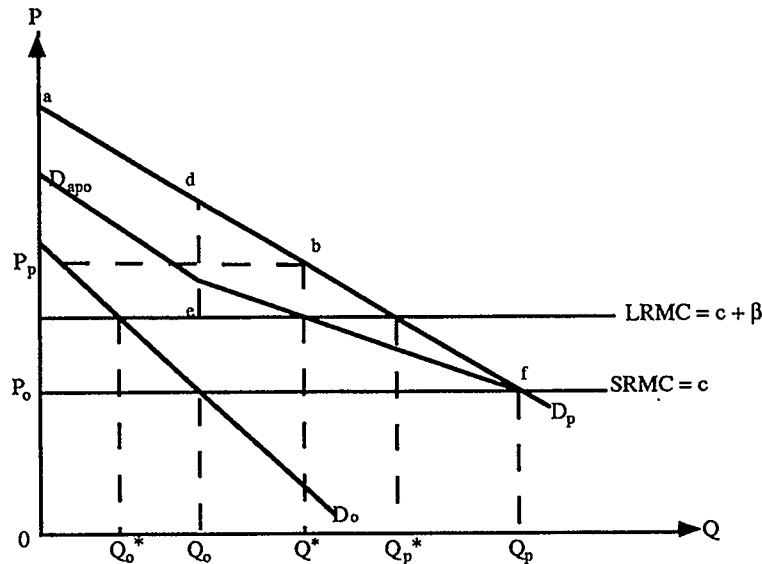


Figure 2.1
Peak and Off-Peak Prices

To clear the market of excess capacity in the off-peak period the optimal price is set to the marginal operating cost, c . At this price variable costs are covered while the use of capacity in the off-peak period is maximized. Thus, given this capacity level, the optimal off-peak price best utilizes installed capacity while covering variable costs of production.

If the optimal off-peak price was charged in the peak period there would be excess demand of $(Q_p - Q^*)$. The available capacity, Q^* , would then have to be rationed among all customers demanding service at this price, customers along \overline{af} . To reduce this excess demand the peak price must be raised to P_p to exactly clear the market of excess demand in the peak period. At this optimal peak price customers from \overline{ab} will demand service. Notice that customers with willingness to pay for service below P_p , those customers along

increases. With economies of scale the social planner will install a higher level of capacity and charge lower prices in both periods in order to utilize this higher capacity.

\overline{bf} , are not served. The price increase ensures that the commodity will be rationed to those customers who value it the most.

In summary, given any capacity, there are two pricing scenarios: (1) demand at the price c is less than capacity and a price of c is charged, and (2) demand at the price c is greater than capacity and the price is raised above c to clear the market.

This second scenario for the off-peak customers is illustrated by Figure 2.2 below. With the installed capacity Q^* and a price of c there will be excess demand in the off-peak period of $(Q^* - Q_o)$.

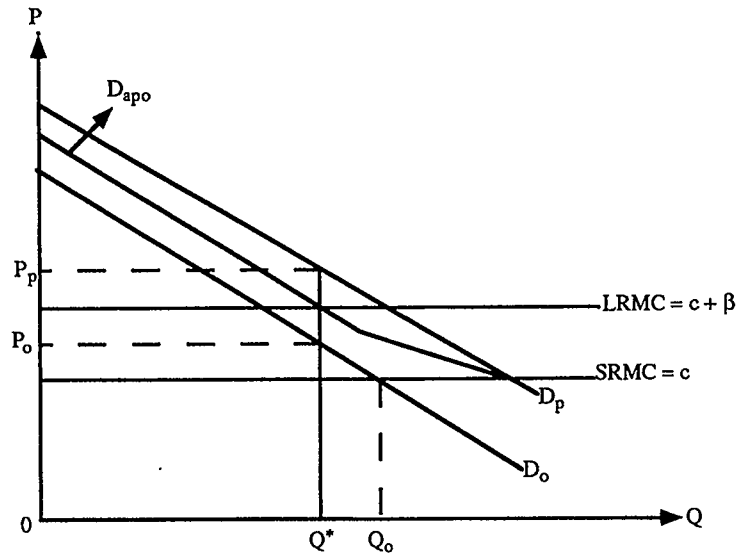


Figure 2.2
Peak and Off-Peak Prices: Case Two

The optimal off-peak price is P_o , which exactly clears the market of excess demand. The optimal price for the peak period is still P_p , since this price clears the market in the peak period.

It is generally assumed in the peak-load model that a uniform production technology is employed so that the same production process is used for all levels of production. With diverse technologies, so that different production technologies are employed for different levels of demand, the possibility that all users contribute to capacity costs increases.⁶

⁶Berg and Tschirhart (1988), p. 175.

In summary, analysis of the peak-load model demonstrates the welfare gains that can be had when non-uniform prices supplant uniform prices. Capacity is more fully utilized in the off-peak periods with the lower off-peak price and the consumer surplus lost due to rationing because supply is insufficient to meet demand in the peak periods is eliminated with the higher peak price.

For any level of installed capacity the optimal prices can be found in all demand periods. Prices will be set equal to the marginal operating cost to utilize available capacity unless this price results in excess demand, in which case the price is raised to just clear the market. The question then is how to set the optimal capacity.

2.3 Optimal Capacity

Should capacity be added so as to meet peak demand with significant excess capacity in off-peak periods or should capacity be invested in so as only to meet off-peak demand, allowing a large portion of peak demand to go unsatisfied? As we will see, the optimal capacity lies somewhere in between these two extremes so that the cost of excess capacity in the off-peak period will be traded off against the costs of excess demand in the peak period.

Once installed, capacity or supply is fixed, and the costs of capacity must be paid regardless of it is fully used or not. Capacity costs are essentially sunk and not recoverable after capacity is installed. Although the optimal capacity and prices are chosen simultaneously, we can think of capacity and prices as being chosen sequentially, such that the optimal capacity is chosen first, and then optimal prices are chosen for this capacity. If we think of the choice of the optimal prices and capacity as a sequential process, we can break the problem down into two parts: first, the optimal prices are chosen for any capacity, and then the optimal capacity level is selected.

The marginal valuation of capacity is the difference between the willingness to pay for service and the short run marginal cost for a unit of output. Capacity is optimal if the amount that consumers are willing to pay for an extra unit of capacity exactly equals the marginal cost of providing this unit. Capacity should be provided as long as consumers are willing to pay more for an extra unit of capacity than the cost. Conversely, capacity should be reduced if the amount that consumers are willing to pay for an extra unit is less than the cost. Referring to Figure 2.1, we can see that if the market consisted only of off-peak demand, the optimal capacity would be Q_o^* because the willingness to pay for capacity is

exactly equal to the costs of production in the long run, the sum of operating and capacity costs. Conversely, the optimal level of capacity would be Q_p^* if the market were characterized only by peak demand.

What would be the optimal capacity if peak users are willing to pay for capacity expansion while off-peak consumers are not? The optimal capacity level is where consumers' willingness to pay for capacity, averaged over all periods, exactly covers the cost per period of capacity expansion. Thus, it is the average marginal willingness to pay for extra capacity that determines the optimal capacity. Even if the valuation of off-peak consumers is not sufficient to cover expansion costs extra capacity might still be desirable if peak consumers value extra capacity sufficiently more than the cost, thereby compensating for the lower valuation of off-peak customers. For extra capacity to be warranted peak customers must value extra capacity sufficiently to pay for it in both the peak and in the off-peak periods where it sits idle and provides no benefits.⁷

The average willingness to pay per period is represented in Figure 2.1 by D_{apo} . At capacity levels beyond Q_o off-peak consumers are not willing to pay for capacity expansion: however, peak consumers are willing to pay \overline{ab} of consumer surplus above the long run marginal cost, thereby justifying a capacity expansion beyond Q_o . Similarly, beyond Q_p peak customers are not willing to pay for added capacity. The optimal capacity level is then Q^* , where the average willingness to pay across both periods is exactly equal to the long run marginal cost of capacity. It is implicitly assumed that capacity is divisible. If capacity were indivisible or "lumpy" and could therefore only be increased by large, discrete amounts, the optimal capacity Q^* might be unattainable.

We can easily see the efficiency gains in terms of total welfare from setting capacity optimally. In Figure 2.3, if price is set to the long run marginal cost (LRMC) and capacity is set at Q_o , total surplus is maximized only if off-peak demand, D_o , is realized. With peak demand there will be a loss of consumer surplus of the shaded triangle abf at this price and capacity. This loss is the opportunity cost borne by unserved consumers due to the rationing of limited supply.

⁷See Train (1991) for a succinct discussion of these issues.

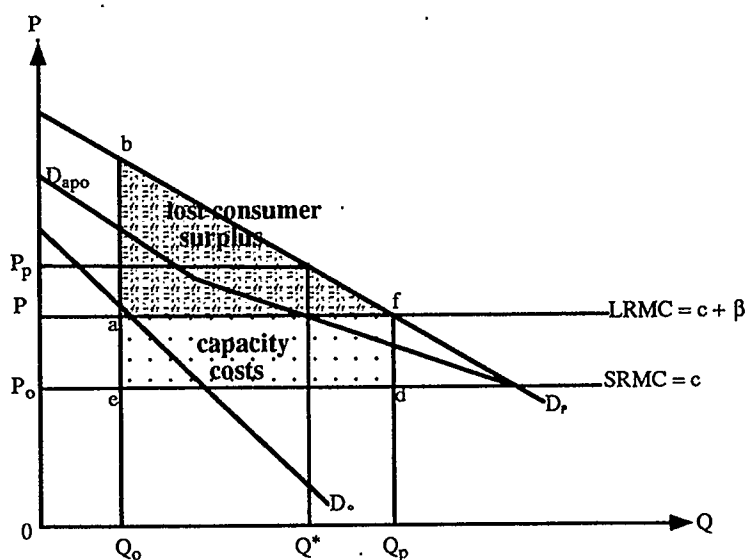


Figure 2.3
Efficiency Losses with Non-Optimal Capacity

Conversely, with a price of LRMC and a capacity of Q_p , total surplus is maximized only if peak demand is realized. If off-peak demand is realized instead, profits will be negative since total revenues, area PaQ_o0 , are insufficient to cover the cost of capacity, the rectangle $afde$.⁸ Notice that even with a price of P_o there will be negative profits with off-peak demand.

The optimal capacity is Q^* since this is the level of capacity at which the trade-off between the loss in consumer surplus due to rationing and capacity costs is minimized. At the optimal level of capacity the average willingness to pay for capacity, evaluated at the margin, is just equal to the marginal cost of capacity. At Q^* peak consumers are willing to pay P_p for capacity expansion while off-peak consumers are not willing to pay anything for capacity expansion. Since the average willingness to pay for capacity is equal to the marginal cost of capacity at the optimal level this means that the willingness to pay for capacity by peak users, $(P_p - c)$, is high enough to compensate for the zero valuation of off-peak users.⁹

⁸From the diagram, we can see that total revenue are less than total cost; $0Q_oaP < 0Q_oaP + afde$.

⁹In this example the willingness to pay for capacity at Q^* by off-peak users is zero. This implies that the valuation of capacity by peak users is $(2\beta + c)$, since the average valuation of capacity is $\{(P_p - c) + (P_o - c)\}/2$, which is then set to the marginal capacity cost, β .

If off-peak users are charged the short run marginal cost, only peak users will contribute to capacity costs. Notice, however, that in Figure 2.2 customers in both periods contribute to capacity costs. Peak customers in Figure (2.2) contribute relatively less to capacity costs than in Figure 2.1 because off-peak customers are contributing relatively more.¹⁰ As the difference in the willingness to pay in two periods converges there is less opportunity for discrimination between periods.¹¹

2.4 A Numerical Example

We can illustrate the optimal pricing and capacity rules in the peak-load model with a simple example. The objective function for this welfare-maximizing peak-load pricing problem is constructed as:

$$W = \int_0^{Q_p} (30 - Q_p) dQ + \int_0^{Q_o} (20 - Q_o) dQ - c(Q_p + Q_o) - 2\beta \cdot Q^* \quad (2.1)$$

where W denotes total social welfare, c and β are the marginal operating and capacity costs and P_p, P_o, Q_p, Q_o are the peak and off-peak prices and quantities as denoted above.

Consumer surplus is the sum of the first two terms on the right less total costs, the last two terms. The third term is the total operating costs and the last term is cost of capacity at the optimal level Q^* for both periods. Figure 2.4 below graphically depicts these areas for the demand curves as drawn in Figure 2.1.

¹⁰With a shifting peak situation, the firm will adjust prices such that at the optimum both peak and off-peak consumers will contribute to capacity costs. A shifting peak is when prices cause the off-peak quantity demanded to exceed the peak quantity demanded. To avoid this, the off-peak price is raised, and the peak price is lowered by equivalent amounts until demand in both periods equals capacity. For a lucid discussion, again see Berg and Tschirhart (1988), Chapter Five.

¹¹Assuming that the periods are of equal length. With periods of unequal length, weights would have to be assigned to each period, proportional to the length of the period, in order to calculate average revenue.

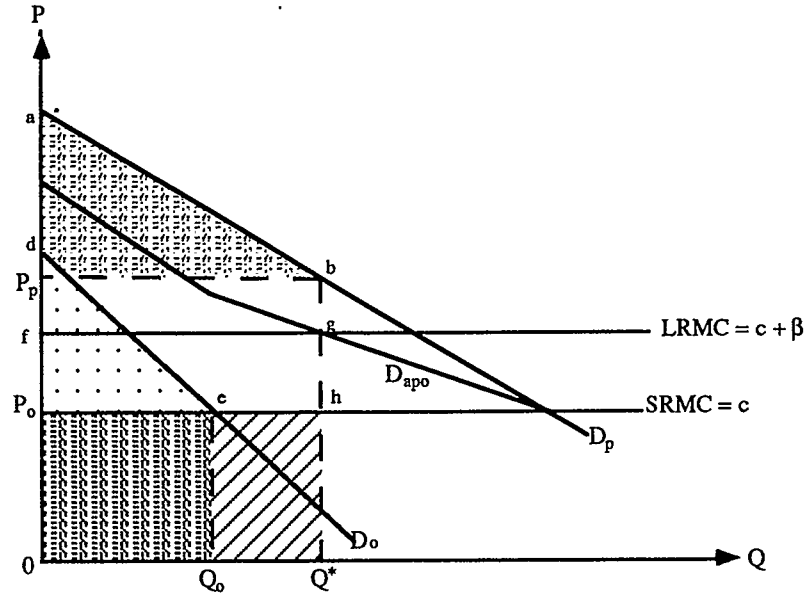


Figure 2.4
Peak-Load Pricing: Total Surplus

For the demand curves as drawn, the consumer surplus for peak customers with a price of P_p and a capacity of Q^* is the triangle abP_p .¹² The consumer surplus for off-peak customers, who are charged a price of P_o , is the triangle deP_o . Operating costs in the off-peak period is the rectangle $P_o e Q_o 0$, while operating costs in the peak period is the larger rectangle $P_o h Q^* 0$. The fixed capacity costs per period are the rectangle $fghP_o$.

We maximize the objective function subject to the feasibility constraint that, in both periods, demand does not exceed installed capacity, Q^* :

$$\begin{aligned} Q_p &\leq Q^* \\ Q_o &\leq Q^* \end{aligned} \tag{2.2}$$

If we assume that marginal operating and capacity costs are;

$$c = 2 \tag{2.3}$$

$$\beta = 4 \tag{2.4}$$

¹²For peak demand as drawn in Figure 2.4 $Q_p = Q^*$.

the optimal prices and capacity are;

$$p_p = 10 \quad (2.5)$$

$$p_o = 2 \quad (2.6)$$

$$Q^* = 20 \quad (2.7)$$

a solution which yields a total surplus of 598 and zero profits, since profits are defined as,

$$\pi = [P_p - c]Q_p(P_p) + [P_o - c]Q_o(P_o) - 2\beta \cdot Q^* \quad (2.8)$$

The optimal capacity is where the valuation of capacity above operating costs is exactly equal to the long run marginal cost. At capacity levels beyond 18, off-peak customers are not willing to pay for additional capacity. At capacity of 20, peak customers are willing to pay exactly the cost of capacity, $(c + 2\beta = 10)$, to cover the cost of capacity in both periods, since

$(Q_p = 30 - P_p = 30 - 10)$ is the willingness to pay of peak customers.

Notice that the optimal peak price is higher than the off-peak price, as expected. The optimal off-peak price is exactly equal to the marginal operating cost because at this price no rationing need occur. Since capacity costs are sunk the optimal price is set at the operating cost to create as much consumer surplus as possible. The optimal peak price is exactly equal to the sum of the long run capacity cost, $(2\beta + c)$. This ensures that there is no excess demand during the peak period.

In the peak load pricing literature it is assumed that both the timing and the magnitude of demand is known in each period. However for many commodities demand is stochastic. While the distribution of possible demand may be known, it is not known what level of demand will be realized at any one time. Because demand is unknown the probability that the commodity will have to be rationed in any period increases. While we have seen in the peak load literature the role that pricing plays in rationing available supply efficiently, the optimal setting of prices becomes more problematic when the optimal price must be set before demand is realized, while the realization of this demand is not known. In the next chapter we will address the question of how to set the optimal level of capacity and prices when demand is stochastic or random.

CHAPTER THREE

MODELING STOCHASTIC DEMAND

3.1 Introduction

Stochastic demand means that demand is uncertain or subject to some random element. This stochastic element could be changes in preferences, incomes, prices of other commodities, or weather. For example, demand for gas is high when the weather is cold but it is not certain exactly when the weather will be cold. This random element causes stochastic demand such that there can only be an expectation about the range of possible demand and the magnitude of actual demand will not be known until it is realized.¹³ Much like the peak load problem capacity and price must be set in advance. Because both capacity and price are fixed before demand is realized, supply cannot be adjusted to match demand since supply is non-storable. Thus, supply must be rationed during shortfalls, and there is the risk that some customers will go unserved or that there will be excess capacity.

We can easily see how stochastic demand complicates the socially optimal capacity and pricing decision. In Figure 3.1 capacity is fixed at K and the price is set at the long run marginal cost, $LRMC$, the sum of the marginal operating cost, c , and the marginal capacity cost, β . The short run marginal cost, $SRMC$, is the marginal operating cost. Both capacity and price are set *ex ante* before demand is realized.

¹³In terms of the actual specification of demand in the literature, this random element may enter demand in a general form, $D(p, \mu)$, or in a specific form. The most common specific forms are additive, $(D(p) + \mu)$, or multiplicative, $\mu D(p)$. Alternatively, this random element could also enter in the availability of supply.

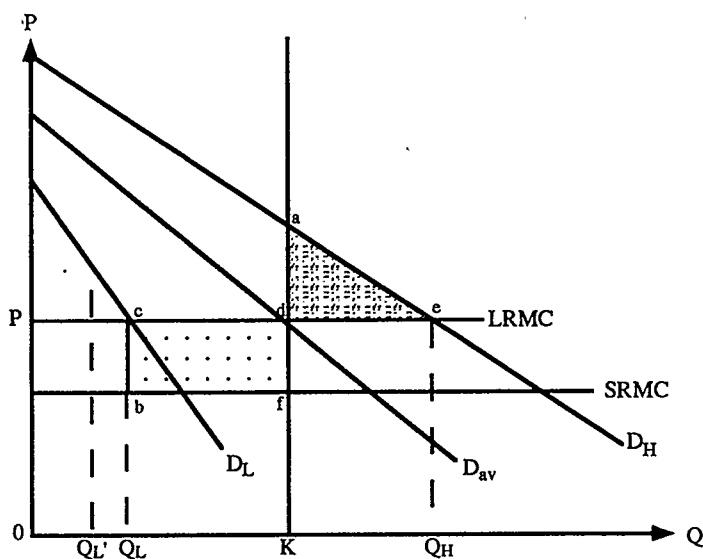


Figure 3.1
Stochastic Demand

If low demand is realized, D_L , capacity costs of the shaded rectangle $cdfb$ will not be covered by revenues because demand will be less than capacity. If high demand is realized instead there will be a loss of welfare of the shaded triangle aed due to rationing, assuming that rationing is efficient. This is the loss in consumer surplus because customers demanding units $(Q_H - K)$ go unserved. There is no loss of consumer surplus when low demand is realized because all customers demanding supply at the price P are served. In choosing the welfare-maximizing optimal capacity when demand is stochastic the costs of excess capacity must be traded-off against the costs of excess demand.

The loss of consumer surplus due to rationing can be reduced by increasing capacity. In Figure 3.2 with a price P and with capacity set at Q_H there will be no loss of consumer surplus when high demand is realized. However there will be large capacity costs of the rectangle $abcd$ which must be paid regardless if sales of Q_H or Q_L are realized. If capacity is set at Q_L capacity costs will always be covered by revenues but there is an opportunity cost of the area aeb if high demand is realized since this is the consumer surplus lost due to rationing with the lower capacity and demand D_H .

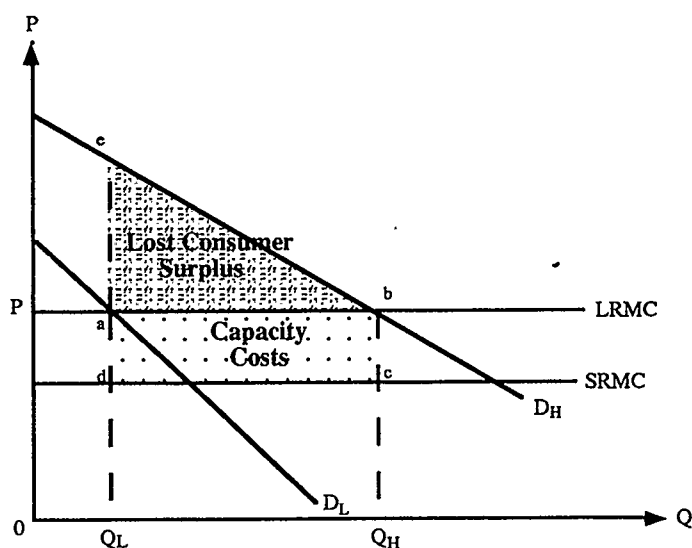


Figure 3.2
Stochastic Demand: Capacity

In setting the socially optimal capacity there are two conflicting incentives. In order to decrease the probability of excess demand and the loss of consumer surplus due to rationing there is the incentive to increase the optimal capacity level. There is also the conflicting incentive to decrease the capacity level in order to reduce the probability of excess capacity and excessive capacity costs.

The costs of rationing depend on the form of rationing assumed. Efficient rationing is where, in periods of excess demand, the consumer with the highest willingness to pay is served first, after which the consumer with the next highest willingness to pay is served, and so on until available capacity is exhausted or all customers are served. With any given capacity and price, efficient rationing allocates resources more efficiently than other forms of rationing because the resource is allocated to those consumers who value it the most. In contrast, the most inefficient form of rationing is where supply is allocated to those consumers who value it the least.

In Figure 3.3 below, supply is rationed first to those customers who value it the most, while in Figure 3.4 available supply is rationed first to those customers who value it the least.

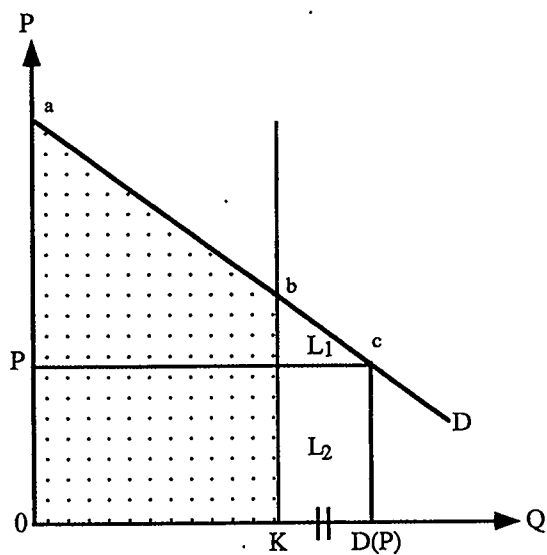


Figure 3.3
Efficient Rationing

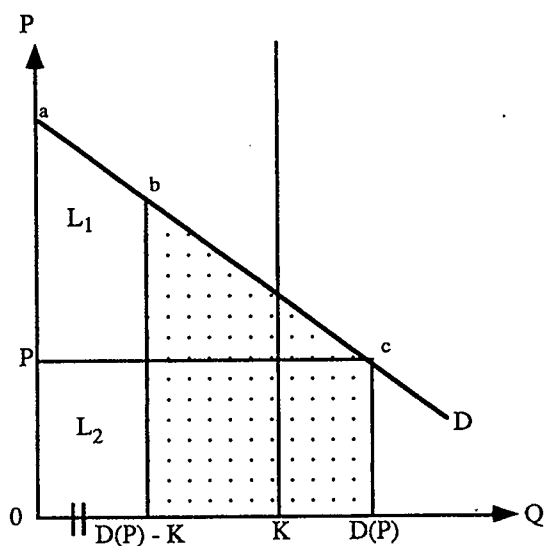


Figure 3.4
Inefficient Rationing

In both Figure 3.3 and 3.4 capacity is fixed at K and P is the price charged to all customers. Capacity must be rationed because at the given capacity level and price because supply is insufficient to fully serve demand of $D(P)$ at the price P and there is excess demand of $(D(P) - K)$. With efficient rationing the commodity is allocated to the customers

from \overline{ab} in Figure 3.3. These customers have the highest willingness to pay for service because they receive the highest level of consumer surplus from consumption. Customers along the segment \overline{bc} go unserved. In Figure 3.4 the available supply to be rationed is the same as in Figure 3.3. With inefficient rationing available supply is allocated first to those customers with the least willingness to pay, those customers along the segment \overline{bc} on the demand curve. The customers with the highest willingness to pay, the customers along \overline{ab} , are unserved.

In Figures 3.3 and 3.4, L_1 is the loss in consumer surplus due to rationing and L_2 is the loss in total revenues. As we can see, the magnitude of the loss of consumer surplus due to rationing and amount of total welfare depends on the form of rationing used. With inefficient rationing the loss of consumer surplus due to rationing is much higher than with efficient rationing because those customers with the highest valuation of service do not receive service.

Other forms of rationing include random rationing, queuing, and pro-rated rationing. A random rationing scheme allocates available supply at random among all those customers willing to buy the product at the quoted price. With random rationing the order of service and the willingness to pay are completely uncorrelated. Queuing (or head-of-the-line rationing) allocates supply in the order of an established queue or order. This type of rationing may converge with efficient rationing if those consumers with the highest willingness to pay are also those consumers with the lowest time opportunity costs, such that it is less costly for them to queue. A pro-rated rationing scheme distributes supply proportionately to all consumers. With this scheme as well, the willingness to pay of consumers is uncorrelated to the order of rationing. Because these other forms of rationing do not necessarily allocate supply to those consumers who value it the most, these forms of rationing are inefficient.

3.2 The Brown and Johnson Model

The stochastic pricing literature begins with Brown and Johnson's 1969 extension of the peak-load model. Risk is incorporated into the peak-load model with the addition of stochastic demand. Capacity and the corresponding prices are set *ex ante*, before demand is realized, and so there exists the risk that some consumers will not be served or that there will be excess capacity.

The welfare-maximizing social planner in this model seeks to maximize total surplus, the sum of (expected) consumer surplus and (expected) profits or producer surplus. During supply shortages the social planner, endowed with full information, rations available supply among consumers. In their model, Brown and Johnson derive the result that the optimal uniform price in a model with stochastic demand is a price equal to the marginal operating cost;¹⁴ this result is surprising since one would expect the results of the peak-load model to still apply when stochastic demand is added.

However, for any given choice of capacity, this is the price that maximizes consumer surplus. When low demand is realized this price minimizes excess capacity, as illustrated in Figure 3.5.

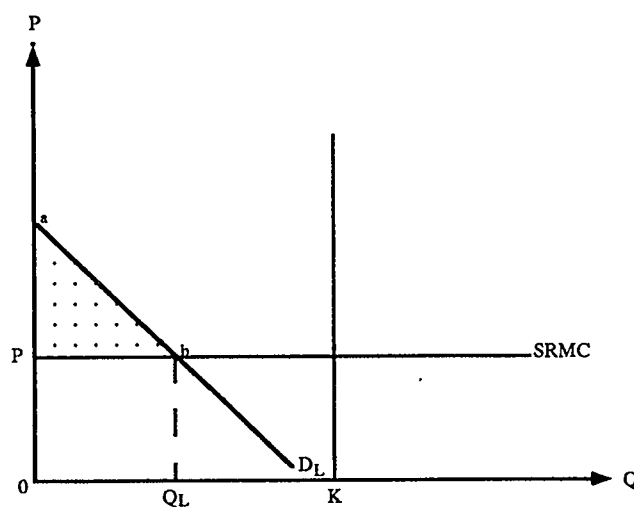


Figure 3.5
Brown and Johnson: Excess Capacity

With a price equal to the short run marginal cost (SRMC), the largest consumer surplus is realized, (area *abc*). With any price higher than *P* this consumer surplus will be less and more capacity will lie idle because fewer customers will demand service. When there is

¹⁴Brown and Johnson examine two forms of demand uncertainty, multiplicative and additive, but they find no substantial difference in the optima between the two functional forms in their welfare-maximizing framework. Carlton (1977) later shows that this similarity in results is due to the form of rationing that is assumed.

efficient rationing, as long as the willingness to pay is greater than the marginal operating cost, consumer surplus will fall as the price is raised.

When high demand is realized consumer surplus is also maximized at this price because with any higher price there is simply a transfer between consumer surplus and producer surplus. In Figure 3.6, with the price set at the short run marginal cost, consumer surplus is the shaded area $dbcP'$.

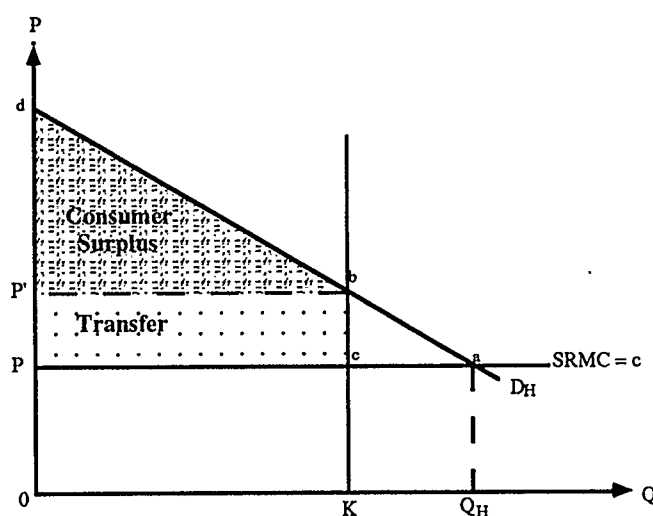


Figure 3.6
Brown and Johnson: Excess Demand

When the price increases to P' the area $P'bcP$ is transformed from consumer surplus to producer surplus and total surplus is unchanged. Thus consumers are just as well off with a price of P as with any price between P and P' because the price does not affect total surplus, and they are better off than with any price greater than P' . With capacity constrained at K consumer surplus will fall with any price above P' because at any price above P' the quantity demanded will be less than capacity. Thus, for any level of capacity less than Q_H , total surplus will be maximized both when there is high demand and when there is low demand by charging a price $P = c$. Because capacity is set *ex ante* capacity costs are essentially sunk when demand is realized. If capacity costs are sunk and unrecoverable the optimal (social welfare-maximizing) pricing policy then is to charge

prices equal to marginal operating costs since this maximizes consumer surplus while utilizing as much available capacity as possible.

With any given capacity level customers can never be worse off than with a price equal to the marginal operating cost because there is no rationing problem in the Brown and Johnson model. While the optimal price in their model introduces the risk of a large loss of consumer surplus during periods of excess demand, this cost is minimized because when capacity is constrained consumers are somehow perfectly rationed by an omniscient social planner in the order of their willingness to pay for service. Because the social planner can costlessly and efficiently ration all customers, the social planner does not need the price mechanism to allocate available supply.

Ideally, capacity would be installed at a level where there would never be excess demand and rationing. However, because capacity is expensive the possible costs of lost consumer surplus due to rationing must be traded-off against the cost of installing an additional unit of capacity.

In the Brown and Johnson model, the optimal capacity is where the average marginal expected willingness to pay for capacity is equal to the marginal capacity cost or the cost of expanding capacity by one unit.¹⁵ This is the same level of optimal capacity as in the peak load model if we assume two levels of demand with an equally likely chance of occurrence. In Figure 3.7 the optimal capacity is shown at Q^* .

¹⁵"...optimal capacity should be chosen such that marginal capacity cost is equal to the truncated mean of the difference between the willingness to pay and the actual price for the marginal disappointed purchaser of the commodity." Brown and Johnson (1969), p. 123.

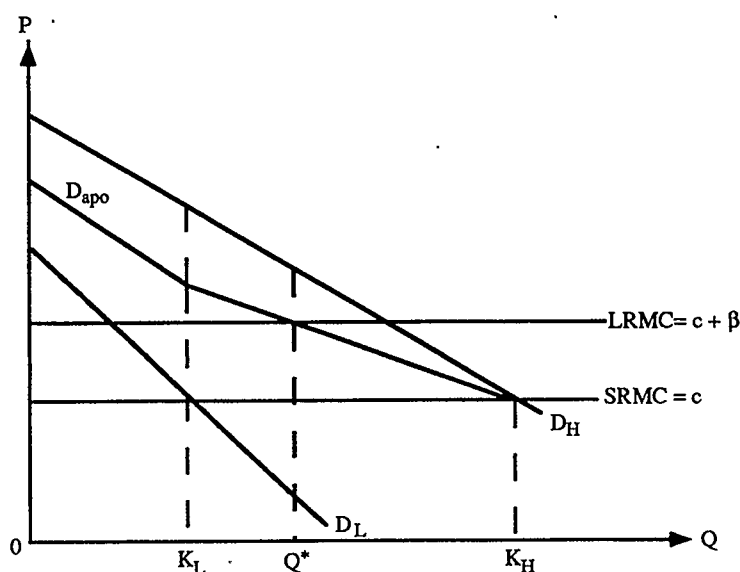


Figure 3.7
Optimal Capacity in the Brown and Johnson Model

In Figure 3.7 we assume two possible states of demand, high and low demand, both with an equal probability of occurrence. The optimal capacity is where the willingness to pay for capacity averaged across all demands is exactly equal to the capacity cost, β . At capacity levels beyond K_L the low demanders, (D_L), are not willing to pay for capacity expansions since their marginal valuation of capacity, the difference between the willingness to pay and the marginal operating cost, c , is zero. Similarly, high demanders, (D_H), are not willing to pay for capacity expansions beyond K_H . In this example it is the high demand customers who determine capacity since beyond K_L low demanders are not willing to pay for capacity expansions. Thus, here the marginal willingness to pay for capacity by high demanders at the optimal capacity level is exactly equal to the cost of capacity expansion. Given that both demands have an equal chance of occurrence, this implies that the willingness to pay for capacity by the high demanders is (2β) . This capacity level optimally trades off the cost of an increase in capacity against the loss of consumer surplus due to rationing as measured by the willingness to pay for service. Like the peak load model, the optimal capacity is determined by the valuations of all consumers who might use it.

Brown and Johnson constructed their model within a welfare-maximizing framework and their optimal capacity and prices are those which maximize total surplus. However, this solution is dependent on various assumptions, including;

- (1) the form of rationing
- (2) revenue constraints
- (3) rationing costs
- (4) reliability

3.2.1 Form of Rationing

The most important assumption of the Brown and Johnson model is that of efficient and costless rationing. While the peak load model uses price to ration available supply among all customers demanding service, Brown and Johnson assume that efficient rationing can occur without the price mechanism. They assume the existence of a social planner who knows the willingness to pay of each customer and can costlessly allocate supply to those customers with the highest willingness to pay. Because the social planner in the Brown and Johnson model does not need price in order to ration supply, the price is chosen to utilize as much capacity as possible.

A price of short run marginal cost is welfare-maximizing in the Brown and Johnson model only when efficient rationing is assumed. If other forms of rationing are assumed this price will no longer be optimal because the amount of consumer surplus lost due to rationing can be reduced by raising the price. We can illustrate this result in the following figure.

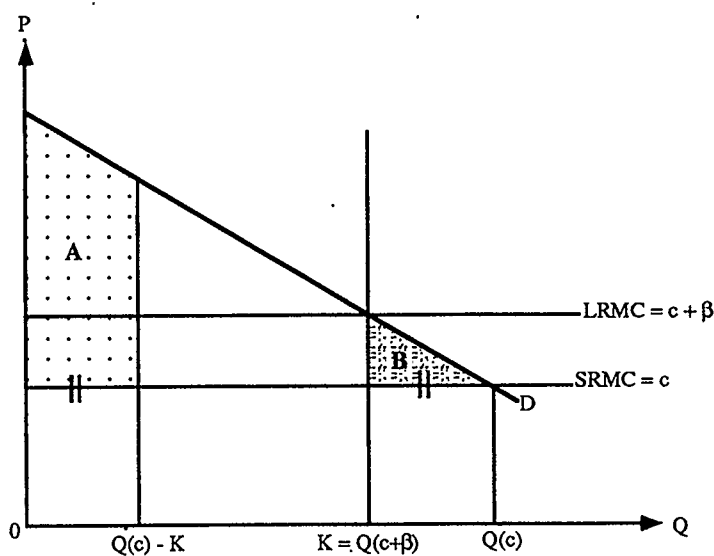


Figure 3.8
Rationing and Price

In Figure 3.8 capacity is installed at $K = Q(c + \beta)$ and the price is set at c , the marginal operating cost. At this price demand is $Q(c)$, and customers from $Q(c)$ to K go unserved when there is efficient rationing. This produces a loss of consumer surplus of the shaded triangle B . With inefficient rationing, customers from $(Q(c) - K)$ to the origin go unserved, where this distance is equal to the distance from $Q(c)$ to K . This produces a welfare loss of the polygon A and by inspection $A > B$ since the distance from 0 to $(Q(c) - K)$ is the same as the distance from $Q(c)$ to $Q(c + \beta)$. When the price is raised to $(c + \beta)$ there is no rationing because demand will exactly equal supply in the figure as drawn, but there is still the loss of consumer surplus B because these customers would have otherwise purchased service at the lower price. With efficient rationing, there will be the loss of area B with price $(P = c)$ due to rationing, as well as with $(P = c + \beta)$ since these customers would otherwise purchase service at the lower price. However, with inefficient rationing the lost consumer surplus with the new price relative to the original price, the area B , is less than the lost consumer surplus due to rationing, area A . Thus, a price increase will decrease consumer surplus less when there is inefficient rationing than when there is efficient rationing.

When there is inefficient rationing those customers with the highest willingness to pay for service go unserved. An increase in price will cause fewer customers to demand service and so there will be less lost surplus of these higher valued customers when capacity is constrained. While the Brown and Johnson price of marginal operating cost maximizes potential consumer surplus when efficient rationing is assumed, when there is inefficient rationing consumer surplus is maximized with a price higher than the optimal Brown and Johnson price.

Visscher (1978) reformulates the Brown and Johnson model with different rationing assumptions. He finds that with inefficient rationing the optimal price is indeed higher than the marginal operating cost; with inefficient rationing the optimal price is equal to the long run marginal cost. Under certain conditions Visscher also shows that an optimal capacity level which is smaller than the Brown and Johnson level is feasible.^{16 17} While an increase in capacity increases consumer surplus by reducing the amount of consumer surplus lost to rationing, with inefficient rationing those customers who benefit from a capacity expansion, those customers who are served first with the lowest willingness to pay, do not value the commodity enough at the margin to justify the cost of the expansion. Although more customers can be served when capacity is expanded, with inefficient rationing relatively more customers with low valuations of service will be supplied than will customers with high valuations of service and it is possible that the increase in capacity will increase consumer surplus less than it will increase costs.

With random rationing the order of service and the willingness-to-pay are completely uncorrelated. The optimal price under random rationing will be higher than with efficient rationing but lower than with inefficient rationing because the average valuation of service with random rationing will lie somewhere in-between the two extremes of efficient and inefficient rationing.

In summary, we can see that the optimal price in the Brown and Johnson model depends heavily on their assumption of perfect and costless rationing. With inefficient rationing this price will be higher.

¹⁶Visscher (1973), p. 226.

¹⁷If the linear demand is sufficiently elastic at the higher price, or if capacity costs are high enough, the optimal capacity might be lower since with inefficient rationing the resource is allocated to those who value it the least (Visscher (1973), p. 227). Also see Carlton (1977) for a discussion of the impact of rationing schemes on the optimal pricing strategy.

3.2.2 Revenue Constraints

Brown and Johnson found the welfare-maximizing pricing strategy in their model was to set the price equal to the short run marginal cost. Although this price maximizes total surplus the producer will suffer negative profits because it will never recover capacity costs. With Brown and Johnson's pricing solution revenues will always be less than costs because the price charged is less than the long run marginal cost. In the long run the producer must be subsidized or it will shut down. One way to ensure that the producer earns non-negative profits is to constrain revenues to cover costs. This constraint will raise the optimal price above short run marginal cost.

The difficulty in constraining revenues is that revenues and costs will be unknown until demand is realized. Only the expected value of revenues and costs can be known. Sherman and Visscher (1978) introduce a risk constraint into the Brown and Johnson model. This constraint restricts the expected value of revenues to be equal to the expected value of costs. For certain realizations of the stochastic variable a surplus or deficit will be possible but this constraint ensures that profits on average will be zero. Sherman and Visscher find that the addition of a revenue constraint raises the optimal price above short run marginal operating cost.

Sherman and Visscher also extend the model to many periods. Uncertainty exists in each period since the magnitude of realized demand is not known until the state of the world is realized. Because expected welfare is maximized by setting expected revenues equal to expected costs, a form of Ramsey prices across the demand periods can then be derived using expected elasticities and the chance constraint,¹⁸ so that price is a function of the expectation of the demand elasticity based on expected sales. With this form of Ramsey pricing the price in one period is raised if a relatively more inelastic demand is expected, thereby minimizing the surplus lost while still covering total costs, since inelastic demanders respond to price changes less sharply than do elastic demanders. In effect, this approach uses non-uniform pricing across periods to raise revenues above costs, and is a generalization of the Ramsey pricing rule in its deterministic form.

Sherman and Visscher's extension of the Brown and Johnson model is interesting because it is an attempt to use peak-load pricing in the stochastic setting, although the

¹⁸See Sherman and Visscher (1978), equation (11), and Berg and Tschirhart (1988), equation (6.14), for specific formulations.

distinction between peak and off-peak demand becomes blurred since demand will shift depending on the value of the stochastic term. Notice that while the price for one period could be set high in expectation of high sales, it is possible that a low demand could be realized instead, resulting in a large loss of total surplus. As we will see in the next chapter, a more interesting question is how to price discriminate within the one-period model.

A second way of dealing with the revenue problem is to use a chance constraint; a certain probability distribution for revenue and costs will be associated with each choice of price and capacity. With some probability distributions the probability of non-negative profits is higher than with other probability distributions. Prices and capacity must be selected such that the probability of non-negative profits is as great as possible, or at least as great as some minimum, arbitrarily set, value. This method once again raises the price in all periods above short run marginal operating costs so that total costs are covered.¹⁹

3.2.3 Rationing Costs

A capacity which is strictly less than demand gives rise to excess demand and to rationing. The problem of excess demand is further aggravated by the low price in the Brown and Johnson model. One cost of excess demand is the loss of consumer surplus incurred by consumers due to rationing. In a welfare-maximizing framework, this cost should be subtracted from total consumer surplus. Another cost imposed by excess demand is the rationing cost, the cost to allocate available supply in periods of shortages, and this cost should be subtracted from total profits.

Brown and Johnson assume both efficient and costless rationing. They assume that consumers are costlessly ranked according to their willingness to pay. However, even with perfect information, so that the social planner does not need to expend resources to calculate the willingness to pay of consumers, administrative costs to rank consumers and ration available supply among them must be incurred. These administrative costs are the rationing costs.

One method of including rationing costs is to charge a constant penalty cost for each unit of excess demand, as Crew and Kleindorfer (1978) have done. The total penalty increases as excess demand increases. If there is no excess demand there will be no penalty charges. In Figure 3.9 a constant penalty cost of c is imposed on each unit of

¹⁹See Berg and Tschirhart (1988), p. 206.

excess demand. With capacity of K and realized demand as depicted, there is a total penalty charge equal to the shaded area. The optimal price will be higher with smaller capacity levels since the higher price mitigates potential rationing costs by limiting the quantity demanded to those consumers willing to purchase the commodity at the higher price. These rationing costs also become smaller as capacity increases so that the occurrence of excess demand becomes less likely. Obviously, this rationing cost will increase as c increases and the possibility of excess demand becomes more expensive. The optimal price will be higher than the short run marginal operating cost with the inclusion of penalty costs since these costs raise total expected costs and make the event of excess demand more expensive.

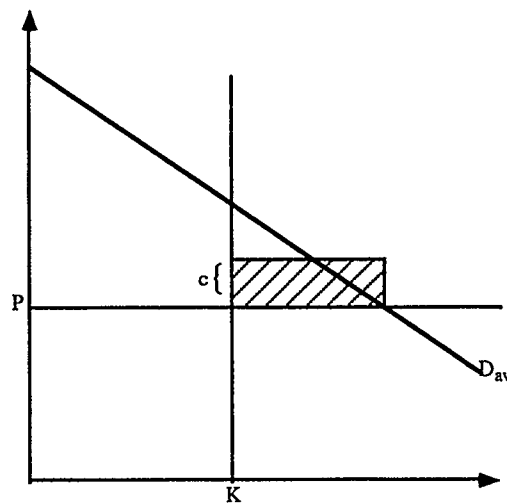


Figure 3.9
Penalty Costs

Both loss of consumer surplus due to rationing and rationing costs should be included in the objective function when welfare is maximized. Since these are both costs, the calculation of total welfare at the optimum will be overstated if they are not included. Notice, however, that the inclusion of rationing costs does not ensure non-negative expected profits because introducing rationing costs in itself does not impose a revenue constraint. Thus, with the higher price and capacity there is less rationing but this does not necessarily cause profits to be non-negative. While penalty costs do decrease the cost of

excess demand they also do not necessarily increase the probability of rationing. If capacity is very expensive it might still be optimal to invest in a low level of capacity and pay penalty costs on excess demand.

3.2.4 Reliability

The Brown and Johnson price of short run marginal cost, together with the assumption of efficient rationing, ensures that the maximum expected welfare and the lowest reliability for the customers with the lowest willingness to pay will be realized for any state of demand. At this low price there is a very high probability of excess demand which in turn implies that there is a low level of reliability for these customers since it becomes more likely that more consumers will demand the product at the given price than can be served. A high level of reliability will not necessarily be supplied because the provision of reliability is costly. If reliability increases as capacity increases and capacity expansion is expensive then reliability itself is expensive.

Brown and Johnson do not address the issue of reliability in their model. Because they have assumed efficient rationing the determination of service reliability is not an issue of interest. The customers with the highest willingness to pay are always served first and have the highest level of reliability and the customers with the lowest willingness to pay are always served last with the lowest level of service reliability. In the Brown and Johnson model only the lowest customers are concerned with reliability since their probability of service depends on how many customers there are with a higher willingness to pay. However, when there is random rationing no one customer receives a specific probability of service and so reliability is a concern for all customers. Available supply is rationed randomly among all customers demanding service so that the customer with the highest willingness to pay has the same probability of service as the customer with the lowest willingness to pay.

Regardless of the rationing method, one method to ensure that a minimum level of reliability is obtained for all customers is to impose a chance constraint or a reliability constraint. A chance constraint requires that all demand is satisfied with a specified probability. Because providing reliability is costly, no class will receive a reliability higher than the quoted level.²⁰ The chance constraint approach is representative of standard

²⁰Meyer (1975), Crew and Kleindorfer (1978) and Nguyen (1978) construct models with such constraints.

regulatory practice in countries where public utilities are required to meet established standards.²¹

The inclusion of such constraints raises prices such that specified standards of system reliability can be met by reducing demand. Whether or not reliability constraints result in a smaller optimal capacity than the Brown and Johnson level depends on the level of reliability required. A price above short run marginal cost implies that capacity could be reduced over the level that would have been required to meet the constraints while charging a price of short run marginal operating cost. However, if the level of required reliability is very high, the optimal capacity may be larger in order to satisfy the requirement.

Although reliability constraints are imposed to improve social welfare, welfare in the Brown and Johnson model actually decreases with their imposition.²² Because demand is assumed to be independent of reliability, there is no increase in social welfare with the imposition of these arbitrary constraints. Thus, the setting a high reliability through the imposition of a constraint does not have a pay-off in terms of social welfare. The imposition of a reliability constraint imposes extra costs but does not realize any extra benefits since demand does not change with the improvement in reliability.

Alternatively, demand can be assumed to be dependent on reliability so that it shifts with a change in reliability. There can be improvements in welfare due to higher reliabilities only when demand is dependent on reliability because then reliability will be endogenously determined. With reliability-dependent demand, the increase in reliability is similar to an increase in the quality of the commodity. *Ceteris paribus*, all consumers will prefer this higher quality commodity. When demand is reliability-dependent, there will be a trade-off between prices, welfare, and reliability that did not exist before.

One way to ensure that reliability is endogenously determined is to make demand a function of reliability. If consumers value reliability such that reliability is a "good" and not a "bad", demand will increase as reliability increases. With a low reliability, consumers will demand less of the commodity. Brown and Johnson assume that demand is independent of reliability; the mean-demand curve is stationary and does not shift with changes in reliability, which is an unrealistic assumption if consumers do value reliability.

Reliability dependent demand will also affect the loss of consumer surplus due to rationing because the probability of excess demand will increase as more consumers

²¹For example, electric power systems in the United States use a "1 day in 10 years" loss-of-load probability (a probability of 99.997 percent) as a reliability target.

²²See Meyer (1975).

demand this higher quality commodity. Thus the level of consumer surplus lost due to rationing when capacity is constrained is greater when demand is a function of reliability. We can see this effect in Figure 3.10. In Figure 3.10 if reliability increases from r to \bar{r} demand will increase from $D(r)$ to $D(\bar{r})$, and total surplus will increase by the shaded area A . With capacity constrained at K and a price of $(c + \beta)$ there will be the loss of consumer surplus of area E due to rationing with the demand $D(r)$. When demand increases to $D(\bar{r})$ with the increase in reliability, the loss of consumer surplus due to rationing will increase to the area $(B + E)$. Thus we can see that the loss of consumer surplus due to rationing is greater when demand is dependent on the reliability of supply.

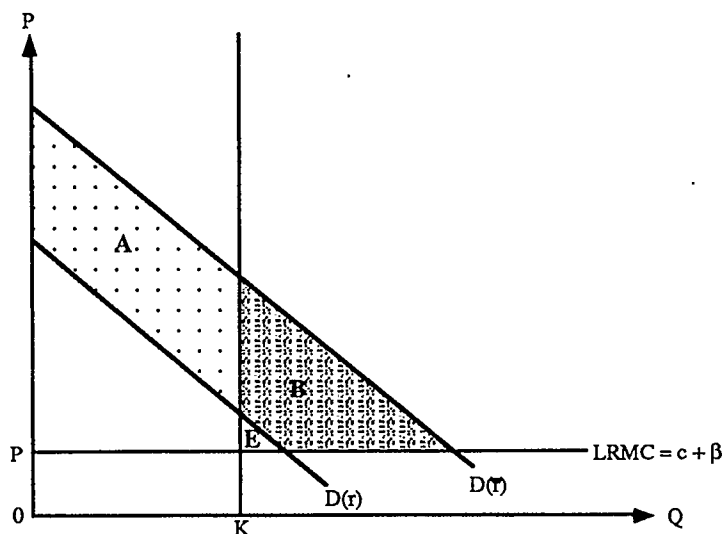


Figure 3.10
Reliability-Dependent Demand

Thus the quantity demanded by consumers will be a function of both price and reliability²³ and the social planner is then confronted by three choice variables: capacity, prices, and reliability.

When reliability-dependent demand is added to the Brown and Johnson model the optimal price will be greater than marginal cost because a greater level of reliability can be supplied with higher prices. With the higher price, fewer customers will demand service

²³ Assuming both that consumers have rational expectations and that the provision of a reliability level and a price will sufficiently enable the consumer to choose an appropriate quantity.

and the commodity can be delivered with a higher level of reliability to customers with a higher willingness to pay. The optimal price and reliability is where the incremental gains in welfare due to higher reliability exactly equal the loss in welfare due to higher prices and capacity.

The model discussed so far is posited within a welfare-maximizing framework using producer and consumer surplus as measures of welfare. Consumer surplus, however, can be criticized 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 commodity.²⁴ This is because expected consumer surplus will not reflect consumer attitudes towards risk in an uncertain market. The probability of obtaining the commodity becomes a characteristic of the commodity and consumers will have different risk preferences. The basic model developed by Brown and Johnson essentially assumes that all consumers have the same risk tolerance. If consumers have different risk preferences a uniform price will be sub-optimal since consumers will trade off the price and the probability of obtaining the commodity, and expected surplus measures will not necessarily capture this trade-off.

Another issue not addressed by Brown and Johnson is stochastic supply. Chao (1983) extends the Brown and Johnson model to include stochastic supply as well as stochastic demand. This assumption brings the model closer to reality because often not only is demand subject to random fluctuations but the availability of installed capacity is also random.²⁵ With electricity generation, for example, generators may break down leaving only part of the installed capacity available to meet demand. Because supply is assumed to be non-storable, the randomness of supply will have an impact on the optimal price.

With the introduction of supply uncertainty the optimal price is the weighted average of marginal operating costs and the marginal loss of consumer surplus due to rationing.²⁶ Thus, the optimal price rises when supply is also stochastic. Chao also adds multiple technologies, which are ranked in ascending order by (constant) operating costs. Because demand is not a function of reliability in Chao's model, reliability is determined by the capacity level. The greater the capacity level, the greater is reliability. In Chao's model, each capacity technology is employed until the expected savings of employing

²⁴See Carlton (1977, 1978).

²⁵It is assumed that the random failures of the generating units are stochastically independent of each other and of all other random variables.

²⁶Chao (1983), p. 186.

another unit of technology is just equal to the marginal capacity cost. This result parallels that of peak-load pricing with diverse technologies.²⁷

In summary, the (unmodified) Brown and Johnson model of stochastic demand yields the lowest price, the highest level of consumer surplus, and largest expected deficit. Others have extended this model to include penalty costs, revenue constraints, reliability constraints, and reliability-dependent demand. In the next section, we will illustrate the differences between these extensions with a numerical example.

3.3 A Numerical Example

In order to illustrate how the various assumptions impact on the results of the Brown and Johnson model, we will construct a simple numerical example. The welfare-maximizing objective function for the Brown and Johnson model with one period is;

$$\begin{aligned}
 W = & \int_{-\infty}^{\infty} f(\mu) \int_P^{X^{-1}(0)} [X(P) + \mu] dP d\mu - \int_{K-X(P)}^{\infty} f(\mu) \int_P^{X^{-1}(K-\mu)} [X(P) + \mu - K] dP d\mu \\
 & + \int_{K-X(P)}^{\infty} f(\mu) K \cdot P d\mu + \int_0^{K-X(P)} P \cdot f(\mu) [X(P) + \mu] d\mu \\
 & - c \left\{ \int_{K-X(P)}^{\infty} K \cdot f(\mu) d\mu + [K - X(P)] \int_0^{K-X(P)} f(\mu) [X(P) + \mu] d\mu \right\} - \beta \cdot K
 \end{aligned} \tag{3.1}$$

where: $f(\mu)$ is the continuous probability density function of the random component of demand as denoted by μ , K is the installed capacity, P is price, c and β are the short run marginal and capacity costs, and $X(P) + \mu$ is demand where $X^{-1}(Q)$ is the inverse demand function. The distribution of the stochastic term is assumed to be known.

The first term in equation is the expected consumer surplus resulting from a price P . The upper limit on the second integral of the first term is the intercept of the demand function with the price axis. In Figure 3.11 below this first term is represented by the shaded area ($A + B + E$).

²⁷Chao (1983), p. 183. See Berg and Tschirhart (1988), pp. 173-6 for a discussion of peak-load pricing with diverse technologies.

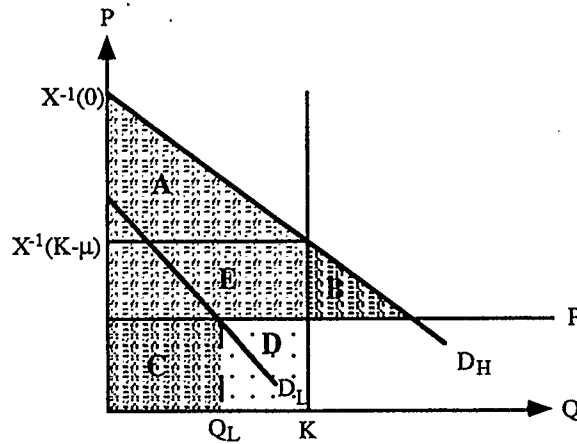


Figure 3.11
Brown and Johnson: Total Surplus

The second term in equation (3.1) is the loss of consumer surplus due to rationing, as represented in Figure 3.11 as the shaded triangle *B* for the price P and the capacity K . This loss must be subtracted from consumer surplus. It will be positive when there is excess demand and zero otherwise.²⁸ This loss is calculated from the given price to the price which would clear the market when there is excess demand, the price $X^{-1}(K - \mu)$.

The third and fourth terms in equation (3.1) are revenues when there is excess supply and when there is excess demand. Revenues when there is excess supply are represented in Figure 3.11 by the area *C* for the demand D_L , while revenues when there is excess demand are represented by the sum of areas *C* and *D*. These terms are integrated using the value of μ such that the market is cleared, $K - X(P)$. Similarly, the fifth and sixth terms are operating costs, again calculated for excess supply and demand for the constant marginal operating cost c . The last term is capacity costs which must be borne regardless of whether quantity demanded exceeds capacity or not. The constant β is the per-unit capacity cost. We assume a simple linear demand of,

$$X(P) + \mu = 25 - P + \mu \quad (3.2)$$

²⁸When the price is such that the market clears, $P = X^{-1}(K - \mu)$, the inner integral of the second term of (3.1) disappears.

where μ is a random variable. We assume that μ is uniformly distributed over the interval $[-5, 5]$ and that operating and capacity costs are $b = 2$ and $\beta = 8$.

Substituting (3.2) into the objective function (3.1) and maximizing over capacity (K), and price (P), the optimal price for this model is equal to the operating cost, as expected from our discussion of the Brown and Johnson model in Chapter 3.1. Thus, the optimal price assumes a value of 2 and optimal capacity a value of 15.351, with an expected total welfare of 112.31 and expected negative profits of -122.31.²⁹ We can construct a parallel profit maximizing model using the same parameters as before but now dropping the first two terms from the objective function (3.1). For this (parallel) profit-maximizing model, the optimal price rises to 16.766, capacity and welfare fall to 7.8161 and 72.79 respectively and profits rise significantly to 37.383.³⁰

We can extend this example to include several of the constraints discussed in Section 3.1 to illustrate the numerical effect. In our example of the welfare-maximizing Brown and Johnson model the firm earns negative profits. One way to allow the monopolist to earn non-negative profits while still maximizing social welfare is to add a revenue constraint. Thus, we add a revenue constraint of,

$$E[R] = E[C] \quad (3.3)$$

to the Brown and Johnson objective function as outlined in equation (3.1) such that expected revenues are constrained to be equal to expected costs. For this formulation, the optimal price rises relative to Brown and Johnson's results to 10.63 while capacity and total welfare fall slightly to 13.88 and 104.61 respectively.

In the peak-load model, the problem of excess demand and rationing was avoided with the addition of a feasibility constraint that demand could not exceed capacity, equation (2.2). This kind of feasibility constraint is not possible in the Brown and Johnson model because demand is stochastic. One way to address the problem of excess demand is to introduce penalty costs into the model, so that a penalty is assessed for each unit of excess demand.

Penalty costs of,

²⁹The actual value will of course depend on the value of the stochastic term.

³⁰Much of the numerical analysis was done using the Prescience software program Theorist vers. 1.51 and 2.0.

$$\phi \int_{K-X(P)}^{\infty} (f(\mu)[\{X(P) + \mu\} - K]) d\mu \quad (3.4)$$

are subtracted from the objective function defined in (3.1), where ϕ is the constant penalty cost per unit of excess demand. This penalty cost is assigned arbitrarily. The introduction of penalty costs essentially increases the costs of excess demand.

If we assume a penalty cost of \$1, the formulation yields a price of 9.57, a capacity of 18.335, and expected values for welfare and profits of 72.14 and -31.53. As can be expected, prices now exceed marginal operating costs. This is because the higher price reduces potential rationing costs. We can expect the price to be higher the smaller is capacity since this reduces the risk of excess demand. However, notice in this example that the optimal capacity is higher than in the original Brown and Johnson solution because a higher investment in capacity reduces the risk of excess demand and rationing costs.

Another approach to the problem of excess demand is to introduce reliability or chance constraints. The reliability constraint can be constructed as;

$$P\{[X(P) + \mu] \geq K\} \leq \varepsilon > 0 \quad (3.5)$$

where the first term denotes the probability that demand is less than supply and ε is the chosen level of system reliability such that higher values of ε reflect higher reliability levels. In general, the reference level of reliability is chosen arbitrarily. This constraint restricts the probability of excess demand to be less than ε .

Adding the constraint (3.5) to the objective function (3.1), and with a chosen reliability level of 0.77,³¹ the model yields a price of 5.5312 and a capacity of 15.239. While capacity remains approximately unchanged for this chosen level of reliability, the optimal capacity may diverge from this level depending on the stringency of system reliability chosen. For this level of reliability, expected welfare also remains relatively unchanged from the base case at 112.37. Again, the price increase necessarily reduces the probability of excess demand. The price increase also means that the level of capacity can be reduced, since less is demanded at the higher price. In this example, reliability can be increased by raising the optimal price and reducing the level of optimal capacity slightly. If

³¹Reliability standards are very much a political issue, and are usually set by regulators to be quite stringent, at least in developed countries. As the stringency of the reliability constraint increases, the value of ε also increases to reflect this stringency.

the imposed level of reliability was higher than the one imposed here it might be possible that capacity will be expanded to fulfill this more stringent capacity constraint.

In the example of the Brown and Johnson model developed above demand was assumed to be independent of the level of reliability. If customers have different preferences for different levels of reliability this is an unrealistic assumption. As well, because demand is assumed to be independent of reliability, the setting of a high reliability through constraints does not realize any substantial payoff in terms of social welfare. If customers value reliability we would expect the setting of reliability to result in substantial welfare gains. Referring to the numerical analysis above we can see that setting reliability constraints or introducing penalty costs, while increasing the level of service reliability, does not significantly change the level of total surplus relative to the basic Brown and Johnson example.

We can adjust the simple linear demand from the Brown and Johnson model, equation (3.2), to reflect changes in reliability such that demand increases with increases in reliability;

$$X(P, r) + \mu = 25 - P + r + \mu \quad (3.6)$$

As well, a constraint must now be added to the objective function such that actual reliability is consistent with quoted reliability;

$$P\{[X(P, r) + \mu] \leq K\} \geq r \quad (3.7)$$

While constraint (3.7) is very similar to the constraint (3.5), reliability is now a choice variable, and so reliabilities are chosen endogenously rather than being arbitrarily imposed. This equation constrains the quoted level of reliability to be no greater than the actual level of reliability supplied. We assume that the constraint (3.7) is binding since the provision of reliability is costly.

Optimizing (3.1) over P , Z and r , the optimal level of reliability is 0.77. The optimal price rises to 6.07, and capacity increases slightly to 15.84. Notice that welfare rises as predicted to 124.09, because there can be improvements in welfare due to higher reliabilities only since reliability is included as an argument in the demand function.

Additional constraints can also be added to this model of endogenous reliability. With the addition of a revenue constraint, such as constraint (3.3), the optimal price rises

dramatically to 10.30, while capacity falls to 13.691. Expected welfare falls to 119.35, but notice that this level of welfare is still greater than in the (unextended) Brown and Johnson solution. Due to the revenue constraint, the producer breaks even with zero expected profits, and the optimal level of reliability drops slightly to 0.769. Notice that even in the profit-maximizing case that expected total welfare is higher -- 82.35 as compared to 72.79 in the reliability independent demand profit-maximizing case. Expected profits are also higher, at 43.53, with a reliability of 0.880 and a (higher) capacity of 8.52. Thus, demand which is dependent on reliability allows the profit-maximizing producer to capture some of the increased consumer surplus due to the increased capacity level and increased level of service reliability.

In summary, we can see that the Brown and Johnson price of marginal operating cost is optimal when demand is independent of reliability. However, welfare gains due to increased reliability can be had only when demand is dependent on the level of reliability chosen.

We can compare the numerical results of the Brown and Johnson model and the various extensions explored in this section with the following table:

Table 3.1
Brown and Johnson and Extensions: Numerical Summary

Case ³²	Price	Capacity	Total Welfare ³³	Profits
Brown and Johnson	$P = 2$	$Z = 15.351$	$E(TW) = 112.31$	$E(\pi) = -122.31$
Profit Maximizing $\beta=8$ $\beta=12$	$P = 16.766$ $P = 35.50$	$Z = 7.8161$ $Z = 20.63$	$E(TW)=72.789$ $E(TW)=512.15$	$E(\pi) = 37.383$ $E(\pi) = 319.33$
Penalty Costs³⁴	$P = 9.5712$	$Z = 18.335$	$E(TW)=72.138$	$E(\pi) = -31.525$
Reliability Constraints	$P = 5.5312$	$Z = 15.239$	$E(TW) = 112.37$	$E(\pi) = -68.204$ $r = 0.77^{35}$
Revenue Constraints	$P = 10.634$	$Z = 13.88$	$E(TW) = 104.61$	$E(\pi) = 0$
Reliability Dependent Demand³⁶ with revenue constraints	$P = 6.0737$	$Z = 15.834$	$E(TW) = 124.09$	$E(\pi) = -62.432$ $r = 0.77$
profit maximizing ³⁷ $\beta = 8$	$P = 10.304$	$Z = 13.691$	$E(TW) = 119.35$	$E(\pi) = 0$ $r = 0.7687$
$\beta = 15$	$P = 17.30$	$Z = 8.52$	$E(TW) = 82.35$	$E(\pi) = 43.53$ $r = 0.880$
	$P = 6.39$	$Z = 19.53$	$E(TW) = 181.76$	$E(\pi) = -339.57$ $r = 0.475$

As we will see in the next chapter, there are also substantial welfare gains in offering differentiated service. Given that consumers may have different preferences and different willingness to pay for service reliabilities, the offering of a menu of service options which more closely match the need of consumers will increase social welfare. Models of differentiated service levels build on Brown and Johnson's model of stochastic demand by adding a spectrum of service options.

³²Assuming per-unit operating and capacity costs of $b = 2$ and $\beta = 8$, and a linear demand of the form $D(P) = (25 - P + \mu)$.

³³Total welfare and profits are expected values. The actualized values depend upon the state of demand realized.

³⁴Assuming a penalty cost of \$1 per unit of excess demand, $\phi = 1$.

³⁵The reference reliability level is chosen arbitrarily.

³⁶Assuming a linear demand of the form $D(P) = (25 - P + r + \mu)$, where r denotes reliability. A constraint must also be added such that actual reliability is consistent with quoted reliability: $P[X(P, r)\mu \leq Z] \geq r$.

³⁷Assuming $b = 2$, $\beta = 15$, and $D(P) = (55 - P + r + \mu)$, such that comparisons with reliability-dependent interruptible pricing are possible.

CHAPTER FOUR

INTERRUPTIBLE PRICING

4.1 Introduction

Brown and Johnson (1969) explore the welfare foundations of peak-load pricing under uncertainty. They find that the optimal price is based on the marginal operating cost and that the optimal capacity is set to equate at the margin the expected loss of consumer surplus due to excess demand with the cost of additional capacity. Brown and Johnson derive their results under very specific assumptions. Their central assumptions are;

- (1) efficient rationing of supply
- (2) the absence of rationing costs
- (3) the absence of revenue constraints, and
- (4) reliability-independent demand

We have seen that their price is no longer optimal when these assumptions change. When any one of these assumptions is invalidated the socially optimal price is higher with a corresponding lower optimal capacity level than in the Brown and Johnson solution.

We will briefly summarize the effect of each of these assumptions on Brown and Johnson's results. In their model, Brown and Johnson assume that the available supply is somehow rationed efficiently. Efficient rationing or allocative efficiency means that supply is distributed first to those customers with the greatest willingness to pay for service. Thus, in periods of supply shortages, consumer surplus is greater with efficient rationing than without because supply is allocated to those who value it the most. The effect of a price increase on consumer surplus will be greater with efficient rationing than with inefficient rationing because the marginal willingness to pay for service, averaged across all served customers, is greater under efficient rationing. The optimal price therefore will be greater under inefficient rationing than efficient rationing because a price increase under inefficient rationing will reduce consumer surplus by a lesser amount. As we will see later on, one way to induce efficient rationing is through service differentiation.

Rationing assumptions will also affect the choice of the optimal capacity. Clearly, if capacity was costless, capacity could simply be built to satisfy maximum demand. It is because capacity investments are expensive that rationing must occur in periods of supply shortages. Capacity decisions are based on the aggregate willingness to pay for service

evaluated at the margin. Because this marginal value is greater with efficient rationing, the optimal capacity will also be greater with efficient rationing than with inefficient rationing.

Turning now to the second assumption, we should again distinguish between rationing costs and the loss of consumer surplus due to rationing. Rationing costs are the costs incurred to implement the rationing scheme, such as administrative or implementation costs, while costs of rationing are the welfare losses in consumer surplus and profits due to insufficient supply. Rationing costs increase the cost of insufficient supply because cost must be incurred to ration supply among consumers. In effect, the presence of rationing costs raises the expected value of operating costs. For any given capacity, the optimal price is higher with rationing costs in order to cover these expenses.

While total welfare is maximized in the Brown and Johnson model the producer suffers negative profits. The introduction of revenue constraints into their model naturally raises the optimal price so that the producer earns non-negative profits.

Even with the high level of optimal capacity in the Brown and Johnson model the level of service reliability for the last customer is very low since the low price in this model increases the probability that there will be excess demand. The addition of a chance constraint on reliability is an attempt to alleviate this problem. The probability of service to all customers is constrained to be no less than some arbitrarily chosen level. This constraint naturally raises the optimal price to increase the service reliability as fewer customers demand service. However, this constraint does not significantly raise the level of welfare because Brown and Johnson assume that demand is independent of reliability. There can only be significant gains in social welfare with improvements in reliability when demand is dependent on reliability.

With reliability-dependent demand both consumer surplus and total revenues will increase with increases in reliability. Now the social planner is subject to a different constraint, the feasibility constraint. Given that consumers are rational and can observe the delivered level of reliability, the social planner is constrained to offer those levels of reliability that it can actually supply.³⁸ Reliability to consumers can be increased in two ways; (1) by increasing capacity, and (2) by raising price. Price increases improve the

³⁸Panzar and Sibley (1978) and Woo (1990) examine a form of interruptible pricing called "self-rationing" which is different from the models examined here. With self-rationing consumers either subscribe to or are assigned a fixed capacity. However, unless all consumers reach their subscription capacity level simultaneously, there will be efficiency losses with this rationing system because a particular consumer may reach her capacity limit while there still exists excess system capacity. This is sometimes referred to as "too much load relief" (Doucet, 1993, p. 94).

level of reliability because fewer customers will demand service. The optimal price and capacity will be higher when demand is dependent on reliability because the improvement in reliability will increase consumer surplus.

The feature that is missing in the Brown and Johnson model is that different customers may value service reliability differently. They may have different preferences or "tastes" for reliability. In the Brown and Johnson model, all customers are charged the same price and in periods of excess demand some customers go unserved. While all customers are offered the same reliability of service some customers may prefer a lower level of reliability and others a higher level.

Another approach to this problem of rationing is the offering of interruptible service. This approach takes advantage of differences in consumer preferences. Two types of service are offered: firm and interruptible service. Interruptible customers are offered service of a lesser reliability, usually at a lesser price, than firm service. This is a form of product differentiation. Although only one commodity is produced, customers are supplied with different levels of reliability, so that the commodity at each level of reliability can be thought of as a separate commodity. Social welfare is increased when a variety of products are supplied, each targeted at a different consumer type. Interruptible service differentiates the quality of service on the basis of reliability. Service with a higher reliability will be of a higher quality and some consumers will prefer this higher quality of service. Product differentiation can be used to exploit heterogeneity among consumer preferences.³⁹

Models of interruptible service rest on the realization that welfare can be increased if, instead of charging a uniform price, customers are charged according to their willingness to pay. This is primarily due to two reasons; (1) efficient rationing, and (2) capacity utilization.

The offering of different classes of service promotes efficient rationing through pricing. Customers who value reliability highly are served with a higher reliability and are charged correspondingly. Thus the rationing order reflects the order of service classes and the ranking of service valuation. One can imagine an infinite array of service classes, each with a reliability and price set for each individual customer, such that the population of customers is perfectly and efficiently rationed. While such a reliability/price menu is not

³⁹When offering interruptible contracts, customers must be able to verify actual supply conditions, such that customers can verify that actual probability of interruption is the same as the contracted probability. Because this is difficult, contingency contracts, stipulating that service will not be delivered under listed objectively verifiable conditions, are often used.

possible due to implementation costs, there are still significant gains in welfare to be had with a finite number of service classes.⁴⁰

The offering of differentiated service allows available capacity to be utilized more effectively. Because interruptible demand is served with a lower level of reliability the social planner can substitute interruptible demand for expensive capacity expansions. Interruptible demand essentially "frees" up extra capacity in periods of shortages because service is shifted to firm customers.

In addition, interruptible service better covers costs of service because it allows a finer extraction of consumer surplus. The firm is less likely to suffer negative profits with differentiated pricing because it can both charge higher prices to customers willing to pay for increased reliability and substitute service to interruptible customers willing to accept a lower level of reliability for expensive capacity expansions.

The model of interruptible service that we discuss here is the Tschirhart and Jen (1979) model. This is a profit-maximizing model based on the assumption that the firm has access to perfect information. Armed with this information, the firm divides customers into classes based on some observable characteristic(s) and then determines the order in which service to these classes is to be interrupted. It is assumed that arbitrage, the re-selling of the commodity between customers, is not possible because the firm is endowed with perfect information. The firm can simply prevent customers from reselling the good and from subscribing to a different demand class than the one to which they have been assigned. The interruption ordering is a decision variable which is only relevant in the stochastic setting; if both demand and supply are non-stochastic an interruption ordering would be superfluous because capacity would never be constrained. In this interruptible model, each class is charged a different price depending on their position in the overall priority ordering. Because each consumer class occupies a different position in the ordering, each class confronts different reliabilities of service.

⁴⁰While having one class for each customer would allow the firm a greater range of discrimination, the implementation costs of such a system is implicitly assumed to be prohibitive. The number of classes implemented depends on the relative trade-off between the costs of adding another class and the gains of greater discrimination. In practice, the number of consumer classes is generally limited by the set of consumer characteristics and by transaction costs.

4.2 Tschirhart and Jen (1979)

Tschirhart and Jen extend the work of Brown and Johnson by dividing consumers into service classes and establishing a priority ordering that specifies the sequence in which service for each class is interrupted when there is excess demand. Thus, each class is served with a different level of reliability. Recognizing that different consumers have different preferences for reliability, the firm discriminates among customers, charging higher prices for service based on the elasticities of demand. Prices for these service classes are constructed subject to a number of conditions. As we will see below, in order to discriminate among customers the firm must allow for price and/or reliability elasticities, a result similar to Ramsey pricing. Prices must also clear the market of excess supply so that capacity is not idle.⁴¹ Given that customers can observe the delivered level of reliability, the firm is subject to a feasibility constraint. Offered reliabilities must be compatible with the quoted level of reliabilities for each class.

Tschirhart and Jen's model of a profit-maximizing monopoly offering interruptible service. In their model, the firm, endowed with perfect information, divides its customers into classes and ranks each class before demand is realized. If supply is insufficient to fully serve a demand class, customers within that class are rationed randomly.

With interruptible service, the firm divides its customers into two classes: the stochastic class, whose demand is subject to random disturbances, and the contractual class. The demand of the contractual class is non-stochastic because these customers contract with the firm for specified levels of service at a given price and reliability of service. These contracts state that the demand of the contractual customer will be satisfied as long as supply is available. It is assumed that the stochastic customers have higher priority and a lower probability of interruption than do contractual customers. The stochastic class is always served first and the contractual class is interrupted on the basis of the priority ordering. Since the demand of the stochastic class of customers is subject to random disturbances, this class then causes stochastic supply conditions for the remaining classes and stochastic demand for the firm.⁴²

⁴¹As long as the willingness to pay for service above operating costs is non-negative, the firm will provide service if there is idle capacity.

⁴²This is similar to the formulation where customers choose a baseload demand which is always served while their remaining demand is stochastic.

If any of the contractual classes have a higher priority than the stochastic class, the firm treats their demand as deterministic, and their service is never interrupted. The optimal price then for this class is simply the non-stochastic price. In most of the literature, however, the stochastic class has the highest priority and all other classes are eligible for interruption.⁴³

With interruptible service the firm rations available supply in the ordering of service interruption of the various classes while within each class supply is allocated randomly. Because expected total surplus depends on the ordering of the classes, the firm treats the ordering as a decision variable. When capacity is restricted,⁴⁴ a large demand from the stochastic class implies that demand from all classes cannot be satisfied and rationing must occur. In this way output allocations for each customer are interdependent because the output allocated to one customer depends on the output allocated to all other customers. This interdependence is sometimes referred to as “congestion”,⁴⁵ since the probability of supplying one individual’s demand decreases as demand as a whole increases.

The firm must first determine the optimal ordering of interruption classes. Once the optimal interruption ordering is determined, the firm then choose the n prices, the n reliabilities and the level of capacity that will maximize expected profits.

With priorities $(1, 2, \dots, n-1, n)$, class $(n-1)$ customers are served before class n customers are served. Thus, if supply is restricted during periods of excess demand such that total demand from all classes cannot be met, the demand of class n is reduced below its contracted amount until the demands of all classes with higher priorities, classes 1 to $(n-1)$, can be met. If this is not possible, the n th class is not supplied at all and the supply to the $(n-1)$ th class is curtailed until the demands of classes one through $(n-2)$ are met. If supply is still insufficient, supply to both the n th and the $(n-1)$ th class is cut off, and supply to the $(n-2)$ th class is restricted until demand of classes one through $(n-3)$ is met. This process of interruption is continued until there is adequate supply to serve the remaining higher priority users. If supply is still insufficient to meet class 1 demand after all contractual classes have been interrupted, then available supply is rationed only among these class 1

⁴³This is because almost all of the interruptible literature is modeled on electricity supply, where the stochastic class, the residential customers, have the highest priority ranking. Contractual customers are commercial and industrial consumers, who have a lower ranking than do residential consumers.

⁴⁴Spulber (1993), p. 243, suggests that positive increasing marginal cost is analogous to binding capacity constraints on total output in that both assumptions cause consumer output allocation to be interdependent.

⁴⁵See Viswanathan and Tse (1989) for a discussion of congestion and how it relates to interruptible pricing.

customers. Thus, the n th class receives the poorest quality of service, while the first class, the stochastic class receives the highest.⁴⁶

This ordering can be sequential or non-sequential. With sequential ordering, customers receiving the highest level of reliability pay the highest prices for service, and customers served with a low level of reliability pay the lowest prices. While this makes sense intuitively, non-sequential pricing is sometimes optimal when demand is reliability-dependent. With non-sequential pricing customers paying the lowest service prices may not necessarily receive the lowest level of reliability. When demand is a function of reliability there will be a trade-off between ordering interruptions to exploit price elasticities and ordering to exploit reliability elasticities. With reliability-independent demand, the optimal interruption ordering is always sequential, but with reliability-dependent demand the ordering may be either sequential or non-sequential depending on the relative strengths of the two elasticities. If the reliability elasticities outweigh the effects of the price elasticities, the optimal prices may be non-sequential.

When determining expected profit the firm must estimate the expected operating revenue⁴⁷ from class I when this class is fully served as well as when class I demand exceeds capacity. However, the expected operating revenue will be different for the interruptible classes because they are served only after the demand of the stochastic class is filled. Given an interruption ordering, the firm maximizes expected profits by choosing n prices, n (quoted) reliabilities, and a level of installed capacity. The firm chooses among all possible interruption orderings to select the ordering that, given the optimal price, reliability, and capacity, yields the maximum profit. Thus the priority ordering is a decision variable and the firm must consider the profit implications of $(n-1)!$ different orderings.⁴⁸ This then becomes a two-stage problem; first, the firm must determine the maximum profit for each ordering, and then it must determine the ordering with the largest profit.

Given that customers can both verify the level of reliability and the probability of obtaining service with a particular level of reliability, the firm is constrained to offer a quoted reliability which is greater than or equal to the actual reliability for each class of

⁴⁶Strauss and Oren (1993) consider a model in which consumers are given the option of early notification of supply interruption, which yields additional gains if customers value early notification.

⁴⁷Operating revenue is defined as the receipts from sales less operating costs.

⁴⁸There are $(n-1)!$ and not $n!$ different orderings because the stochastic class, (class one), is always served first.

service offered.⁴⁹ In practice this constraint operates as an equality because there is no motivation for the firm to act otherwise. Actual reliabilities will not be higher than quoted reliabilities because the provision of reliability is costly. Similarly, the firm will not quote reliabilities which are lower than the actual reliabilities because this might result in excess supply. With a lower reliability, customers may demand less service, and the firm may be left with idle capacity in some contingencies. Thus, the constraint will be satisfied at a given capacity only when the quoted prices and reliabilities result in demands which yield a compatible set of actual reliabilities. Furthermore, if class n is quoted a positive reliability of service, all classes are quoted positive reliabilities since class n receives the lowest quality of service.

By changing the quoted reliability, the firm changes the riskiness of excess demand to customers and so customers, given that they value reliability, will adjust their demand. Demand is therefore dependent upon reliability and reliability will be determined endogenously.

The firm constructs a menu of price bundles where each price offered, p_i , is bundled with a corresponding reliability, r_i , which is the quoted minimum probability of service for each class. Before demand is realized and production occurs, each class of customers is offered the price and reliability option designed for them. After consumers face this price option and supply is realized consumers are either served or their service is interrupted, depending on the ranking of their class within the ordering. If a class is not fully served, supply is rationed randomly within the class. When the demand for a class is supplied, that class is then fully served. Available supply can range anywhere between zero and the actual class demand, depending on the demand of the first class. With a price/reliability bundle (p_i, r_i) , the customer pays p_i per unit for all units of supply consumed with a probability of supply interruption of r_i . Thus, the firm can loosely be thought of as discriminating among customers on the basis of reliability.

The optimal profit-maximizing prices in this model exceed operating costs for all classes and the price for class 1, the class with the highest reliability of service, exceeds operating plus capacity costs. The firm charges the highest price to those customers with the highest reliability of service and the highest willingness to pay. This result is closer to the (non-stochastic) peak-load prices than it is to the uniform price in the Brown and Johnson model. Like the peak-load model, interruptible pricing exploits the difference in

⁴⁹Thus the constraint is $G(h(i)) \geq r_i$, where $G(h(k))$ is the actual probability that the demand of the first k classes does not exceed capacity and r_i is the reliability for class i . (Tschirhart and Jen (1979), p. 248.)

the willingness to pay between customers and charges higher prices to those customers who value the quality of service the most. This higher price helps the firm to satisfy the feasibility constraint by reducing demand and easing the strain on capacity. Class I customers contribute more to capacity costs than other customers because it is these customers who are served with the highest reliability and who place the greatest demand on capacity. The lower prices for the contractual class help to utilize capacity in periods of excess supply. It is this pricing result that ensures that the firm does not incur negative profits and covers operating costs for all service classes.

Reliability is increased in either of two ways; (1) through price increases, so that demand falls and capacity is freed up, or (2) through capacity expansion. A change in price alters demand and supply for all lower priority classes. For example, a change in price for class i changes the quantity demanded by class i as well as the supply available to the lower priority classes $(i+1)$ through to n . As capacity is expanded, actual and quoted reliability increases, causing increases in demands and revenues. This effect is captured in the operating revenue, the receipts from sales less operating costs, for each demand class as sales increase with the increase in supply. There is an additional effect captured by marginal revenues which is the increase in the demand for service as reliability increases. As reliability of service increases the quality of service will increase and customers will want more of this higher quality commodity, given that the price is the same.⁵⁰ The optimal profit-maximizing capacity is chosen such that the sum of the expected marginal operating revenue from increasing supply and the expected marginal revenue from increasing service reliabilities is equal to the marginal cost of expansion, the per-unit marginal capacity cost. Capacity is increased until the increase in revenues from increasing capacity no longer covers the cost of capacity expansion.

The firm must choose the ordering with the largest profit among the $(n-1)!$ possible orderings. As mentioned previously, the ordering of prices can be either sequential or non-sequential. With sequential prices, those classes receiving a lower quality of service are charged lower prices than those classes receiving higher qualities of service. If reliability is valued by consumers, sequential prices will seem to be more equitable to customers. In contrast, non-sequential prices do not follow any particular price/reliability pattern.

⁵⁰*Ceteris Paribus*. Referring to Figure 3.10, at the constant price P , as reliability increases from r to \bar{r} , demand increases from $D(r)$ to $D(\bar{r})$.

If demand is reliability-independent the optimum price menu will be sequential because the marginal revenue loss due to excess demand will be the smallest for the class which is charged the lowest price. With reliability-independent demand, consumers will not change their demand in response to changes in quoted reliability. Thus, low prices are coupled with low reliabilities and the demand of any class will be invariant with respect to the position that class occupies in the ordering. In general, this will not be true with reliability-dependent demand because changes in the quoted reliabilities will now affect demand. Consumers whose demand is very sensitive to changes in reliabilities may be charged a low price but receive a high level of reliability. In effect the firm must now consider both price and reliability elasticities of demand. Non-sequential ordering may be profit-maximizing with reliability-dependent demand depending on the trade-off between these two elasticities.

With reliability independent demand, prices for the various service classes can be constructed in a manner similar to the Ramsey pricing rule such that prices are inversely proportional to demand elasticity (where demand or price elasticity is the percentage change in the quantity demanded with respect to the percentage change in price). The price of service for class i is inversely proportional to class i price elasticity and directly proportional to the prices for lower priority classes, classes $((i+1), \dots, n)$ when demand is independent of reliability.⁵¹ The pricing rule for reliability-dependent demand is more complex. With reliability-dependent demand, it may be more advantageous for the firm to set prices using reliability, rather than price, elasticities.^{52,53} In addition, the firm also faces a trade-off between the cost of providing a greater level of reliability and the higher revenues from the increase in demand.

When demand is a function of reliability the profit-maximizing ordering will depend on both price and reliability elasticities. These two elasticities are inverse with respect to the quantity demanded; as price increases, the quantity demanded will decrease, while as reliability increases the quantity demanded increases because the quality of service increases. There is a trade-off between ordering interruptions to take advantage of price elasticities and ordering to take advantage of reliability elasticities. If demand for a

⁵¹See Tschirhart and Jen (1979), p. 252.

⁵²Mathematically, price elasticity is defined here as $\frac{\partial D(p_i, r_i)}{\partial p_i} \cdot \frac{p_i}{D(p_i, r_i)}$ and $\frac{\partial D(p_i, r_i)}{\partial r_i} \cdot \frac{r_i}{D(p_i, r_i)}$ is defined as the reliability elasticity.

⁵³Thus, consumers are both price and reliability sensitive. In response to a change in price consumers will move along their price curve, but with a change in reliability the demand curve itself will shift. (See Figure 3.10).

particular class is highly elastic in both prices and reliabilities it may be profitable to charge this class a low price but assign it a high priority. The assignment of priorities will depend on the relative strengths of the two elasticities. The firm has an incentive to assign classes which are reliability elastic a high priority because such classes are more responsive to changes in reliability, while there is a similar incentive to charge classes which are price elastic low prices because these classes are more responsive to changes in prices. In general, the optimal price ordering will tend towards low prices to classes with price elastic demands and high reliabilities for classes with reliability elastic demand.

With non-sequential orderings of priorities it is essential that the firm is able to prevent arbitrage between customers. All consumers will want the low priced, high reliability commodity because it is a higher quality commodity provided at a lower price. The firm cannot offer this commodity to all consumers and still satisfy the feasibility constraint and so it must be able to prevent customers from consuming commodities other than the one provided for them. If welfare is maximized with sequentially ranking it can be assumed that the effect of reliability elasticities is over-ridden by the effect of price elasticities.

We can illustrate these conflicting incentives in the following table;

Table 4.1
Price/Reliability Elasticities and Priority Orderings

Price Elasticity	Reliability Elasticity	Price Incentive	Reliability Incentive	Ordering
elastic	elastic	decrease	increase	non-sequential
elastic	inelastic	decrease	decrease	sequential
inelastic	elastic	increase	increase	sequential
inelastic	inelastic	increase	decrease	non-sequential

If a class has a high price elasticity, there is the incentive for the firm to charge a relatively low price and assign that class a low priority. However, if the same class also has a relatively high reliability elasticity the firm might want to assign that class a higher priority instead. From Table 4.1 we can see that with reliability-dependent demand sequential ordering is optimal only when price and reliability elasticities are opposing. This is because these elasticities are inversely related.⁵⁴

⁵⁴Tschirhart and Jen (1979), p. 255.

In summary, in structuring the optimal price menu, the firm must choose prices to;

- (1) to allow for price/reliability elasticities
- (2) to clear the market of excess supply, and
- (3) to be compatible with the quoted levels of reliabilities

Thus the firm is confronted by a multitude of decision variables and trade-offs between variables.

The optimal price structure becomes even more complex for the welfare-maximizing monopolist. Because consumer surplus must be calculated for each class in order to find the social optimum, demand curves must be specified for each class. In any ordering where class n 's service is completely cut-off before any curtailment to class $(n-1)$, for a social optimum it must be true that the willingness to pay of class n for the last unit cut-off is less than the willingness to pay of the first unit to be curtailed in class $(n-1)$. Because this is not always true, a social optimum may require that service be rotated among classes.^{55 56}

4.3 A Numerical Example

Building on the models constructed in the previous section, a similar example can be built for the Tschirhart and Jen model. In this example, we will assume only two classes of demand; firm (class 1), and interruptible (class 2). Class 1 demand is always served first so the optimal ordering will not be a choice variable in this example. Class 1 demand is constructed as;

$$X_1(p_1) = 30 - p_1 + \mu \quad (4.1)$$

where μ is the stochastic element of demand and is distributed uniformly between $[-5, 5]$. Class 2 demand is specified as;

⁵⁵Berg and Tschirhart (1988), p. 227.

⁵⁶The Hamlen and Jen (1983) model is similar in construction to the Tschirhart and Jen model except that instead of interrupting lower-ranked classes altogether some pre-fixed fraction of each consumer's demand is interrupted. This fraction is dependent upon ranking. Thus, Hamlen and Jen distinguish between the probability of interruption and the extent of interruption, both of which affect consumer demand differently. Hamlen and Jen also generalize the stochastic demand component of the Tschirhart and Jen model to all consumer groups, not just the first class. One of the more interesting results of the Hamlen and Jen model is that the (expected) profit-maximizer will always offer a lower price to groups with a greater extent of interruption, while the expected welfare maximizer may not. (Hamlen and Jen, (1983), p. 1110).

$$X_2(p_2) = 25 - p_2 \quad (4.2)$$

such that the stochastic element of demand, (μ) , is captured only by class 1⁵⁷ and demand is independent of reliability levels.⁵⁸ We assume only one interruptible class of demand here so that only one possible class ordering must be considered by the firm.

The availability of capacity is dependent on the magnitude of realized demand, so available capacity is denoted as;

$$h(k) = K - \sum_{i=1}^k X_i, \quad k = 1, 2 \quad (4.3)$$

where K is the installed capacity and k is the number of demand classes. Once the demand class X_i has been served, $h(k)$ is the remaining capacity available to serve classes $(k - i)$. In this one period model the firm maximizes (expected) profits of:

$$\begin{aligned} E[\pi] = & \int_{\underline{\mu}}^{h(1)} [(p_1 - c)(X_1 + \mu)f(\mu)]d\mu + \mu \int_{h(1)}^{\bar{\mu}} [(p_1 - c)K \cdot f(\mu)]d\mu + \\ & \int_{\underline{\mu}}^{h(2)} [(p_2 - c)X_2 \cdot f(\mu)]d\mu + \int_{h(2)}^{h(1)} [(p_2 - c)(K - \mu - X_1)f(\mu)]d\mu - \beta \cdot K \end{aligned} \quad (4.4)$$

We assume constant per-unit operating and capacity costs of c and β , K is the installed capacity, and μ is the stochastic term. The stochastic term is distributed uniformly over the interval $[\underline{\mu}, \bar{\mu}]$. In the above equation the expected profits from the first class is the sum of the first two terms. The first term is operating revenue when class 1 is fully served, and the second is revenue when this class is partially served. Similarly, the third and fourth terms are the operating revenues from the interruptible class when this class is fully served and when is it partially served. The limits of integration are lower for this class because class 2 demand is served only once class 1 is fully served. The last term is capacity costs. The firm maximizes profits as given by (4.4) over the choice variables p_1 , p_2 , and K .

Assume marginal operating costs and capacity costs of;

⁵⁷As in the Brown and Johnson model the stochastic element enters additively.

⁵⁸This last assumption is changed later on.

$$\begin{aligned} c &= 2 \\ \beta &= 15 \end{aligned} \tag{4.5}$$

The optimal capacity is 9.420, with a price of service for class 1 of 23.38 and a corresponding reliability of 0.780. The price for class 2 is 18.91, with a corresponding reliability of 0.172. Given these values the level of expected consumer surplus is 73.70 and expected profits are 350.18.

Notice that in the numerical example of the Tschirhart and Jen model, the prices are indeed sequential; the price is higher for the service class with the higher reliability. Comparing the (absolute) elasticities, we can see that these prices conform to the inverse elasticity rule,⁵⁹ and the (relatively) more elastic demand is charged lower prices:

$$18.91 = p_2 < p_1 = 23.38 \quad \text{and} \quad 3.67 = e_1 < e_2 = 6.53 \tag{4.6}$$

where ε_i is the expected price elasticity. In order to find these expected elasticities for each class, we must use the expected quantity consumed by each class, using a formulation similar to that as in equation (4.4), because sales do not always equal the quantity demanded when demand is stochastic. Thus we can see from (4.6) that the sequential pricing for this solution conforms to the order of the own price elasticities of demand. This is similar to a multi-product firm which prices the product with the more elastic demand lower than other products with less elastic demands. Notice also that the optimal price for the first class does exceed the sum of marginal operating and capacity costs, $(c + \beta)$, as expected, and the price for the second class exceeds marginal operating costs. This ensures that both classes contribute to profits.

Our discussion of the Tschirhart and Jen model thus far indicates that there will be welfare gains when interruptible pricing supplants uniform pricing. In order to create a clear picture of these welfare gains, we will compare the numerical results of the Tschirhart and Jen model with that of the profit-maximizing Brown and Johnson model. This comparison is possible since we assume an aggregate demand of,

$$D(p) = 55 - p + \mu \tag{4.7}$$

⁵⁹Berg and Tschirhart (1988), p. 226.

in the Brown and Johnson model, which is the additive demand of (4.1) and (4.2). Substituting this into the example developed in Chapter 3 and optimizing over p and K produces expected profits of 319.33 and an expected total welfare of 512.33 in the Brown and Johnson profit-maximizing extension. Notice that expected profits increase with the introduction of service differentiation. The firm benefits by more closely matching services with the needs of consumers with the offering of priority service. Consumer surplus is higher in the Brown and Johnson model than in the Tschirhart and Jen model because the former is a welfare-maximizing model, while the latter is profit-maximizing.

As expected the optimal level of capacity is lower in the Tschirhart and Jen model than in the Brown and Johnson extension. The optimal capacity in the Tschirhart and Jen model is 9.42 as compared to 20.02 in the Brown and Johnson model. This is because the firm can substitute interruptible service for expensive capacity additions while still supplying class 1 customers with the same level of reliability. In contrast, a firm offering a single level of reliability must rely on the *average* willingness to pay for capacity, as evaluated at the margin, to determine the optimal capacity level. If this firm increased the level of reliability, it must increase reliability to all customers, necessitating an investment in a larger capacity.

We now extend the numerical example developed above of the Tschirhart and Jen model to include reliability dependent demand. Altering the demand curves to include reliability, the demand functions now become;

$$X_1(p_1) = 30 - p_1 + r_1 + \mu \quad (4.8)$$

$$X_2(p_2) = 25 - p_2 + r_2 \quad (4.9)$$

With reliability dependent demand, the firm must ensure that quoted levels of reliability are consistent with actual levels, so that equation (4.4) is subject to the following feasibility constraint;

$$F(h(i)) = \int_{\underline{\mu}}^{h(i)} f(\mu) d\mu \geq r_i, \quad i = 1, 2 \quad (4.10)$$

where $F(h(i))$ is the quoted reliability for class i ,⁶⁰ and r_i is the actual level of reliability supplied to that class.

With these new demand functions, the firm will charge essentially the same prices for the two demand classes as with reliability-independent demand; $p_1 = 23.87$ for class 1, and $p_2 = 18.92$ for class 2, as compared to 23.38 for firm service and 18.91 for interruptible service with reliability independent demand. However, the reliabilities that each class receives increase to $r_1 = 0.85$ and $r_2 = 0.22$ because the optimal capacity increases to 10.46. Since changes in reliability now directly affect demand and, through demand, marginal revenues, the firm can increase profits by increasing reliability to both classes with a capacity expansion.

Notice that sequential pricing is optimal due to the elasticities of demand. As mentioned above, these prices need not be sequential if the reliability and price elasticities work against each other. That they are sequential, however, does tell us that the effect of the price elasticities outweighs the effect of reliability elasticities. Calculating these (absolute) elasticities yields:

$$\begin{aligned}\varepsilon_1 &= 3.48 < 5.63 = \varepsilon_2 \\ \eta_1 &= 0.124 > 0.065 = \eta_2\end{aligned}\tag{4.11}$$

where ε_i is the (own) price elasticity and η_i is the reliability elasticity using expected quantities. As we can see from equation (4.11) the firm charges higher prices to the less price elastic, more reliability elastic demand. While the ordering of demand classes is fixed in this example because there is only two demand classes, prices need not be sequential. Even though the first class will always be served with a higher level of reliability, it might be profitable to charge them a lower price if that class is relatively more price-elastic.

With reliability dependent demand, expected profits will fall dramatically to 50.02 and total welfare will also fall to 120.95. This is due to the dependence of demand on reliability. With reliability dependent demand the monopolist must consider the effect of a low quoted reliability on demand. The low reliability quoted to class 2 will now affect the demand for this service since demand is dependent on reliability. In order to provide higher levels of reliabilities to increase class 1 demand the firm must invest in expensive capacity expansions. Together these factors will effect total profits.

⁶⁰The quoted reliability for class i is defined over the range $[\underline{\mu}, h(i)]$ since class i is served in the range of the lower bound of the stochastic term ($\underline{\mu}$) to the level at which interruption of class i occurs ($h(i)$).

Again we can compare the Tschirhart and Jen example with reliability dependent demand to that of the Brown and Johnson model of a profit-maximizing monopolist with reliability dependent demand.⁶¹ The Brown and Johnson monopolist offers a uniform price of 36.00 with a uniform reliability of 0.985 and a capacity of 20.618. The Brown and Johnson firm offers a high level of reliability because of the gains in revenues through the increase in demand. Notice that, as expected the Brown and Johnson monopolist must invest in a higher level of capacity than in the interruptible model. The Brown and Johnson monopolist must invest in a much higher level of capacity because it does not have the option of interrupting its lower valued customers to serve higher valued customers in periods of shortages.

In the Brown and Johnson example expected profits are higher with reliability-dependent demand than with reliability-independent demand. Expected profits are now 337.09 as compared to 319.33 previously. Expected profits in the Brown and Johnson model are -339.57, as compared to profits of 50.02 in the Tschirhart and Jen model and to 319.33 previously. Expected total welfare is 181.76 in the Brown and Johnson model, as compared to 70.92 in the Tschirhart and Jen model, but this is primarily due to the price discriminating⁶² nature of the Tschirhart and Jen model, which captures more of consumer surplus as profits. We can summarize the numerical results of the Tschirhart and Jen model in the following table;

⁶¹Again here we must constrain quoted reliabilities to be consistent with actual reliabilities.

⁶²Again, the firm is discriminating more across products rather than prices, since each level of reliability can be viewed as a different good. In the truest sense, the price-discriminating firm discriminates across different customers through price differentials for the same good.

Table 4.2
Summary of Interruptible Pricing and Extensions

Case ⁶³	Price	Reliability	Capacity	Expected Consumer Surplus	Expected Profits
Tschirhart and Jen • Reliability Independent Demand ⁶⁴	$P_1 = 23.38$ $P_2 = 18.91$ $\varepsilon_1 = 3.67$ $\varepsilon_2 = 6.53$	$r_1 = 0.77996$ $r_2 = 0.17116$	$Z = 9.4206$	$E(CS) = 73.707$	$E(\pi) = 350.18$
• Reliability Dependent Demand ⁶⁵	$P_1 = 23.87$ $P_2 = 18.92$ $\varepsilon_1 = 3.48$ $\varepsilon_2 = 5.631$	$r_1 = 0.84812$ $r_2 = 0.21832$ $\eta_1 = 0.124$ $\eta_2 = 0.065$	$Z = 10.462$	$E(CS) = 70.927$	$E(\pi) = 50.021$

As we have seen above with the comparison of the welfare results of the profit-maximizing Brown and Johnson model with the Tschirhart and Jen model, the introduction of service differentiation provides gains in profits over uniform pricing. The above analysis, however, assumes that the demand for service with the highest reliability is independent of the demand for interruptible service. If customers expect excess capacity, the demand for interruptible service will increase with this expectation. With excess capacity, customers pay a lower price with interruptible service for roughly the same amount of reliability that they would receive if they had purchased firm service. Conversely, if customers expect excess demand we could expect the demand for higher priority service to increase. With excess demand the reliability level of interruptible service is very low, and so if customers value reliability they will pay a higher price and purchase service with a higher reliability. Thus, firm and interruptible service are substitutes and there exists an incentive problem such that customers will substitute away from firm service when they expect periods of excess capacity. The firm must then take into account this incentive problem when choosing the optimal capacity, ordering and price bundles. One way to build this incentive problem into the price menu is to allow self-selection among

⁶³Assuming $b = 2$ and $\beta = 8$, and firm demand of $D_1(p_1) = (30 - p_1 + \mu)$ and interruptible demand of $D_2(p_2) = (25 - p_2)$.

⁶⁴Assuming a profit-maximizing firm and that $b = 2$ and $\beta = 15$: β must be increased from 8 to 15 because at $\beta = 8$ the marginal willingness to pay for capacity is excess when compared to the marginal cost and a corner solution results. With $b = 2$ and $\beta = 8$, $p_1 = 19.88$, $p_2 = 17.113$, $r_1 = 1.2192$, $r_2 = 0.4305$, $Z = 17.312$, $E(CS) = 115.44$, and $E(\pi) = 136.47$.

⁶⁵These prices need not necessarily be sequential. That they are sequential, however, does tell us that the effect of price elasticities outweigh the effect of reliability elasticities.

service choices by consumers. In allowing consumers to select among the offered service classes self-arbitrage among customers must also be prevented. In order for the firm to supply the promised reliabilities customers must choose the service option designed for them and not some other option. As we will see in the next chapter one way to prevent this self-arbitrage is to design service tariffs based on customers' valuation of service so that those customers with the highest valuation of service pay the highest tariffs. While this prevents self-arbitrage among customers it also implements the efficient rationing order since those customers with the lowest willingness to pay are interrupted first.

CHAPTER FIVE

A MODEL OF PRIORITY PRICING

5.1 Introduction

While customers value reliability of service, they may be willing to accept degradation in the level of reliability in exchange for corresponding decreases in price. In effect we can think of a kind of demand function where consumers pay high prices for high levels of reliability and correspondingly pay low prices for low levels of reliability, with variations in between. In the Jen and Tschirhart model, there was one class of firm demand, served without interruption, and n classes of interruptible demand. One can imagine that there exists a kind of trigger-reliability between classes; that is, some level of reliability at which customers find it beneficial to switch to a different demand class because the price/reliability trade-off in their current class no longer maximizes their utility. Switching between demand classes could be triggered by changes in either price or reliability. If one imagines each level of reliability as a different commodity with a different price, one could see that there is some point at which certain reliability/price bundles are substitutes; it is at this point that the marginal consumer materializes. This marginal consumer is indifferent between two (adjacent) classes of priority. However, Tschirhart and Jen do not address this issue in their model. In their model customers are not allowed to switch between priority classes in response to changes in prices or reliabilities.

One way to incorporate this substitutability into the interruptible model is to allow self-selection of demand classes among consumers. Consumers, acting in their own self-interest, will select the demand class best suited to their needs. Given that the offered price/reliability menu is optimally structured, supply will be efficiently allocated through this self-rationing.

Tschirhart and Jen assume that there is full information in their model. The firm knows the preferences of each customer and can segregate customers into the appropriate demand class. This is possible because the firm knows the demand preference of each customer and because the firm can prevent arbitrage, the reselling of the commodity, between customers.

However, in general only the distribution of demand is known, while the preferences of each consumer are private information. Customers cannot simply be asked for their willingness to pay for service because consumers have a strong incentive to claim

that they have a low willingness to pay so that they will be charged a low price. With asymmetric (incomplete) information, customers cannot be prevented from choosing a demand class which was designed for another consumer. Clearly, consumers with high willingness to pay for reliability will want to purchase the price/reliability bundle offered to other consumers if this bundle offers the same reliability at a lower price. This introduces a "self-selection" or "incentive-compatibility" constraint so that customers choose the price/reliability bundle designed for them and not some other bundle.

One way to get consumers to reveal their true type or preference is through their selection from the price menu. The optimal price menu is designed such that only customers with higher willingness to pay select firm service, while customers with a lower willingness to pay select interruptible service. The higher type consumers are prevented from purchasing the consumption bundle designed for lower consumer types because lower type consumers are offered a "sub-optimal" bundle; that is, by offering them a lower reliability than they would otherwise purchase.

Personal arbitrage is the selection by consumers for allocations other than the allocation designed for them. Personal arbitrage violates the self-selection constraint described above. Personal arbitrage can be prevented by inducing interruptible consumers to select a lower reliability than they otherwise would. Arbitrage can be prevented either offering a very low reliability relative to price to interruptible customers,⁶⁶ or by offering a very high reliability relative to price to higher type customers.

The optimal price menu is designed to exactly ration available supply. The self-selection constraint is important because it helps ensure that the promised reliabilities can be supplied. If some customers do not choose the price/reliability bundle designed for them, there might not have enough capacity to supply customers with the level of reliability originally promised. It is essential that customers select the service class designed for them since demand for interruptible service allows higher reliability promised to firm customers to be supplied without investing in expensive capacity expansions. Notice the interdependence of demand in this model: if all customers choose the top priority, no

⁶⁶By excluding demands with low valuations of service, the social planner can improve the service reliability to higher valued demand. Chao et al (1988) points out that this exclusion is due to the efficiency losses caused by a finite number of priority classes. If the social planner was able to offer a continuum of priorities, it might be able to serve these low valued customers while still maximizing total welfare. Such a continuum is appealing because it would fully exploit the type of product differentiation that priority service offers. However, the implementation costs of such a system would be prohibitive. The number of classes implemented depends on the relative trade-off between the costs of adding another class and the gains of greater discrimination. In practice, the number of consumer classes is generally limited by the set of consumer characteristics and by transaction costs.

customer has top priority. Whether or not the promised reliabilities can be supplied ultimately depends on the level of reliability that consumers select.

For the model described in this section, we assume two classes of demand, firm and interruptible. Firm customers are charged higher prices for a higher reliability of service than interruptible customers. We assume in this model that firm service customers are subject to interruptions, although the probability of interruption for firm customers is less than that for interruptible service customers. Thus, the use of "firm" and "interruptible" to describe the classes of priorities is somewhat misleading.⁶⁷ If firm customers are not promised service with a reliability of 100 percent, the amount of installed capacity can be reduced, thereby saving on costs. This becomes more important the more expensive capacity is to install.

Consumers self-select between these two service offerings, acting to minimize expenditures subject to a utility constraint. This expenditure function approach is essentially equivalent to the utility maximization used by Tschirhart and Jen.⁶⁸ The optimal price menu is formulated to maximize total social welfare, the sum of consumer and producer surplus.

We compare the offering of two service offerings to the offering of a single level of service with a single uniform price. We will see that the availability of two service offerings achieves greater social welfare than does uniform pricing. Consumers benefit both from lower capacity costs and from the offering of service options which more closely match their price/reliability trade-offs. A smaller investment in capacity is possible with priority pricing than under uniform pricing because the reduced demand of interruptible customers "frees up" capacity, thereby enabling firm service to be supplied with a greater probability of service. Priority pricing exploits the heterogeneity of preferences in the consumer population. With uniform pricing and random rationing, there will be some consumers who prefer more (or less) reliability, and are willing to pay a higher (lower) price for this reliability. Intuitively, we can see that a closer match of service options with customer preferences will increase social welfare.

The model described in this section is a one period model. Within the period, the optimal capacity is chosen, and then the optimal price menu is designed subject to this

⁶⁷We do not assume that firm demand is met with a reliability of 100%. See Section 7.1 for a discussion of the implications of such an assumption.

⁶⁸See Varian (1992), p. 105, for a discussion of the identities that tie the expenditure function to the indirect utility function and the Marshallian and Hicksian demand functions.

capacity. Customers then select from the offered price menu, the state of the world is realized, and supply is allocated among customers.

The remainder of this chapter describes the model in detail. Section 5.2 outlines the specification of consumer preferences and derives consumer behavior while defining the marginal consumer. Section 5.3 specifies the characteristics of supply and derives the probability of service for firm and interruptible service. Section 5.4 sets up the social welfare maximization problem, and in Section 5.5 this solution is compared to that of uniform pricing. Capacity as a decision variable is added in Section 5.6, and finally, the model results are summarized and conclusions are drawn in Section 5.7.

5.2 Consumer Specification

We assume a general equilibrium model with two commodities; an "inside commodity", q , and a competitively supplied "outside" or composite commodity, the numéraire N . Consumers divide their income, Y , between these two commodities. We assume that all consumers have the following concave utility function;

$$\mu = (v + h^2)q + N \quad (5.1)$$

where μ is utility, v is a positive constant, h is an index for consumer preference or type, q is the quantity consumed of the inside commodity, and N is the numéraire commodity.⁶⁹ Notice that demand is non-stochastic; the stochastic element in this model comes from the fluctuation in available supply.

Heterogeneity in the customer population is represented by the taste parameter, h . The preference of each consumer is private knowledge known only to the consumer. This taste parameter is uniformly distributed across the total population of customers according to the density function $f(h)$ and the cumulative distribution function $F(h)$. Preferences are distributed uniformly on the continuous interval $[\underline{h}, \bar{h}]$, where, for simplicity, we assume the unit interval. The taste parameter ensures that customers have different

⁶⁹This utility function is quadratic in preferences. While a function which is linear in preferences is simpler to manipulate, it does not allow for a large enough improvement in social welfare with the introduction of interruptible service to avoid a corner solution. Because we assume uniform distributions throughout this model, with linear preferences the probabilities of service do not change enough with the switching of customers between classes to justify service differentiation.

willingness to pay for the commodity. The willingness to pay for the commodity by the lowest type consumer, for $h = 0$, is,

$$\mu = v \cdot q + N \quad (5.2)$$

while the willingness to pay for the commodity by the highest type, for $h = 1$, is,

$$\mu = (v + 1)q + N \quad (5.3)$$

The willingness to pay for service for consumers in between the highest and lowest types ranges between the two extremes represented by equations (5.5) and (5.6). We can see these equations that as h increases the benefit from the consumption of the inside commodity also increases.

We assume perfectly inelastic unit demand so that the consumer either buys one unit, ($q = 1$), or does not buy at all, ($q = 0$). The consumer will purchase the commodity if the utility received from purchasing is greater than the price, x ;

$$v + h^2 > x \quad (5.4)$$

The consumer will then pay the price x to gain $(v + h^2)$ of utility.⁷⁰ Using equation (5.1), we can see that if the consumer buys from the commodity, q , the utility received is,

$$\mu = (v + h^2) + N - x \quad (5.5)$$

where x is the price of service, and if the customer does not buy from the commodity, the utility received is,

$$\mu = N \quad (5.6)$$

since only the numéraire then is purchased.

Consumers minimize their total expenditures subject to the original utility level, $\bar{\mu}$, so that all consumers face the following problem;

⁷⁰We assume throughout that $v > x$.

$$\min_{q,N} \quad x \cdot q + N \quad s.t. \quad \bar{\mu} = (v + h^2)q + N \quad (5.7)$$

where x , the price of service, is p for firm customers and w for interruptible customers. Solving the constraint on the right-hand side of (5.7) for the numéraire and substituting it into the objective function on the left-hand side of (5.7) yields consumer expenditures,

$$E(\bar{\mu}, x) = \bar{\mu} - (v + h^2)q + x \quad (5.8)$$

We can easily see that this expenditure function is homogeneous of degree 1 in prices and is monotonically increasing in the original utility level, $\bar{\mu}$.

Equation (5.8) is customer expenditures if service was certain; that is, if customers were served with 100% reliability. If all customers were served with a reliability of 100%, there would be no trade-off between price and reliability and all customers would prefer the service with the lowest price, w^{71} , and a uniform price would be offered. However customers are offered some probability of service because supply is stochastic, and so they must minimize their expected expenditures.

We assume that customers pay prices p and w if they are served, and that they must incur some expenditure to maintain their original level of utility, $\bar{\mu}$, if they are not served. Because supply is subject to stochastic fluctuations, the expected expenditures of customers is the sum of two possible events; (1) the receipt of service and payment for this service, or (2) service interruption occurs and consumers must purchase more of the numéraire commodity. The expected expenditure for service will differ for firm and interruptible customers because they will be served with different levels of reliability.

Let r_f denote the probability of service for firm customers, and r_i the probability of service for interruptible customers, so that the *expected* expenditures for the two classes of service, assuming that all customers are risk-neutral, are as follows;

$$FE = [\bar{\mu} - (v + h^2) + p]r_f + (1 - r_f)\bar{\mu} \quad (5.9)$$

⁷¹There is also a problem in setting the price of service, p and w , so high so that consumers spend all their money on the numéraire good and none on the inside good. We assume that the willingness to pay for the good is sufficiently high relative to the operating cost that it is socially optimal to supply the good. Thus, we make the assumption that $v > c$, that the minimum willingness to pay, v , is greater than the marginal operating cost, c .

$$IE = [\bar{\mu} - (v + h^2) + w]r_i + (1 - r_i)\bar{\mu} \quad (5.10)$$

The first term in equation (5.9) and (5.10) is the expected expenditure if service is received, while the second term is the expected expenditure to maintain the original level of utility if interruption occurs. If interruption occurs, the consumer must consume enough of the numéraire commodity, N , to reach the original level of utility, $\bar{\mu}$. Thus, expected customer expenditure is the sum of the expected expenditure on service if there is available supply and the expected cost to maintain the utility level if service is interrupted.

All customers self-select between the two service classes, choosing the service class that minimizes their expected expenditures while maintaining their original level of utility. The marginal consumer, \hat{h} , is defined as the consumer who is indifferent between the two classes of service. The marginal consumer is defined where the expected expenditure for both service classes is the same;

$$FE(\hat{h}) = IE(\hat{h}) \quad (5.11)$$

Substituting in equations (5.9) and (5.10) in equation (5.11) yields;

$$(\bar{\mu} - (v + \hat{h}^2) + p)r_f + (1 - r_f)\bar{\mu} = (\hat{\mu} - (v + \hat{h}^2) + w)r_i + (1 - r_i)\bar{\mu} \quad (5.12)$$

Solving equation (5.12) for \hat{h} , the marginal customer can then be described by the following relationship;

$$\hat{h}(p, w, r_f, r_i) = \sqrt{\frac{p \cdot r_f - w \cdot r_i}{r_f - r_i} - v} \quad (5.13)$$

We can see from equation (5.13) that the marginal customer is a function of both the prices and the reliabilities of the two demand classes. The (expected) expenditure functions will shift in response to a change in price or reliability, and the value of the marginal customer will change as well. However, as we will see in Section 5.3, the probabilities of service also depend on \hat{h} . Thus, the design of the price menu is crucial for determining the subscription rates for interruptible service.

The marginal customer embodies the idea of self-selection in this model. While the marginal customer is indifferent between the two service classes, other customers are not. Firm customers, for example, elect to purchase firm service because this minimizes their expenditures subject to the utility constraint. If this were not the case, firm customers would purchase interruptible service instead.

From Figure 5.2 we can see that for preferences $h < \hat{h}$, (expected) expenditures for interruptible service are less than expenditures for firm service. Conversely, for preferences $h > \hat{h}$, (expected) expenditures for firm service are less than for interruptible. The marginal customer, \hat{h} , is defined where (expected) expenditures for the two service classes are equal. The expenditure function will shift with changes in prices and in probabilities of service. The expenditure functions will shift downwards with an increase in the price and will shift upwards with an increase in service probabilities. Changes in prices or reliabilities will change the expected value of service expenditures and the value of the marginal customer will also change in response.

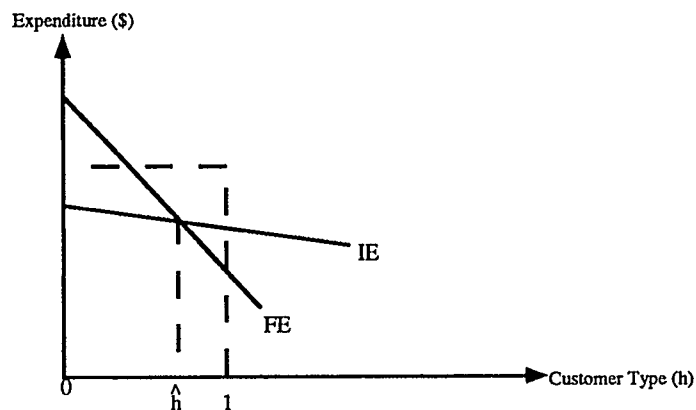


Figure 5.1
Expenditure Functions

With total demand normalized to 1 we can define the percentage of total demand which is served as interruptible as \hat{h} , while the percentage of total demand that is firm is $(1 - \hat{h})$. We can conceptualize total demand as in Figure 5.2, where firm and interruptible demand are measured along the unit interval $[0, 1]$,

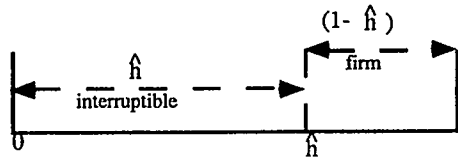


Figure 5.2
Unit Demand

The number of customers purchasing service for each class will depend on the probabilities of service that are offered. As the probability for a service class increases, the demand for this class will also increase. These probabilities, however, will in turn depend on the number of customers purchasing service for each class. For example, if many customers purchase interruptible service the probability of receiving firm service will be high. We assume throughout this model that when a certain level of reliability is offered consumers can confirm that this reliability has been delivered.

The substitutability between classes implicitly places an upper constraint on the prices of the two service classes. If the price of firm service is too low, for example, too many customers may choose this option and the firm may not be able to satisfy its reliability constraint. At the limit, all customers may choose firm service and there will only be one class of service.

The optimal price menu is designed to satisfy the self-selection constraint, that customers choose the price/reliability bundle intended for them. Consumers choose a service class on the basis of the offered service prices and reliabilities but, as mentioned above, these reliabilities in turn depend on the choices consumers make. For any given capacity the reliability of firm service falls as more and more customers choose subscription to this service class. In the next section we will explicitly derive these probabilities of service.

5.3 Supply Specifications

For simplicity, demand is assumed to be non-stochastic, while supply is assumed to vary stochastically due to outages. For example, outages could be due to changes in weather which cause random fluctuations in supply.

Outages are represented by the variable a and are uniformly distributed over the interval $[0, k]$, where k is the installed capacity. The total available capacity, I , is the supply which is on-hand after outages have affected the availability of installed capacity. Thus, total available capacity can be represented as;

$$I(k, a) = k - a \quad (5.14)$$

The installed capacity would be fully available only if no outages occur; that is, if $a = 0$. Conversely, if outages reach the upper bound of supply fluctuations, (if $a = k$), no capacity will be available.

Given that all levels of outages have an equally likely chance of occurring, outage fluctuations will have the positive, continuous density function,

$$f(a) = 1/k \quad (5.15)$$

The cumulative density function for this distribution would then be,

$$F(a) = \int_0^a \frac{1}{k} da = \frac{a}{k} \quad (5.16)$$

We assume that the extent of outages are common knowledge and are easily observable by all agents so that customers can form rational estimates of the probabilities of service for the two classes.

5.3.1 Probabilities of Service

5.3.1.1 Firm Service

In this section we will derive the explicit probabilities of service for each service class. As before, interruptible customers are served only if there is available capacity after all of firm demand is met.

The probability of firm service consists of two parts: (1) the probability that all firm demand is met, and (2) the probability that firm demand is only partially met. Full service for the class of firm customers is possible if available supply is at least as large as firm demand;

$$k - a > 1 - \hat{h} \quad (5.17)$$

and, rearrangement of equation (5.17) yields the inequality,

$$k + \hat{h} - 1 > a \quad (5.18)$$

Substituting the definition of a given in equation (5.18) in the definition of the cumulative density function given in equation (5.16), the probability that firm customers are fully served is:

$$P(k + \hat{h} - 1 > a) = \int_0^{k + \hat{h} - 1} f(a) da = \frac{(k + \hat{h} - 1)}{k} \quad (5.19)$$

This is the probability that outages are small enough so that all firm demand can be met.

The probability of the opposite event -- that the magnitude of outages outstrips firm demand -- is the probability that firm demand is greater than supply. Firm demand is greater than available supply when the following holds;

$$1 - \hat{h} > k - a \quad (5.20)$$

and rearrangement of equation (5.20) yields the inequality,

$$a > k - 1 + \hat{h} \quad (5.21)$$

We assume that when supply is insufficient to meet demand within a service class, the available supply is rationed randomly among all class customers. Thus, the available $(k - a)$ units of supply will be distributed among the $(1 - \hat{h})$ purchasers. Since outages are distributed along the interval,

$$a \in [(k - 1 + \hat{h}), k] \quad (5.22)$$

the probability of partial service given random rationing can then be represented as,

$$P(\text{random rationing} | k - a < 1 - \hat{h}) = \int_{(k-1+\hat{h})}^k \left[f(a) \cdot \frac{(k-a)}{(1-\hat{h})} \right] da \quad (5.23)$$

Substituting in equation (5.15) and integrating, equation (5.23) reduces to,

$$P(\text{random rationing} | k - a < 1 - \hat{h}) = \frac{(1-\hat{h})}{2k} \quad (5.24)$$

The total probability of firm service is then the sum of these two events; the event that firm demand is fully served, (equation (5.19)), and the event that a firm customer is served, given that there is partial service and random rationing within the demand class, (equation (5.24)). We denote the total probability of firm service as r_f ,

$$r_f = P(k - a > 1 - \hat{h}) + P(\text{random rationing} | k - a < 1 - \hat{h}) \quad (5.25)$$

Substitution yields,

$$r_f = \frac{(1-\hat{h})}{2k} + \frac{k + \hat{h} - 1}{k} = \frac{2k + h - 1}{2k} \quad (5.26)$$

5.3.1.2 Interruptible Service

In contrast to firm service, interruptible customers are served only if there is available capacity once firm demand is met. There will be full service of interruptible demand when,

$$k - a - (1 - \hat{h}) > \hat{h} \quad (5.27)$$

Rearrangement of equation (5.27) yields the inequality,

$$k - l > a \quad (5.28)$$

Substituting the definition of a given in equation (5.28) in the definition of the cumulative density function given in equation (5.16), the probability that interruptible customers are fully served is;

$$P(k - l > a) = \int_0^{k-l} f(a) da = 1 - \frac{l}{k} \quad (5.29)$$

Interruptible demand will be met only after firm demand is satisfied. Thus, interruptible demand will not be filled at all when supply is insufficient to meet firm demand;

$$l - \hat{h} > k - a \quad (5.30)$$

There will be partial service of interruptible customers if outages are in-between the two extremes of equations (5.30) and (5.28), such that;

$$k - l < a < k + \hat{h} - l \quad (5.31)$$

Part of interruptible demand will be served if outages are such that firm demand can be served, but available supply is still insufficient to serve all of interruptible demand. Again it is assumed that when there is partial service to the interruptible class, there will be random rationing of supply within this class such that the available supply, $k - a - (l - \hat{h})$, is rationed among the \hat{h} interruptible customers when outages are distributed along the interval,

$$a \in [(k - l), (k + \hat{h} - l)] \quad (5.32)$$

and the probability of partial service given random rationing for interruptible customers can be described as;

$$P(\text{random rationing} | k-1 < a < k + \hat{h} - 1) = \int_{k-1}^{k+\hat{h}-1} \left[f(a) \left(\frac{k-a-1+\hat{h}}{\hat{h}} \right) \right] da \quad (5.33)$$

Using the definition of the density function given by equation (5.15), this reduces to,

$$P(\text{random rationing} | k-1 < a < k + \hat{h} - 1) = \frac{\hat{h}}{2k} \quad (5.34)$$

As for firm service, the probability of interruptible service, r_i , is then the sum of the probability of these two events: (1) the interruptible class is fully served, (2) that the interruptible class is partially served. Using equations (5.34) and (5.29), the probability of serving interruptible demand therefore can be described as;

$$r_i = \frac{k-1}{k} + \frac{\hat{h}}{2k} = \frac{2k + \hat{h} - 2}{2k} \quad (5.35)$$

From equation (5.35) and (5.26) we can see that the probabilities of service for both classes explicitly depend on the level of installed capacity, as well as on the number of customers choosing each class, as represented by \hat{h} . Consumer behavior is therefore defined by the three equations; the probabilities of service, (5.35), (5.26), and the marginal customer (5.13).

By construction, the probability of firm service is greater than the probability of interruptible service;

$$r_f > r_i \quad (5.36)$$

$$\frac{(1-\hat{h})}{2k} + \frac{k+\hat{h}-1}{k} > \frac{k-1}{k} + \frac{\hat{h}}{2k} \quad (5.37)$$

which reduces simply to $-1 > -2$, which necessarily holds true. A higher probability of service is offered to firm customers than to interruptible customers. Again, due to the self-selection of service, there will be a dependence between the probabilities of service. The more customers who elect to participate in firm service, the lower will be the probability

that interruptible customers are served because firm customers are served with a higher priority than are interruptible customers. Conversely, the more customers who elect to purchase interruptible service, the higher will be the probability that firm customers are served. At the limit, if all customers select firm service, no customer has top priority.

5.4 Social Welfare Maximization

The social planner must maximize total welfare, the sum of consumer and producer surplus. Producer surplus is simply profits and we use compensating variation to calculate the level of consumer surplus. Consumer surplus is then the difference between (expected) expenditures under uniform pricing with random rationing and (expected) expenditures when service options are offered, given the probabilities derived in the previous section and the definition of \hat{h} . The maximization problem is therefore;

$$\max_{\hat{h}, k} TW = WTP + \pi \quad (5.38)$$

$$s.t. \quad r_i = \frac{k-1}{k} + \frac{\hat{h}^2}{2k^2}, \quad r_f = \frac{(1-\hat{h})^2}{2k^2} + \frac{k+\hat{h}-1}{k}, \quad \hat{h} = \sqrt{\frac{p \cdot r_f - w \cdot r_i}{r_f - r_i} - v}$$

Expected profits, π , are as follows;

$$\pi = (p-c)(1-\hat{h})r_f + (w-c)\hat{h} \cdot r_i - \beta \cdot k \quad (5.39)$$

where p and w are the per-unit prices of firm and interruptible demand, respectively, c is the marginal constant operating cost per unit of supply, and β is the capacity cost per unit of capacity.

The first term in equation (5.39) is the contribution to the firm's profits from firm sales, while the second term is the contribution from interruptible sales. The last term represents total capacity costs. Capacity is expensive; otherwise enough capacity would be installed to meet maximum demand. The marginal capacity cost is assumed to be larger than the marginal operating cost.

Total consumer surplus is the aggregate willingness to pay for service differentiation; that is, the difference between consumer expenditures before service differentiation less firm and interruptible expenditures after class differentiation;

$$WTP = \int_0^1 FE(\bar{p}, \mu) dh - \int_{\hat{h}}^1 FE(p, \mu, \hat{h}) dh - \int_0^{\hat{h}} IE(w, \mu, \hat{h}) dh \quad (5.40)$$

where μ is the original utility level, \hat{h} is the marginal customer, and \bar{p} is the original (uniform) price before service differentiation. We can use the compensating variation approach to measuring consumer surplus because we have assumed quasi-linear utility functions, and so there is no income effect. The first term in equation (5.40) is customer expenditures before service differentiation, while the second and third terms are the expenditures on firm and interruptible service after service differentiation. The difference between total expenditures for all customers before and after service differentiation is the total willingness to pay for service differentiation. We can see from equation (5.40) that if the offering of priority classes does not provide a saving on total expenditure, the willingness to pay for service differentiation will be non-positive. Conversely, the opposite is also true: if the offering of priority classes provides a lower expenditure on service, the willingness to pay for priority service will be positive.

Substitution of equations (5.9) and (5.10) in (5.40) yields (5.41);

$$WTP = \int_0^1 \left[(\mu - (v + h^2) + \bar{p}) \bar{r}_f + (1 - \bar{r}_f) \mu \right] dh - \int_{\hat{h}}^1 \left[(\mu - (v + h^2) + p) r_f + (1 - r_f) \mu \right] dh \\ - \int_0^{\hat{h}} \left[(\mu - (v + h^2) + w) r_i + (1 - r_i) \mu \right] dh$$

where \bar{r}_f is the probability of service with uniform rationing. This probability is a constant because it depends on the optimal capacity chosen under uniform pricing where as the optimal capacity of interest here is the capacity under uniform pricing.⁷² This probability is a constant for this problem because although the probability depends on the capacity

⁷²As we will see later on, with capacity $k = 1$, the probability of service under uniform pricing is simply $1/2$.

chosen, the relevant capacity is the capacity the minimizes expenditures under uniform pricing. Integrating equation (5.41) yields (5.42),

$$WTP = \left(\bar{p} - v - \frac{1}{3} \right) \bar{r}_f + \frac{1}{3} \hat{h}^3 \cdot r_i + \hat{h} \cdot r_i (v - w) - r_f (p - v) (1 - \hat{h}) - \frac{1}{3} r_f (\hat{h}^3 - 1) \quad (5.42)$$

Using equations (5.39) and (5.42) the objective function as defined by (5.38) reduces to the following;

$$TW = \bar{r}_f \left(\bar{p} - v - \frac{1}{3} \right) + \frac{1}{3} r_i \cdot \hat{h}^3 + \hat{h} \cdot r_i (v - c) - \frac{1}{3} r_f (\hat{h}^3 - 1) + r_f (1 - \hat{h}) (v - c) - \beta \cdot k \quad (5.43)$$

We can easily see that (5.43) is concave in \hat{h} . This concavity guarantees that an interior solution exists. The first and second derivatives with respect to \hat{h} are;

$$\frac{\partial TW}{\partial \hat{h}} = (r_i - r_f) (\hat{h}^2 + (v - c)) \quad (5.44)$$

$$\frac{\partial^2 TW}{\partial \hat{h}^2} = (r_i - r_f) 2\hat{h} \quad (5.45)$$

From (5.44) the condition for concavity is simply,

$$r_i < r_f \quad (5.46)$$

which is true by construction and is proven by equation (5.37).

Total welfare is the sum of the total willingness to pay for priority service and profits. The social planner maximizes total welfare given the choice variables, p , w , and k . In the next section, we derive the results from optimization.

5.5 Results

5.5.1 Pricing

A numerical solution can be found using this formulation of the interruptible model. First, a solution is found assuming that capacity is fixed. This would be the case in the

short run or with sunk capacity. This assumption removes the last term from our profit function in equation (5.39). By removing capacity from the set of choice variables, we are able to concentrate on the optimal pricing solution. For this solution, capacity is set at the value I , ($k = I$), the maximum value of customer preferences. Using (5.14), available capacity is defined as;

$$I = I - a \quad (5.47)$$

This assumption will affect the probability of service within each class. Given (5.46) we can derive the new probabilities of service for each class using (5.26) and (5.35).

$$r_f|_{k=I} = \frac{\hat{h} + I}{2} \quad (5.48)$$

$$r_i|_{k=I} = \frac{\hat{h}}{2} \quad (5.49)$$

Under uniform pricing, all customers are served if the available capacity is greater than total demand;

$$k - a > I \quad (5.50)$$

and with capacity fixed at 1 the probability that demand is fully served is then;

$$\overline{r_f}|_{k=I} = \frac{I}{2} \quad (5.51)$$

Substituting the probabilities of service, equations (5.48) and (5.49) and (5.50), into the objective function for non-uniform pricing, equation (5.43), produces a cubic in \hat{h} ;

$$TS|_{k=I} = \alpha + \frac{I}{6}(-\hat{h}^3 + \hat{h} + 3(v - c) + I) \quad (5.52)$$

where the constant, α , is defined as,

$$\alpha = \bar{r}_f \left(\bar{p} - v - \frac{1}{3} \right) \quad (5.53)$$

Maximizing (5.52) over \hat{h} gives the first-order condition,

$$\left. \frac{\partial TS}{\partial \hat{h}} \right|_{k=1} = \frac{-\hat{h}^2}{2} + \frac{1}{6} = 0 \quad (5.54)$$

which yields a value for the marginal consumer of 0.577. Because customer preference is distributed along the interval $[0, 1]$, we can interpret this value of \hat{h} such that 57 percent of the customer population elects to participate in the interruptible program. We can use this value for \hat{h} to find the probability of service for each class. Thus the probability of service for firm customers is 0.788 and the service probability for interruptible customers is 0.288.⁷³ Given these values for \hat{h} , r_f and r_i , equation (5.13) yields a relationship for the optimal price, p , for firm service as defined by the price for interruptible service, w ,

$$p = 0.366w + 0.275 \quad (5.55)$$

As we can see, there is no unique solution to this problem. For every value of w , there exists a corresponding value of p such that \hat{h} is held constant.

From (5.55) we can see that firm customers pay higher prices than do interruptible customers. Notice that this relationship is positive; as the price for interruptible service increases, so does the price of firm service, although by smaller increments. This is due in part to substitutability between the two classes of service. The increase in price of firm service is necessary to prevent too many interruptible customers from switching to firm service when the price of interruptible service increases. If the price menu is designed to be incentive-compatible, the self-selection by customers of service classes imposes a constraint on the relative prices of the service options. The price of firm service cannot be raised too high, or the price of interruptible service set too low, if benefits from the offering of priority service are to be realized. If the price of firm service is set too high relative to the offered reliability, all customers will choose the interruptible service option, and there will be one class of service. Conversely, if the price of interruptible service is set too low relative to the offered reliability, all customers will subscribe to interruptible service, and

⁷³These probabilities do not sum to 1 due to rounding.

again there will only be one class of service. As the price of interruptible service increases, more interruptible customers will switch to firm service. This will affect the probability of firm service, and so the price of firm service need not increase by as much as the increase in w since the lowered reliability will make this option less attractive to firm customers.

One unique solution is to set the price of interruptible service to the marginal operating cost, c . With the price of interruptible service set to zero the price of firm service is 0.275.

Notice that these price/reliability bundles are sequential, such that customers pay for higher levels of reliability. This condition is necessary for self-selection. While prices are sequential for this solution, the ordering of the priority classes is not a choice variable in this model as it was in the Tschirhart and Jen model. In this model, priority classes cannot be ordered non-sequentially because consumers cannot be ordered using price or reliability elasticities, since these elasticities are private information. As well, it is clear that the incentive-compatibility constraint would not hold with a non-sequential ordering; if a service class with a lower price and higher reliability was offered, all consumers would have an obvious incentive to switch to the lower priced, higher reliability service class. Non-sequential priority orderings can therefore only occur with full information and no arbitrage.

In order to find numeric values for profits and consumer surplus, we assume values for the constants:

$$v = 0.1 \tag{5.56}$$

$$c = 0 \tag{5.57}$$

For this solution, excluding capacity costs, the level of profits is 0.092 with a total surplus of 0.1808 and consumer surplus of 0.189.

We can easily see that the offering of differentiated service increases social welfare over uniform pricing with random rationing. We know that the offering of differentiated service increases social welfare in this model because the marginal consumer exists. If priority service did not increase social welfare, this marginal consumer would not exist because there would be no advantage from choosing one class of service over another; either $\hat{h} = 0$ or $\hat{h} = 1$. Another way to see that the introduction of priority classes increases social welfare is to look at the total willingness to pay for service differentiation, that is, the difference between expenditures before service differentiation less the firm and interruptible

expenditures incurred when different service classes are offered. The marginal consumer materializes because as long as the total willingness to pay is positive, customers will incur greater costs with uniform pricing to maintain their original level of utility than they will with priority service.

The model developed here can only show that there are gains in social welfare from moving from uniform pricing to priority pricing. Because the social planner in this model maximizes the difference in expenditures between the two pricing schemes it is not possible to draw a comparison with a uniform pricing model which minimizes total social expenditures. In the next chapter we develop parallel models for both uniform pricing and for priority pricing in order to draw a comparison between the two of the optimal capacity choice. These models will allow us to compare the optimal capacity levels for the two pricing strategies.

5.6 Capacity Choice: A Comparison with Uniform Pricing

The optimal capacity needed to serve all customers can be reduced when service is differentiated by reliability. The offering of interruptible service frees up capacity so that firm customers can be served with a higher reliability level without expensive investments in capacity. In this section we will develop a model for uniform pricing and for priority pricing which minimizes total social expenditures and allows the optimal level of capacity to be chosen. Having developed these models, we then can compare the optimal level of capacity for the two pricing strategies.⁷⁴

With capacity choice, the decision sequence is as follows: (1) the optimal level of capacity is chosen, (2) the optimal price/reliability menu given the optimal level of capacity is designed, (3) this menu is presented to customers, who then select among the offered services, and (4) the state of the world is realized, and profits and total welfare are calculated on the basis of available supply. While this process can be thought of sequentially, all four steps happen simultaneously in this one-period model.

⁷⁴We assume here that capacity is not so expensive that the social planner cannot offer firm service.

5.6.1 Uniform Pricing

With uniform pricing only one class of service with one level of reliability and one price will be offered. There are no marginal consumers in this case since there is only one service class. With outages distributed along the interval $[0, k]$ the density function is,

$$f(a) = \frac{1}{a} \quad (5.58)$$

as before. All customers who demand the commodity will be served if available supply is greater than demand,

$$k - a > l \quad (5.59)$$

since we have assumed unit demand. All customers are served when outages are such that;

$$k - l > a \quad (5.60)$$

Thus the probability of full service is simply,

$$r_f^u = \int_0^{k-l} f(a) da = \frac{k-l}{k} \quad (5.61)$$

If they are served, expenditures of,

$$\int_0^l (\mu - (v + h^2)) dh = \mu - v - \frac{l}{3} \quad (5.62)$$

are incurred, so the total expected expenditure when there is full service for all customers is,

$$FS^u = \left(\mu - v - \frac{l}{3} \right) r_f^u = \left(\frac{k-l}{k} \right) \left(\mu - v - \frac{l}{3} \right) \quad (5.63)$$

However, when available supply is less than demand such that,

$$l > k - a \quad (5.64)$$

supply will be rationed randomly among all customers demanding service. The total social expenditure when there is excess demand is,

$$PS^u = \int_{k-l}^k f(a) \left[\int_0^l (k-a) (\mu - (v + h^2)) dh + \int_0^l (l-k+a) \mu \cdot dh \right] da \quad (5.65)$$

since there will be partial service when outages are distributed along the interval $[(k-l), k]$. The first term in (5.65) is the expenditure for those customers who are served. These expenditures occur with the probability $(k - a)$ since the available supply, $(k - a)$, must be rationed among the unit demand. The probability of not receiving service is then simply $(l - (k - a))$. Those customers who do not receive service in periods of rationing must still incur expenditures of μ to maintain their original level of utility. Integrating (5.65) with respect to h and a yields,

$$PS^u = \frac{l}{2k} \left(2\mu - \left(v + \frac{l}{3} \right) \right) \quad (5.66)$$

The total expected social expenditure that the social planner must incur is then expenditures when there is full service, expenditures when there is partial service, and capacity cost,

$$E(TE^u) = FS^u + PS^u + \beta \cdot k = \frac{l}{2k} \left(2\mu - \left(v + \frac{l}{3} \right) \right) + \frac{l}{k} \left(\mu - \left(v + \frac{l}{3} \right) \right) + \beta \cdot k \quad (5.67)$$

Minimizing (5.67) over k yields the first-order condition,

$$\frac{\partial E(TE^u)}{\partial k} = \frac{3v + l - \mu}{2k^2} + \beta = 0 \quad (5.68)$$

Using value for v and c from (5.56), (5.57) and assuming that $\beta = 0.1$, equation (5.68) yields the optimal capacity level,

$$k = \sqrt{5(4\mu - 1.3)} \quad (5.69)$$

such that $\mu \geq 0.325$ so that k is non-negative. Notice that the optimal level of capacity depends on the original level of utility, μ . This is because the customers must incur expenditures to maintain their original utility level when they are interrupted. If this original level of utility is very high, the cost of maintaining the original utility level will be very high for interrupted customers and these expenditures can be minimized through the investment in a higher level of capacity. Thus, the optimal level of capacity is dependent on the original level of utility such that the optimal capacity increases as the original utility level increases. We will now develop a total social expenditure function similar to (5.67) for priority pricing so that we can compare the optimal capacity levels between the two pricing strategies.⁷⁵

5.6.2 Priority Pricing

Under priority pricing all customers receive service when available supply is greater than total demand,

$$k - l > a \quad (5.70)$$

so that all customers receive full service when outages are distributed along the interval,

$$a \in (0, k - l) \quad (5.71)$$

The probability that all customers are served is,

$$r_f = \int_0^{k-l} f(a) da = \frac{k-l}{k} \quad (5.72)$$

and the total expected social expenditure is,

⁷⁵Prices are not solved for in the models developed in 5.6.1 and 5.6.2 because from the point of view of total social welfare, price simply redistributes surplus from the producer to the consumer. Thus, in terms of total social welfare, price is irrelevant, and consumers will purchase service as long as the price is less than their willingness to pay for service, $p < v$, and service will be supplied as long as the price received is greater than the marginal operating cost, $p > c$.

$$FS = r_f \int_0^l (\mu - (v + h^2)) dh = \frac{k-l}{k} \left(\mu - v - \frac{l}{3} \right) \quad (5.73)$$

There is full service to firm customers when outages are such that full service to all customers is not possible but available capacity is still greater than firm demand,

$$k - l < a < k - l + \hat{h} \quad (5.74)$$

so the probability that rationing among interruptible customers occurs is,

$$r_f^f = \int_{k-l}^{k-l+\hat{h}} f(a) da = \frac{\hat{h}}{k} \quad (5.75)$$

Expected expenditures incurred by firm customers when they are fully served is then,

$$FS_f = r_f^f \int_{\hat{h}}^l (\mu - (v + h^2)) dh = \frac{\hat{h}}{k} \left((\mu - v)(l - \hat{h}) - \frac{l}{3}(l - \hat{h}^3) \right) \quad (5.76)$$

When firm customers are randomly rationed, the probability of service is,

$$\frac{k-a}{l-\hat{h}} \quad (5.77)$$

and the probability of interruption is then,

$$1 - \frac{k-a}{l-\hat{h}} = \frac{l-\hat{h}-k+a}{l-\hat{h}} \quad (5.78)$$

Expected expenditures when firm customers are rationed is then,

$$PS_f = \int_{k-l+\hat{h}}^k f(a) \left[\int_{\hat{h}}^l \left(\frac{k-a}{l-\hat{h}} \right) (\mu - v - \hat{h}^2) dh + \int_{\hat{h}}^l \left(\frac{l-\hat{h}-k+a}{l-\hat{h}} \right) \mu \cdot dh \right] da \quad (5.79)$$

where outages are distributed along the interval $a \in [k - l + \hat{h}, k]$. Inside the first integral of (5.79) we are integrating along h using Leibnitz's Rule⁷⁶. Interruptible customers are fully served when there is available supply greater than interruptible demand such that,

$$k - a - (1 - \hat{h}) > \hat{h} \quad (5.80)$$

so the probability that interruptible customers are served given that rationing occurs is simply,

$$r_i^f = \frac{k - a - (1 - \hat{h})}{\hat{h}} \quad (5.81)$$

and the probability that interruptible customers are not served is simply $(1 - r_i^f)$. The total expected expenditure for interruptible customers when there is partial service and random rationing is the sum of expenditures when interruptible customers are served and when they are not served,

$$PS_i = \int_{k-l}^{k-l+\hat{h}} f(a) \left[\int_0^{\hat{h}} \left(\frac{k - a - l + \hat{h}}{\hat{h}} \right) (\mu - v - \hat{h}^2) dh + \int_0^{\hat{h}} \left(\frac{a + l - k}{\hat{h}} \right) \mu \cdot dh \right] da \quad (5.82)$$

Interruptible customers are not served at all when outages are greater than firm demand,

$$a > k - l + \hat{h} \quad (5.83)$$

so that the total expected expenditure to maintain the original level of utility for interruptible customers when the class of interruptible customers is not served is,

$$NS_i = (1 - P(k - l + \hat{h} > a)) \mu \cdot \hat{h} = \left(\frac{1 - \hat{h}}{k} \right) \mu \cdot \hat{h} \quad (5.84)$$

⁷⁶See Sydsæter, 1981, Theorem 4.2, p. 175, for a description of Leibnitz's Rule.

Thus under priority pricing, total social expenditures are,

$$TE_p = FS + FS_f + PS_i + PS_f + NS_i + \beta \cdot k \quad (5.85)$$

We can summarize the range of outages for which the different demand classes receive service with the following figure;

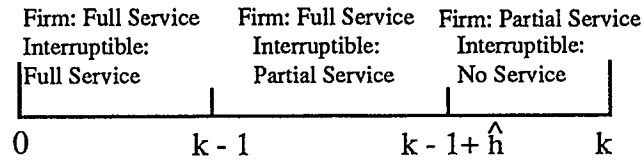


Figure 5.3
Priority Service

For outages along the interval $[0, (k - 1)]$, there is full service for both firm and interruptible customers. Along the interval $[(k - 1), (k - 1 + \hat{h})]$ there is full service for firm customers and partial service for interruptible customers since they are served only once firm customers are served. Finally, along the interval $[(k - 1 + \hat{h}), k]$ there is only sufficient supply to partially serve firm customers and interruptible customers go unserved. Substituting in equations (5.73), (5.76), (5.82), (5.84), (5.79) into the objective function (5.85) and minimizing over \hat{h} and k , yields expected expenditures of 0.314 and an optimal capacity of 1.235 with a marginal consumer of 0.577 as before, assuming a level of original utility of 0.5. If we assume the same level of utility so that we can compare across the two models, the social planner must expend resources of 0.374 and chooses an optimal capacity level of 1.871 under uniform pricing. We can summarize the numerical results of the models formulated in this chapter with the following table;

Table 5.1
A Summary of Results

Case	Price	Reliability	Capacity	Expected Consumer Surplus	Expected Profits
Priority Pricing⁷⁷					
- k=1	$w = 0.03^{79}$ $p(w) = 0.29$	$r_f = 0.789$ $r_i = 0.289$ $\hat{h} = 0.577$	$k = 1^{80}$	$E(CS) = 0.181^{81}$	$E(\pi(w)) = 0$
- k as a choice variable⁷⁸	$w = 0.12$ $p(w) = 0.27$	$r_f = 0.829$ $r_i = 0.424$ $\hat{h} = 0.577$	$k = 1.235$	$E(CS) = 0.1834$	$E(\pi(w)) = 0$
- Capacity Comparison uniform			$k = 1.871$	$E(TE) = 0.374$	
- priority pricing		$\hat{h} = 0.577$	$k = 1.235$	$E(TE) = 0.313$	

The offering of priority service allows a smaller investment in capacity than would otherwise be necessary under uniform pricing with random rationing. Interruptible demand, because it is served with a lower level of reliability than is firm demand, essentially "frees-up" capacity, allowing a smaller investment in capacity while still serving firm demand with a higher level of reliability than interruptible demand.

Notice also that the level of total social expenditures fall when priority pricing is introduced. The lower level of optimal capacity allows the social planner to avoid expensive capacity expansions and reduce expenditures. There is also a saving on expenditures because the offering of priority pricing more closely matches service options with customer preferences than does uniform pricing. This allows the social planner to

⁷⁷With marginal operating cost $c = 0$, constant $v = 0.1$ and $\beta = 0.1$. This is a social welfare maximizing model.

⁷⁸We can reformulate the model developed in Section 5.4 and allow capacity, (k), to be a choice variable. Trivially, social welfare will increase by a matter of course when we allow capacity choice. However, as noted above, given the formulation of this model, comparison with uniform pricing are not possible.

⁷⁹ w is set arbitrarily since there is no unique solution in this case.

⁸⁰Arbitrarily chosen.

⁸¹Assuming that capacity costs are not sunk.

expend fewer resources to compensate those customers with a high willingness to pay for service interruption.

With uniform pricing, because the preferences of consumers are heterogeneous, some customers will prefer a reliability which is lower than the offered reliability. These customers will choose interruptible service when they offered this choice, allowing a smaller investment in capacity than would be otherwise necessary. When offering priority classes, the costs of additional capacity must be traded-off against the costs to consumers of service interruption. While capacity is expensive, it is not necessarily true that some amount of interruption is optimal. If consumers have a very high willingness to pay for service relative to the cost of capacity it might be optimal only to have one class of service and a large amount of installed capacity. In our model this would mean $v \gg \beta$, where again β is the marginal capacity cost and v is the minimum utility to the consumer from receiving service. We assume here that v is small enough relative to β such that a trade-off exists.

One of the advantages of this model is that the self-selection by customers of service classes reveals information about (customers') valuation of capacity. This is one of the most important advantages of this model, because in the long run, the optimal capacity level must be chosen. If the cost of capacity is very large or very lumpy, the choice of capacity will have a significant impact both on profits and on the level of social welfare.

5.7 Conclusion

As we have seen in the previous sections, priority pricing offers significant improvements in social welfare over uniform pricing. Random rationing across all customers is clearly inefficient, as some customers suffer more than others from the interruption of service. In contrast, priority pricing more closely matches the needs of consumers with service options, thereby providing welfare gains. This matching of preferences also allows a lower investment in capacity, thereby significantly reducing total costs. The provision of a lower level of reliability to interruptible customers reduces the pressure on the capacity constraint while still allowing the provision of a higher reliability level to firm customers.

While the alternative model was formulated to model two classes of demand, firm and interruptible, it could easily be extended to model an array of demand classes, as in the Tschirhart and Jen model. Although a finer graduation of classes would add a certain

amount of complexity to the model, it would not significantly alter the model results. As we will see in the next chapter, this is especially true in light of the result that the bulk of the welfare gains resulting from the implementation of priority pricing are achieved through the offering of two or three service classes. The alternative model demonstrates the simplicity of priority pricing while embodying the two most important elements of the literature; self-selection and the revelation of the marginal valuation of capacity.

The priority pricing model developed in this section can be compared against other priority pricing models, such as Wilson's models of priority pricing. In the next chapter we will discuss further extensions of priority pricing, including; the linkage between priority pricing and product differentiation, spot markets, and insurance; the importance of rationing; two-part tariffs; and market organization. However, first we will briefly compare the priority pricing model developed above with the model which it most closely resembles -- the Woo and Toyama (1986) model.

CHAPTER 6

PRIORITY PRICING: FURTHER DISCUSSION

6.1 Introduction

In this chapter we will compare the model of priority pricing developed in Chapter 5 to a model of interruptible pricing which closely resembles it, the model developed by Woo and Toyama (1986). The results of these two models are quite different due to the different assumptions on which the models are based. We will then compare the priority pricing model to models of priority pricing developed by other authors, focusing on the model developed by Chao and Wilson (1987). This model differs in that a continuum of priority classes is offered instead of the two service classes offered in the model developed in Chapter 5. With a continuum of service classes priority pricing approaches perfect product differentiation because a different service class is offered to every customer. Because the offering of service classes then exactly matches the needs of consumers, priority pricing with a continuum of reliabilities realizes further gains in social welfare. This finer discrimination among customers is in part achieved with the introduction of two-part tariffs so that customers pay a fixed fee based on their valuation of reliability and a variable fee based on their consumption.

This fixed fee is the expectation of the service price had a spot or futures market existed. If a competitive market had been implemented after demand and supply had been realized, the spot price is the price that would exactly clear this market. Priority prices are simply the expectation of this spot price. Priority prices are preferable over spot prices because of the significant start-up and implementation costs in establishing a spot market. Priority pricing is also preferable if customers are risk-averse since customers can "hedge" the risk of interruption by choosing a higher price and a higher level of reliability. However, priority pricing provides only partial insurance against the risk of interruption, and for full social optimization, priority prices must be augmented by insurance. Finally, in this chapter we also briefly discuss various methods of implementing priority pricing, including demand subscription, service insurance and priority points.

6.2 Woo and Toyama (1986)

Like the model developed in Chapter 5, Woo and Toyama offer only two service options; firm service and interruptible service. Firm service customers are served with a higher reliability than are interruptible customers. However, unlike the priority pricing model, firm customers are not subject to interruption. Although both supply and demand fluctuate stochastically, in this model firm customers are supplied with a reliability of service of 100 percent. In this model customers do not choose between levels of reliability. They instead choose between certain service and uncertain service.

Woo and Toyama allow for asymmetric information. The preferences of consumers are private knowledge, known only to consumers. Consumers self-select between service options, choosing the service option which minimizes their expenditures while maintaining their level of utility. The total level of consumer surplus in this welfare-maximizing model is then the difference between total expenditures under uniform pricing and expenditures with interruptible pricing. Woo and Toyama also assume constant per-unit operating and capacity costs.

The optimal pricing strategy in the Woo and Toyama model is to charge peak-load prices. Woo and Toyama also find that the introduction of interruptible service does not alter the optimal capacity choice. These results are due to the assumptions on which the model is based.

In the priority pricing model developed in Chapter 5 it is assumed that firm demand is subject to interruption; firm demand is not served with a reliability of 100%. If service to firm customers is guaranteed in all possible contingencies, these customers must be served regardless of the magnitude of the supply outages. Recall that in order to fully serve firm customers, available capacity must be greater than firm demand;

$$(k - a) - (1 - \hat{h}) > 0 \quad (6.1)$$

If firm service is offered with 100% reliability, equation (6.1) must hold true for $(a = k)$, the maximum outage level. Substituting this into (6.1) yields,

$$\hat{h} - 1 > 0 \quad (6.2)$$

which cannot hold true in this model since we restrict \hat{h} to lie within the interval $[0,1]$. In the priority pricing model firm customers cannot be guaranteed service. It is not possible to provide 100% reliability if there exists some possibility, however small, that all capacity will be unavailable given that capacity is expensive. This assumption is one reason why the Woo and Toyama (1986) model of interruptible service does not provide an interior solution.

Given that firm customers are served with a reliability of 100%, Woo and Toyama mis-specify their production feasibility constraint. This constraint restricts firm demand to be less than available capacity. Woo and Toyama, however, formulate this constraint such that firm customers are actually served with a reliability of 100% only on average across all firm customers.⁸² A reliability level which is supplied on average is quite different from a reliability level which is supplied at all times for all customers. Turning again to the priority pricing model, the introduction of this assumption would then be reflected in the following constraint,

$$\int_{k+\hat{h}-1}^k [(k-a) - (1-\hat{h})] da \geq 0 \quad (6.2)$$

where the variables are as set out previously in Chapter 5. This constraint assumes that on *average*, firm demand must be met with 100% reliability. Assuming that this constraint is binding, we can integrate to get;

$$(k + \hat{h} - 1) \left(\frac{k - \hat{h} + 1}{2} \right) - \frac{k^2}{2} = 0 \quad (6.3)$$

This equality is true only if $\hat{h} = 1$; that is, if no customers subscribed to firm service. If the marginal customer is defined at the upper bound of customer preferences, there will be only one class of service and the offering of different priorities of service will be irrelevant. With one class of service, the firm must charge one price and all customers are subject to random rationing and the solution collapses to the Brown and Johnson solution. The optimal capacity would then be such that all customers can be served on average and the

⁸²Woo and Toyama (1986), p. 132, equation (21).

optimal price for such a large capacity would be a uniform price equal to the operating cost since this maximizes consumer surplus.

Woo and Toyama further assume that interruption of all interruptible customers occurs when all customers, the sum of firm and interruptible customers, cannot be served. There is no partial service to interruptible customers. For example, if total available capacity is 10 units and firm demand is 5 units, then this assumption implies that interruptible demand can only be met if it is no larger than 5 units; it is not possible that when interruptible demand is 7 units, only 5 units are served and 2 units are left unserved. Thus, there are only two possible supply scenarios in the Woo and Toyama model; (1) both firm and interruptible demand are served, and (2) firm demand is met but none of interruptible demand is served.

Interruptible demand serves only to utilize capacity that would otherwise sit idle when there is supply excess in of firm demand. Due to these assumptions, the optimal capacity in the Woo and Toyama model is much larger than in the priority pricing model and Woo and Toyama conclude that the optimal capacity is determined only by firm demand.⁸³ The introduction of interruptible service does not alter the reliability planning criterion. This would also explain the increase in welfare with the introduction of interruptible service in this model. If service to firm customers is never interrupted, the value of social welfare will increase by matter of course when interruptible service is offered because there will be gains in welfare when capacity is utilized which would otherwise lie idle. With this specification of service reliabilities, the needed capacity will be much larger than if all customers were subject to interruption because the investment in capacity must be large enough to serve firm customers regardless of the magnitude of supply outages. In this model, interruptible demand does not free-up capacity by reducing the demand on available capacity.

Woo and Toyama also find that the optimal prices are peak-load prices. Again, this conclusion makes sense in light of their assumptions regarding reliability of service. If interruptible customers are only served when there exists significant excess capacity, the optimal price would just cover operating costs in order to utilize as much excess capacity as possible. Capacity costs would then be covered by the higher prices charged to firm customers.

⁸³Woo and Toyama (1986), p. 135.

In the Woo and Toyama model and in the priority pricing model, only two levels of service reliability are offered; firm and interruptible. Given heterogeneity of preferences among consumers, there are obvious welfare gains in offering a finer gradation of service classes, as was done in the Tschirhart and Jen model. The interruptible model of self-selection described here and in Chapter 5 can be extended through the offering a wider spectrum of service classes.

6.3 Extensions to Priority Pricing

The alternative model set out in Chapter 5 is a simple model of priority pricing with two service classes. As was demonstrated numerically in Chapter 5, priority service provides gains in social welfare over uniform prices with random rationing. Chao et al (1988), Chao and Wilson (1987), and Wilson (1989a) obtain similar results for a priority pricing model with a continuum of priority levels.⁸⁴ We will focus here on the model developed by Chao and Wilson.

In the Chao and Wilson model, demand is stochastic. Again, each customers' valuation of service is private information, known only to the consumer. In their model, Chao and Wilson offer a continuum of service options so that every consumer is offered an individual service option designed specifically to meet their needs. There is no random rationing because every customer is served with a different level of reliability, with the result that customers are perfectly rationed during supply shortages. This perfect rationing yields gains in social welfare because every customer is ranked and interrupted according to their valuation of service. In addition to these welfare gains, priority pricing saves on rationing costs. Rationing costs are minimized with priority pricing because priority pricing determines the order of service before each contingency is realized. The menu of service options is offered to consumers before the state of the world is realized. It is less costly to establish a single rationing order which holds for all contingencies than to establish a rationing order for each possible contingency.

A continuum of service classes allows for a greater efficiency of rationing and greater welfare gains because the preferences of each customer can be exactly matched with a corresponding reliability. The addition of extra priority classes increases social welfare because it allows for a better utilization of capacity. If there is some possibility of excess

⁸⁴Viswanathan and Tse (1989) also discuss the connection between priority pricing model with finite and infinite classes.

supply an additional priority class at a lower level of priority can be offered without investing in additional capacity. Thus, additional service at a lower priority can be supplied without affecting the reliability and surplus to higher priority consumers. The offering of a continuum of service classes will also help to prevent self-arbitrage. With a continuum of priority classes every customer can select exactly that service class which best meets their needs, and so arbitrage is unnecessary. Finite priority classes are inefficient as compared to continuous classes because if a class cannot be fully served, supply is rationed randomly within that class. However, such a continuum will not be possible for any practical implementation of priority pricing due to the costs involved in the creation of such a continuum.

Chao et al, Chao and Wilson, and Wilson (1989a), derive the welfare losses due to the offering of a finite number of discrete reliability levels. In general they find that the welfare loss due to a finite number of priority classes is of order $1/n^2$, where n is the number of priority classes.⁸⁵ This result suggests that most of the welfare gains achievable with a continuous ordering of priority service can be obtained with two or three priority classes. Thus, the model described in Chapter 5 achieved 75 percent of the possible welfare gains with the two classes of service offered.

In Chapter 5 we have shown that the offering of priority service increases social welfare over uniform pricing. Chao and Wilson (1987) show that priority pricing is Pareto superior to a random rationing scheme with a fixed price because the efficiency gains that are realized can be distributed to increase every customer's expected net benefit without affecting customer's incentives to self-select efficiently. One reason that priority pricing results in welfare improvements is that it provides incentives for the firm to increase output under all capacities, inducing the firm to better utilize its capacity.⁸⁶ If there is some possibility of excess supply the firm can offer lower reliabilities of service without investing in additional capacity because by supplying additional service at a lower priority the firm does not affect the level of reliability and surplus to higher priority consumers.

Another reason why welfare increases under priority pricing is that a wider range of service options are offered. Priority service can be viewed as a form of product differentiation⁸⁷ because demand is segmented into a range of priority classes. While

⁸⁵ Assuming linear demand. See Chao et al (1988), p. 86, and Chao and Wilson (1987), p. 910.

⁸⁶ Viswanathan and Tse (1989), p. 162.

⁸⁷ Priority pricing differs from "ordinary" product differentiation in that the reliabilities are obtained endogenously. The quality or reliability obtained for one particular customer depends on the number of

random rationing in effect imposes an uniform (expected) reliability, priority rationing provides a spectrum of reliabilities, each priced to induce an efficient selection by customers according to their service valuation. If different customers value reliability differently there will be welfare gains in offering different levels of reliability. The closer the match in the offered reliabilities and preferences, the greater will be the gain in welfare, and so the greatest level of welfare will be achieved when every customer is offered exactly their preferred level of reliability. Furthermore, by matching service options more closely with preferences, customers are more efficiently rationed, and capacity is more efficiently utilized, thereby increasing social welfare.

6.3.1 Optimal Prices: Two-Part Tariffs

Each option in the priority menu in the Chao and Wilson model specifies a service or per-unit charge, $(s(v))$, and a priority charge, $(p(v))$ for the supplied level of reliability. These charges are based on the willingness to pay for service, v . The priority charge is a fixed charge, paid in advance, while the service charge is the variable charge for each increment of output, payable as service is received. Thus, priority tariffs can be thought of as a two-part tariff. The total tariff for each customer is then $[p(v) + s(v)]$.⁸⁸

The optimal variable charge is set equal to the marginal (short run) operating cost. This price ensures that as much capacity is utilized as possible. The social planner discriminates among customers through this priority charge. Each customer selects a priority charge based on their willingness to pay for service. Thus, consumers with higher willingness to pay for service select the higher reliability service option and pay higher prices. The selection by a customer of a high priority charge then reveals that that customer has a high valuation of service.⁸⁹ The priority charge selected by any one customer is the expected cost of compensation to all those customers receiving lower reliabilities in order to provide a higher level of reliability to this customer. The highest priority charge chosen by all customers is then the marginal valuation of capacity since this is the loss of consumer surplus to all other customers if they are interrupted. Customers' self-selection of priority options reveals important information about the distribution of customers' willingness to pay for service and capacity.

other customers selecting the same or higher priorities. With ordinary product differentiation, one customer's selection of product will not affect the offering of supply to other customers.

⁸⁸In this model, unit demand is assumed. (Chao and Wilson, 1987, p. 902)

⁸⁹Chao and Wilson (1987), p. 902.

If capacity is very costly or lumpy, the charging of fixed fees may be necessary to raise sufficient revenues in order to cover total costs. The tariff can then be directly constructed to meet net revenue requirements. The fixed (priority) fee is designed to recover capacity costs, while the per-unit service charge recovers operating costs, thus covering net revenue requirements. However, if an individual customer imposes no fixed costs on the producer, there then is no advantage from charging a fixed subscription or access fee. The fixed fee also determines the degree of market penetration, since it will prevent customers from entering the market who would otherwise make purchases had there been no fixed fee.⁹⁰

6.3.2 Optimal Capacity

In general, the implementation of priority pricing lessens the need to expand capacity. Priority pricing allows the substitution of low-priority demands having relatively low value to customers for expensive additions to capacity that would otherwise be required to maintain reliable service. The contracts for low-priority service essentially free-up capacity to meet supply shortfalls, thereby protecting the higher reliability expected from high-priority contracts.

Customers' self-selection reveals the marginal valuation of a service improvement from a capacity expansion, thereby allowing the selection of the optimal capacity. Without the revelation of this information the firm must otherwise depend on customers' valuation of average service reliability. Obviously, if only a single class of service is offered, customer selections will reveal little about their valuations of reliability. With the offering of several service classes, large demand for the highest priorities reveals that customers are willing to pay for extra capacity.

Capacity is expanded if the priority charge of the highest-priority customer is sufficient to pay for it,⁹¹ since this maximum charge measures the aggregate willingness to pay for improvements in reliability.⁹² This is because the price for reliability of service paid is the amount that one customer must pay to outbid customers obtaining lower service reliabilities.⁹³ This price is the amount that compensates all other customers selecting

⁹⁰See Tirole, p. 153.

⁹¹Wilson (1989a), p. 28.

⁹²This is also true since the profit from an incremental capacity unit is equal to the total surplus of the marginal buyer of that increment. Oren, Smith, and Wilson (1985), p. 558.

⁹³Wilson (1993), p. 239.

lower priorities whose service reliabilities were degraded by service to the higher priority customer. This price essentially captures the lost consumer surplus due to rationing at the given capacity level. The optimal capacity is set at the level where the marginal valuation of reliability is just equal to the marginal cost of capacity. This is similar to the optimal capacity chosen in the peak-load model. In the peak-load model capacity is expanded as long as the valuation of service by peak customers is sufficient to cover expansion costs.⁹⁴

For any given capacity a spot market will ration available supply efficiently. The optimal priority charge, however, not only reveals each customer's valuation of service, but it is also the expectation of the spot price for comparable service. In the next section we will explore the relationship between spot prices and priority prices and explain why priority prices are preferable.

6.4 Spot Markets

The fixed priority charge is the expectation of spot prices for comparable service. If a market had existed for the demand and supply that were realized, the spot price is that price which would exactly clear the market. The fixed fee in the two-part tariff is the expectation of the spot price had a spot market existed since it is the price that one customer must pay to outbid all other customers obtaining lower priority levels. Corresponding to each priority class is an imputed reservation price that is the maximum price at which the customer would make spot purchases. As the level of uncertainty decreases, priority prices and spot prices converge.⁹⁵

Priority pricing is preferable over spot markets when; (1) supplies are non-storable; (2) customers' valuation are stable over time;⁹⁶ and (3) transaction costs of a spot market are significant.

⁹⁴Oren, Smith and Wilson (1985) and Chao and Wilson (1987) derive the conditions for optimal capacity when there is multiple supply technologies and increasing costs. The optimal capacity for each technology is then where the expected benefit from the employment of another unit of technology is equal to its incremental installation cost.

⁹⁵However, revenues under the two schemes are the same only if marginal demand is stochastically independent of both aggregate demand and supply (Chao and Wilson (1987), p. 906). This is because revenue from spot pricing is the expectation of the spot price and spot demand, while revenue from priority pricing is the result of the expectation of the spot price with the expectation of the spot demand. As well, the spot demand and price tend to be correlated, which further increases the difference between the expected revenues of the two pricing schemes.

⁹⁶Customers preferences may change between the time when the priority menu is offered and when supply is allocated. Doucet (1994) addresses this problem by augmenting priority pricing service with a second-stage market implemented after the contingency has been realized. If customer's valuations have changed in

Priority pricing is preferable over spot markets when supplies are non-storable because priority pricing determines an *ex ante* rationing rule for any realized state of supply. If customers' valuations of service are stable over time, then this rationing rule need not be updated frequently and recontracting costs can be avoided. Because the priority menu needs only be offered once *ex ante* for each state of demand while a spot market must be implemented for each realization, priority pricing saves on implementation costs.

An added advantage of priority pricing over spot markets is that spot prices must be revised continuously while priority service contracts cover set time periods. While a spot market must be set up in all periods to reflect changes in demand or supply conditions, priority contracts establish prices for all contingencies. If customer preferences are serially correlated or persistent over time,⁹⁷ the gain in welfare from the establishment of a spot market in every period will be less than the cost. Priority pricing imposes fewer transaction costs on customers than do spot markets because of the high cost of continually monitoring spot prices and adjusting demand accordingly. In contrast, priority pricing is a predetermined rationing rule for the contract period. The optimal length of the priority contract period will depend on the correlation or persistence of customer preferences. In the extreme, if customer's valuations are invariant, permanent contracts could be offered. If customer's valuations are imperfectly correlated, limited period contracts will be more efficient. In general, the length of the optimal contract period is dependent on three factors; (1) the cost of recontracting; (2) the serial correlation of customer valuations; and (3) the number of priority classes. The more priority classes are offered, the shorter will be the contract period, because finer classes allow larger gains from more frequent contracting.

The finer the gradation of classes, however, the more expensive a priority menu will be to design and implement, and the more expensive will be the rationing costs. However, these costs can be minimized since most of the gains from priority pricing can be

the time period between choosing from the price menu and the realization of service, this second market will realize gains from trade among customers. This second-stage market would also be welfare-improving if customers were unsure at the time of contracting of their (later) preferences. Doucet (1994) examines a two-stage interruptible model where a spot market is implemented to facilitate gains from trade after demand is realized and supply is produced. These gains are possible because the ranking of the willingness to pay of consumers in this model varies through the contract period. It is unclear, however, if a spot market is indeed economically feasible why it would not be used in the first stage. There might also be incentive problems with this formulation; why would consumers subscribe to service in the first stage if they can be served in the second (at the time of interruption)?

⁹⁷This is particularly true if the stochastic element of demand is weather related, such as the demand for water or residential gas.

achieved using a small number of classes. Given that most of the welfare gains from the implementation of priority pricing can be implemented with a small number of priority classes, priority pricing can cheaply supplant the continuous variation in prices and monitoring required by a spot market.

Priority pricing is also preferable over spot markets if the spot market is very thinly traded. In a thinly traded market, market participants may not believe that the offered prices are an accurate reflection of the underlying probabilities. Such an illiquid market would be less likely with priority pricing because prices would be offered for all possible supply or demand realizations, not just for a particular realization, as in a spot market.

Another important advantage of priority pricing over spot markets is that customer's selections from the price menu reveal the benefit of capacity expansion. While a spot market is essentially an algorithm to determine the maximum reservation price in a particular contingency, it reveals only customers' valuations for that contingency. In contrast, the process of self-selection through priority pricing by consumers allows the inference of the allocation rule for all contingencies simultaneously. This is because priority service is generally offered as a forward contract over a longer period while spot prices must be revised instantaneously as supply and demand conditions change. In the long run, the integration of priority pricing with capacity planning allows the priority tariffs to reflect the associated capacity and operating costs.⁹⁸

6.5 Insurance

One limitation of all of the priority pricing models is the assumption that all customers are risk-neutral. If consumers were instead risk-adverse, a fully efficient pricing scheme would have to provide insurance in order to compensate customers for the costs of service interruption. Fair insurance is offered when premiums are based on the actuarial or true probability of interruption as offered in a competitive market. Optimally, such contracts would specify the payout of the contract once interruption had occurred as the difference between the marginal valuation of that unit of supply, v , and the marginal service charge s , which is the variable component of the two part tariff.⁹⁹ The premium which compensates the underwriter for the actuarial risk is the sum of actuarial risk and the

⁹⁸Chao and Wilson (1987), p. 901.

⁹⁹The payout is the same in all possible states of the world because customers are risk-neutral.

priority service charge.¹⁰⁰ Thus, the optimal premium is $(v - s)/[1 - r(v)]$, where $r(v)$ is the probability of service.

The insurance premium for each customer in the continuum priority model is based on the marginal willingness to pay for service for the reliability supplied to that customer. Those customers with higher service valuations and higher reliabilities of service will pay higher premiums. Given that the actuarial value of incremental insurance coverage for each increment in the service order is described in the insurance schedule, no cross-subsidization among risk classes is needed.¹⁰¹

If actuarially fair insurance rates are offered customers will purchase insurance. If contracts are traded continuously, risk-adverse customers can minimize their price risk by purchasing their contracts early. Because full insurance coverage equates the marginal utility of each customer across all contingencies, the efficient service order will result from customers minimizing their total payments.

Perfect insurance offered at actuarially fair premia will allocate supplies according to customer's valuations of service. Since the insurance premiums are the difference between the marginal valuation of service and the marginal priority charge, the provider of insurance has an incentive to minimize the compensation paid out to customers during shortages. The underwriter will interrupt those customers with the lowest valuations of service first because the compensation that must be paid to these customers is smaller than to other higher valued customers.

Priority pricing itself can be thought of as supplying partial insurance. A priority price can be decomposed into two parts; the price of the commodity, (the variable charge of the two-part tariff), and the price of insurance against being rationed, (the fixed charge of the two-part tariff). This latter price is the customer's valuation of reliability. Those customers valuing a high reliability will choose to pay a higher priority price. Thus, priority pricing can be thought of as providing insurance against shortages because it offers what are essentially contingent-forward contracts against interruption.

Notice, however, that priority pricing effectively ensures the customer only for the probability of getting a particular level of reliability and not against the event of not receiving service. For example, priority pricing can be compared to house insurance; there is a difference between insuring against the probability of your house burning down and the provision for compensation in the event that your house should actually burn down.

¹⁰⁰Chao and Wilson (1987), p. 909.

¹⁰¹Wilson (1989a), p. 27.

6.6 Market Organization

Chao and Wilson (1987) discuss three different implementations of priority pricing: (1) demand subscription;¹⁰² (2) service insurance; and, (3) priority points.

Demand subscription is the purchase by each consumer of several units of supply, each with a different valuation such that these units can be ranked. A base unit with a high valuation would be ranked higher than the marginal or peak unit with a lower valuation. By selecting different reliability levels and a corresponding tariff for each demand increment, customers have the option of receiving some minimum supply because their entire demand will not be subject to interruption.

However, this implementation assumes that the supplier has perfect information regarding customer preferences because with demand subscription the supplier is responsible for estimating the chances of interruption and for interpreting contractual obligations. A mis-specification of the price menu might have too few customers selecting low priority options to enable the provision of high priority service. In addition, the supplier might encounter difficulty in enforcing contractual arrangements if the contracts themselves are not specified in terms of observable events. Demand subscription has been used successfully in the electric power industry where utilities have built up knowledge regarding customer's preferences and where contracts can be easily specified in terms of observable outages. Demand subscription would be more difficult to implement for industries in which such knowledge has not been accumulated, or in industries which are limited by the level of available monitoring technologies, such as gas transportation, since gas entering the system cannot be directly traced.

Another form of priority pricing implementation is service insurance.¹⁰³ Customers purchase insurance expecting to be compensated for an interruption by an amount dependent on the premium paid in advance. In the event of a supply shortage customers with the lowest relative coverage would be interrupted first. This scheme differs from the supplementary insurance described above in that the supplier is not committed to

¹⁰²Demand subscription is similar to Spulber's reference point pricing and linear pro-rated service. See Spulber (1992, 1993).

¹⁰³A similar market organization would be Gedra and Varaiya's (1993) model of callable forwards contracts. A callable forward consists of two parts; a forward contract and a call. The forward contract, owned by the customer, guarantees the delivery of one unit of supply at a particular date. The call on the same unit of supply confers the right, but not the obligation, to purchase supply at a given price. This call portion of the contract can be sold by the consumer back to the supplier.

providing a certain probability of service although these probabilities are used to design the optimal insurance menu. Instead, the supplier is committed to the priority ranking as determined by the risk premium or interruption compensation offered in the insurance contracts. This scheme requires less monitoring and control than demand subscription since the supplier commits to the ranking determined by the risk premium outlined in the service contract but does not have to supply a guaranteed probability of service.¹⁰⁴

A third implementation of priority pricing is the provision of priority points. Priority points are offered and in the event of a supply shortage customers with the fewest points are interrupted first. This scheme is advantageous in that the creation of a market allowing customers to trade their acquired priority points relieves the supplier of the responsibility of developing a price menu and assessing the probability of interruption. Market transactions of these points will provide information about both the distribution of customer valuations and about the valuation of capacity.

However, as Chao and Wilson (1987) point out,¹⁰⁵ this scheme requires that consumers' valuations be rational in that their selection of points is based on reliability assessments which are consistent with probable events.¹⁰⁶ Customers must be informed and rational. One useful aspect of this scheme is that the supplier can vary the price of priority points to reflect daily or seasonal variation in demand and supply conditions. An alternative to the priority points market is the auctioning to brokers of a limited supply of points, with the price determined by market conditions. This market would provide an incentive to brokers to accurately assess probability distributions while enabling a market in futures contracts.

In summary, priority pricing efficiently rations available supply in periods of shortfalls because it ranks customers in order of their valuation of service. Spot markets would also perfectly ration available supply by allowing demand and supply to clear the market. However there are significant implementation and monitoring costs in operating such markets. Priority pricing minimizes these costs by offering forward contracts for service. Rather than contracting for immediate delivery, forward contracts are agreements with the obligation to deliver service at some future date. Priority tariffs can be broken down into two parts; a variable charge for quantity and a priority charge which is the

¹⁰⁴Probabilities of service would, however, be used to design the contract menu and to inform customers about the predicted consequences of contract choices.

¹⁰⁵Chao and Wilson (1987), p. 914

¹⁰⁶See Chao and Wilson (1987), Proposition 8.

consumer's valuation of service. This priority charge is the expectation of what the spot price would be in the events for which service is delivered under the selected priority.¹⁰⁷

The priority pricing models assume that consumers are not averse to the risk of service interruption. If consumers are not risk-neutral priority pricing must be supplemented with interruption insurance for a socially optimal outcome and customers must be compensated for the consequences of service interruption. Risk-averse consumers will prefer full risk coverage if insurance premiums offered at actuarially fair rates. This insurance can be offered separately or bundled with priority service. When bundled with priority service the supplier will still adhere to the efficient rationing order because it will want to minimize the compensation that it pays out. Even with risk-aversion, we can again see the robustness of the welfare gains in priority pricing.

¹⁰⁷Chao and Wilson (1987), p. 906.

CHAPTER SEVEN

CONCLUSION

While we have examined some of the extensions and issues related to the priority pricing model, one extension which has not been addressed in the literature is storage of supply and how it affects the optimal pricing and capacity rules. Recall that it is the uncertain nature of demand which makes a mix of technologies attractive. When demand is variable and output is non-storable, idle capacity becomes inevitable, making less capital-intensive technologies more attractive. The mix of technologies needed to minimize overall production cost may become more varied if storage of supply is possible. With uncertainty, storage will lower optimal peak prices and increase optimal off-peak prices, thus "smoothing" prices as the range of excess demand lessens. Storage in effect "hedges" against demand uncertainty, such that the supplier can draw upon storage reserves in periods of high demand and add to reserves in periods of low demand. Storage reserves can then be used to meet demand instead of rationing supply through prices. Although none of the models thus far discussed integrated the possibility of storage into the optimization problem, this possibility becomes more intriguing if priority pricing is applied to situations of stochastic demand and supply other than electricity,¹⁰⁸ such as water allocation, where storage is a more viable possibility. The introduction of storage also brings into play the interdependence between time periods since storage built up in one period can be used in the next.

In this thesis we have seen that the differentiation of service through priority pricing improves social welfare. The efficiency gains from the increased variety of service options are a direct result of the increased efficiency of the resulting rationing of scarce supplies and, in the long run, from the efficient provision of capacity. While we have seen these welfare gains in the literature and in the modeling, the question remains as to how applicable priority pricing is in the real world.

In an idealized world customers would respond instantaneously to spot prices which reflect the market conditions. However, the implementation and set-up costs of such a market, especially when there are many customers, is daunting. Through the offering of fixed period contracts priority pricing would avoid the monitoring and transaction costs involved in running a spot market. If these contracts are transparent and the reliability of

¹⁰⁸As mentioned previously, most of the literature on interruptible pricing is modeled on electricity markets.

service is easily verifiable there then will be efficiency gains in offering priority pricing. Certainly part of the challenge of offering priority pricing in the real world is in designing contracts which are both workable and easily understood by customers. These contracts do not have to be individualized since most of the welfare gains in offering differentiated service are achieved with two or three classes. We can see evidence of such contracts in the real world, such as with the transportation of gas where the pipeline offers an array of tariffs with various probabilities of interruption or auctions off priority rights for certain transportation right-of-ways.¹⁰⁹ With deregulation, many electrical utilities are also experimenting with service differentiation.¹¹⁰ In the increasingly competitive environment that utilities are now finding themselves in, the offering of innovative service options which are responsive to the different needs of customers provides utilities with another tool with which to capture new market segments.

In terms of practical problems, the implementation of priority pricing is limited by computer, communication and metering technology. The supplier must be able to ration supplies efficiently and effectively in order for priority contracts to provide benefits. There have been rapid technological advancements in the electricity industry with the growing use of non-utility generation. As other types of utilities are also forced to operate in a more competitive environment there might also evolve this focus on new monitoring technologies. In addition to these limitations, the supplier must have enough information about the distribution of customer preferences to design a workable priority menu and prevent self-arbitrage among customers. If the supplier has not built up enough information regarding its customer base the implementation of priority service might take several iterations of offered price menus before the optimal menu is found.

In this paper we have briefly addressed the different forms of market organization that priority pricing might take. The auctioning of priority points in a established market has intuitive appeal as it minimizes the informational requirements of the firm. The recent growth of electronic bulletin boards for the exchange of gas transportation contracts might be an indication that this form of market organization is preferable.¹¹¹ However, a market in priority points will not be efficient if the market is thinly traded or if participants are reluctant to participate.

¹⁰⁹These tariffs are of course more complex than implied here but a detailed discussion of the different pipeline tariffs is beyond the scope of this paper.

¹¹⁰Chao (1991).

¹¹¹For example, Pacific Gas and Electric are currently examining the potential of allowing an auction of priority rights at the line 400 interconnect.

The potential of priority pricing to produce welfare gains has been explored in the literature. The real test now is to see how the lessons learnt from the literature can be applied in the real world.

Bibliography

Amundsen, Erik Schrøder, and Balbir Singh. 1992. "Developing Futures Markets for Electricity in Europe." The Energy Journal, 13(3):95-112.

Arnott, Richard, André De Palma, and Robin Lindsay. 1993. "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand." American Economic Review, 83(1):161-79.

Berg, Sanford V., and John Tschirhart. 1988. "Pricing and Capacity Under Stochastic Demand." Natural Monopoly Regulation. Cambridge University Press; Cambridge, pp. 193-235.

Boland, John J. 1993. "Pricing Urban Water: Principles and Compromises." Water Resources Update, 92:7-10.

Brown, Gardner and M. Bruce Johnson. 1969. "Public Utility Pricing and Output Under Risk." The American Economic Review, 59:119-128.

Burness, H. Stuart, and James P. Quirk. 1979. "Appropriate Water Rights and the Efficient Allocation of Resources." American Economic Review, 69(1):25-37.

Burness, H. Stuart and Robert H. Patrick. 1991. "Peak-Load Pricing: Continuous and Interdependent Demand." Journal of Regulatory Economics, 3:89-106.

Carleton, Dennis W. 1977. "Peak Load Pricing with Stochastic Demand." American Economic Review, 67:1006-1010.

Carleton, Dennis W. 1978. "Market Behavior with Demand Uncertainty and Price Inflexibility." American Economic Review, 68:571-87.

Chao, Hung-Po. 1983. "Peak Load Pricing and Capacity Planning with Demand and Supply Uncertainty." Bell Journal of Economics, 14 (1): 179-90.

- Chao, Hung-Po, Shmuel S. Oren, Stephen A. Smith, and Robert Wilson. 1986. "Multilevel Demand Subscription Pricing for Electric Power." Energy Economics, 4:199-217.
- Chao, Hung-Po, Shmuel S. Oren, Stephen A. Smith, and Robert Wilson. 1988. "Priority Service: Market Structure and Competition." The Energy Journal, 9:77-103.
- Chao, Hung-Po, and Robert Wilson. 1987. "Priority Service: Pricing, Investment, and Market Organization." The American Economic Review, December, pp. 899-916.
- Chao, Hung-Po. 1991. "Reliability Pricing: A Theoretical Utopia?" Electrical World, February, p. 27.
- Coate, Stephen, and John C. Panzar. 1989. "Public Utility Pricing and Capacity Choice Under Risk: A Rational Expectations Approach." Journal of Regulatory Economics, 1:305-17.
- Colby, Bonnie G. 1993. "Reallocating Water: Evolving Markets, Values and Prices in the Western United States." Water Resources Update, 92:27-34.
- Crew, M. A., and P. R. Kleindorfer. 1976. "Peak Load Pricing with a Diverse Technology." Bell Journal, 7:207-31.
- Crew, M. A. and P. R. Kleindorfer. "Reliability and Public Utility Pricing." American Economic Review, March 1978, pp. 31-40.
- De Vany, A. S. and T. R. Saving. 1977. "Product Quality, Uncertainty, and Regulation: The Trucking Industry." The American Economic Review, 67 (4):583-594.
- Doucet, Joseph A., Kyung Jo Min, Michel Roland, and Todd Strauss. 1994. "A Two-Stage Mechanism to Improve Electricity Rationing" Paper read at the Canadian Economics Association Annual Meeting, June 10, 1994, Calgary, Alberta.

Doucet, Joseph A., and Michel Roland. 1993. "Efficient Self-Rationing of Electricity Revisited." Journal of Regulatory Economics, 5:91-100.

Feldman, Stephan. 1975. "On the Peak-Load Pricing of Urban Water Supply." Water Resources Research, 11(2):355-6.

Gedra, Thomas W., and Pravin P Varaiya. 1993. "Markets and Pricing for Interruptible Electric Power." IEEE Transactions on Power Systems, 8 (1):22-8.

Halverson, Philip Bernard. 1990. "Economic Impact of Interruptible Water Markets on Columbia Basin Project Irrigated Agriculture." Diss. Washington State University, Department of Economics.

Hamilton, Joel R., Norman K. Whittlesey, and Philip Halverson. 1989. "Interruptible Water Markets in the Pacific Northwest." American Journal of Agricultural Economics, 71(1):63-84.

Hamlen, W. A., Jr., and F. Jen. 1983. "An Alternative Model of Interruptible Service Pricing and Rationing." Southern Economic Journal. 49:108-1121.

Hanemann, W. Micheal. 1993. "Designing New Water Rates for Los Angeles." Water Resources Update, 92:11-21.

Harris, Milton, and Artur Raviv. 1981. "A Theory of Monopoly Pricing Schemes with Demand Uncertainty." The American Economic Review, 71:347-65.

Hogan, William W. 1992. "Contract Networks for Electric Power Transmission." Journal of Regulatory Economics, 4:211-42.

Howe, Charles W., and Mark Griffin Smith. 1994. "The Value of Water Supply Reliability in Urban Water Systems." Journal of Environmental Economics and Management, 26:19-30.

Howe, Charles W. 1993. "Water Pricing: An Overview." Water Resources Update, 92:3-6.

Joskow, Paul L. 1976. "Contributions to the Theory of Marginal Cost Pricing." Bell Journal, 7:31-40.

Kleindorfer, Paul R., and Chitru S. Fernando. 1992. "Peak-Load Pricing and Reliability Under Uncertainty." Journal of Regulatory Economics, 5(1):5-23.

Lane, M. N., and S. C. Littlechild. 1976. "Weather-Dependent Pricing for Water Resources in the Texas High Plains." Water Resources Research, 12(4):599-604.

Lee, Seong-Uh. 1993. "Welfare-optimal Pricing and Capacity Selection Under An Ex Ante Maximum Demand Charge." Journal of Regulatory Economics, 5:317-335.

Loury, Glenn, and Tracy R. Lewis. 1986. "On the Profitability of Interruptible Supply." American Economic Review, 76(4):827-30.

Marchand, Maurice. 1974. "Priority Pricing." Management Science, 20 (7):1131-1140.

Meyer, R. 1975. "Monopoly Pricing and Capacity choice under Uncertainty." American Economic Review, 65:326-37.

Morris, John R. 1990. "Pricing for Water Conservation." Contemporary Policy Issues, 8 (4):79-91.

Nieswiadomy, Micheal, and Steven L. Cobb. 1993. "Impact of Pricing Structure Selectivity on Urban Water Demand." Contemporary Policy Issues, 11:101-11.

Nguyen, D. T. 1978. "Public Utility Pricing with Stochastic Demands: A Note." Applied Economics, 10:43-47.

Oren, Shmuel, Stephen Smith, and Robert Wilson. 1985. "Capacity Pricing." Econometrica, 53 (3):545-66.

Panzar, John C., and David S. Sibley. 1978. "Public Utility Pricing Under Risk: The Case of Self-Rationing." American Economic Review, 68 (5):888-895.

Ravid, S. Abraham. 1992. "Reliability and Electricity Pricing." Journal of Economics and Business, 44:151-159.

Renzetti, Steven. 1994. "Water Pricing in Canada." Paper read at the Canadian Economics Association Annual Meeting, June 12, Calgary, Alberta.

Riley, John G. and Charles R. Scherer. 1979. "Optimal Water Pricing and Storage with Cyclical Supply and Demand." Water Resources Research, 15 (2):233-239.

Ring, Brendan, and Grant Read. 1994. "Short Run Electricity Pricing in Competitive Electricity Markets." Paper read at the Canadian Economics Association Annual Meeting, June 12.

Spulber, Daniel F. 1993. "Monopoly Pricing of Capacity Usage Under Asymmetric Information." The Journal of Industrial Economics, 59(3); 241-57.

Spulber, Daniel F. 1992a. "Optimal Nonlinear Pricing and Contingent Contracts." International Economic Review, 33(4):747-72.

Spulber, Daniel R. 1992b. "Capacity-Contingent Nonlinear Pricing by Regulated Firms." Journal of Regulatory Economics, 4:299-319.

Sydsæter, Knut. 1981. Topics in Mathematical Analysis. London: Academic Press Inc., Ltd.

Strauss, Todd, and Shmuel Oren. 1993. "Priority Pricing of Interruptible Electric Service with an Early Notification Option." Energy Journal, 14(2):175-96.

Tan, Chin-Woo, and Pravin Varaiya. 1993. "Interruptible Electric Power Service Contracts." Journal of Economic Dynamics and Control, 17:495-517.

Tirole, Jean. 1988. "Price Discrimination." In The Theory of Industrial Organization. Cambridge, Mass: MIT Press, pp. 132-68.

Tischer, A. 1993. "Optimal Production with Uncertain Interruptions in the Supply of Electricity: Estimation of Electricity Outage Costs." European Economic Review, 37:1259-1274.

Train, Kenneth E., and Nate Toyama. 1989. "Pareto Dominance Through Self-Selecting Tariffs: The Case of TOU Electricity Rates for Agricultural Customers." The Energy Journal, 10(11):91-109.

Train, Kenneth E. 1991. Optimal Regulation: The Economic Theory of Natural Monopoly. Cambridge, Mass: MIT Press.

Tschirhart, J. and F. Jen. 1979. "Behavior of a Monopoly Offering Interruptible Service." Bell Journal of Economics, 10:244-258.

Varian, Hal R. 1992. (Third Edition) Microeconomic Analysis. New York: W. W Norton & Co.

Viswanathan, N., and Edison T. S. Tse. 1989. "Monopolistic Provision of Congested Service with Incentive-Based Allocation of Priorities." International Economic Review, 30(1):153-74.

Visscher, Micheal L. 1973. "Welfare-Maximizing Price and Output with Stochastic Demand: Comment." American Economic Review, 63:224-229.

Visscher, Micheal, and Roger Sherman. 1978. "Second Best Pricing with Stochastic Demand." American Economic Review, 68:41-53.

Wilson, Robert B. 1989a. "Efficient and Competitive Rationing." Econometrica, 57(1):1-40.

Wilson, Robert. 1989b. "Ramsey Pricing of Priority Service." Journal of Regulatory Economics, 1:189-202.

Wilson, Robert B. 1993. Nonlinear Pricing. Oxford University Press, Oxford.

Woo, Chi-Keung. 1994. "Managing Water Supply Shortage: Interruption vs. Pricing." Journal of Public Economics, 54:145-160.

Woo, Chi-Keung. 1991. "Capacity Rationing and Fixed Cost Collection." The Energy Journal, 12(2):153-64.

Woo, Chi-Keung. 1988. "Optimal Electricity Rates and Consumption Externality." Resources and Energy, 10:277-292.

Woo, Chi-Keung, and Kenneth W. K. Lo. 1993. "Factor Supply Interruption, Welfare Loss and Shortage Management." Resource and Energy Economics, 15(4):339-52.

Woo, Chi-Keung, and Dewey Seeto. 1988. "Optimal Off-Peak Incremental Sales Rate Design in Electricity Pricing." The Energy Journal, 9(1):95-113.

Woo, Chi-Keung, and Nate Toyama. 1986. "Service Reliability and the Optimal Interruptible Rate Option in Residential Electricity Pricing." The Energy Journal, 7 (3):123-36.

Zarnikau, Jay. 1994. "Spot Market Pricing of Water Resources and Efficient Means of Rationing Water During Scarcity (Water Pricing)." Resource and Energy Economics, 16:189-210.

Zeitouni, Naomi, Nir Becker, and Mordechai Shechter. 1994. "Two Models of Water Market Mechanisms With an Illustrative Application to the Middle East." Discussion paper, Natural Resource and Environmental Research Center and Department of Economics, University of Haifa, Israel.

Zilberman, David, David Sunding, Richard Howitt, Ariel Dinar and Neal MacDougall.
1994. "Water for California Agriculture: Lessons from the Drought and New
Water Reform." Choices, First Quarter: 25-28.