Econometrica, Vol. 54, No. 5 (September, 1986), 1185-1195

CONSUMER INFORMATION IN MARKETS WITH RANDOM PRODUCT QUALITY: THE CASE OF QUEUES AND BALKING[1]

BY REFAEL HASSIN

We consider a revenue maximizing server who has the opportunity to suppress information on actual queue length, leaving demanders to decide on joining the queue on the basis of the known distribution of waiting times. We address the following second best problem: If suppression, but not pricing, can be socially controlled, is it socially optimal to prevent suppression? We show that it may be, but is not always, socially optimal to prevent suppression, and that it is never optimal to encourage suppression when the revenue maximizer prefers to reveal the queue length.

KEYWORDS: Consumer information, product quality, queues with balking.

1. INTRODUCTION

IT IS NOT UNUSUAL for the quality of a good sold by a given firm to vary considerably over time, while its price remains constant. This may happen when estimating the quality and charging each instant of the good differently involves high costs. The firm then faces the following dilemma: If it allows customers to examine the good, or supplies by itself the relevant information, it can expect variations in the demand and its rate of profit will change with the changes in quality. On the other hand, if the firm withholds this information from its customers, it can expect a steady demand, based on the statistical distribution of the good's quality.

Quality of service is especially likely to vary when that quality depends on the number of people who obtain the service. Such is the case with public transportation and with recreation facilities, where the greater the number of people obtaining service at a particular time, the lower benefit each person obtains, or the longer each customer must wait for service. In this paper we analyze a queueing system, where service quality is measured by the length of customers' wait. This case is of interest both because it is omnipresent and because there exists a well developed theory that can be used to obtain more definite results than can be expected in the general case.

We consider a revenue maximizing server who has the opportunity to (costlessly) suppress information on actual queue length, leaving demanders to decide on joining the queue on the basis of the known distribution of waiting times. The queueing model that we use has been analyzed by Naor (1969) and extended by Yechiali (1971), Knudsen (1972), Edelson and Hildebrand (1975), and Hassin (1985). Naor analyzed equilibrium in a single (costless) server queue where identical risk neutral demanders arrive according to a Poisson process and servicing takes exponential time. Where individuals decide whether to join the

[1] This research was supported by the Office of Naval Research Contract ONR-N00014-79-C-0685 at the Center for Research on Organizational Efficiency, Institute of Mathematical Studies in the Social Sciences, Stanford University, Stanford, California. My thanks to Amihai Glazer for his comments on an earlier draft of this paper.

queue by comparing the expected time cost of waiting with the benefit from being served, there is too much waiting relative to the social optimum (defined as the best steady state). In contrast, with a toll set by a revenue maximizing server, there is too little waiting relative to the social optimum. The model with suppressed information has been analyzed by Edelson and Hildebrand (1975), Littlechild (1974), and Balachandran and Shaefer (1980). In contrast to Naor's model, in this case the policy of a revenue maximizing server coincides with the socially optimal one.

In this paper we address the following second best problem: If suppression, but not pricing, can be socially controlled, is it socially optimal to prevent suppression? We feel that this is the right problem to address in this context since price regulation is in many cases impossible or undesirable while information suppression may be more easily prevented. In view of the results obtained by Naor and by Edelson and Hildebrand, the answer to this problem is by no means obvious, since although disclosure of the queue length prevents customers from joining very long queues, others will join the queue when it is socially preferred that they balk. We show that it may be (but is not always) socially optimal to prevent suppression, but that it is never optimal to encourage suppression when the revenue maximizer prefers to reveal the queue length.

In the next section we list our notation and describe the model. In Sections 3 and 4 we describe some preliminary results concerning queues with and without balking, respectively. In Section 5 we analyze the revenue maximizer's choice between the two options, and give conditions for making suppression revenue maximizing. In Section 6 we compare social welfare under revenue maximizing pricing with and without suppression. Section 7 contains a repetition of this discussion when the firm can establish more than a single facility, and in the final section we mention some related topics and make suggestions for future research.

2. THE MODEL

We adopt Naor's assumptions and notation. We consider a single server system where: (i) The potential demand for service consists of a stationary Poisson stream of risk neutral customers, with parameter λ (not all of this demand is actually served, as part of it may either arrive at the facility and balk or decide not to arrive at all, as will be explained in the next two sections). (ii) The station renders service in such a way that the service times are independently, identically, and exponentially distributed with intensity parameter μ. We write $\rho \equiv \lambda / \mu$. (iii) On successful completion of service, the customer obtains a reward of R, $R > 0$. (iv) The cost to a customer of staying in the system (either waiting or being served) is c per unit time, $c > 0$. We assume that $R \geqslant c/\mu$, since otherwise customers would not wait even for completion of their own service. (v) The queue discipline is first-come first-served.

When balking is possible, we also assume that the newly arrived customer is required to choose one of two alternatives; either he joins the queue (in which case it follows from the model's assumptions that he will never reverse his decision before completion of his service) or he refuses to join the queue, incurring no gain or loss. In the latter case we assume that the customer never returns to obtain the service (i.e., waiting "at home" for service is as costly as waiting in the system).[2]

We summarize now the main notation used throughout this paper.

λ: potential demand for service per unit time;
μ: rate of service;
R: reward obtained upon completion of service;
c: customer's waiting cost per unit of time;
$\rho = \lambda/\mu$: system's utilization factor;
$\nu_s = R\mu/c$;
n_s: maximum size of the queue when there is no admission toll;
n_o: maximum size of the queue under socially optimal toll;
n_r: maximum size of the queue under revenue maximizing toll;
Z_B^R, Z_{NB}^R: expected firm's revenue per unit of time under revenue-maximizing toll with and without balking, respectively;
Z_B^P, Z_{NB}^P: expected social welfare per unit of time *under revenue maximizing toll* with and without balking, respectively.

3. QUEUES WITH BALKING

In this section we first summarize some of Naor's results and then investigate revenue and welfare as functions of λ. If no admission toll is imposed, then a customer will join the queue only if the queue length, i, satisfies $(i+1)c/\mu \leq R$. The maximum possible length of the queue is thus $[\nu_s]$ where $\nu_s \equiv R\mu/c \geq 1$, and $[\nu]$ means the largest integer not exceeding ν. A revenue-maximizing firm will impose a toll of size $(c/\mu)(\nu_s - n_r)$, where n_r is the maximum queue length under this toll, defined by

$$(3.1)\qquad \begin{gathered} n_r = [\nu_r], \quad \text{and } \nu_r \text{ satisfies} \\ \nu_r + \frac{(1-\rho^{\nu_r-1})(1-\rho^{\nu_r+1})}{\rho^{\nu_r-1}(1-\rho)^2} = \nu_s. \end{gathered}$$

The average revenue per unit of time will be in this case

$$(3.2)\qquad Z_B^R = \lambda R \frac{1-\rho^{n_r}}{1-\rho^{n_r+1}}\left(1-\frac{n_r}{\nu_s}\right),$$

while social welfare, defined as the sum of rewards obtained from service minus

[2] A model which takes into account the possibility of customers' searching for the right time to arrive is analyzed by Glazer and Hassin (1983).

the waiting costs, is

$$
\begin{aligned}
Z_B^P &= \lambda R \frac{1-\rho^{n_r}}{1-\rho^{n_r+1}} - c\left[\frac{\rho}{1-\rho} - \frac{(n_r+1)\rho^{n_r+1}}{1-\rho^{n_r+1}}\right] \\
&= \lambda R\left\{\frac{1-\rho^{n_r}}{1-\rho^{n_r+1}} - \frac{1}{\nu_s}\left[\frac{1}{1-\rho} - \frac{(n_r+1)\rho^{n_r}}{1-\rho^{n_r+1}}\right]\right\}.
\end{aligned} \tag{3.3}
$$

It is immediate from equation (3.1) that $\nu_s \geqslant \nu_r$, and Naor has shown that the socially optimal maximum queue size, n_o, is equal to $[\nu_o]$ where $\nu_s \geqslant \nu_o \geqslant \nu_r$. As a function of λ, ν_s is constant, ν_o is decreasing from ν_s at $\lambda = 0$ to 1 as $\lambda \to \infty$, and ν_r increases from 1 as λ increases until it reaches its maximum value at $\lambda = \mu$, and then it decreases toward 1 as $\lambda \to \infty$. For $\lambda < \mu$, when either n_o or n_r changes, the difference between these values decreases. However for $\lambda > \mu$, this difference may be increased when n_o changes and decreased when n_r changes, since both are decreasing step-functions.

The functions Z_B^R and Z_B^P of equations (3.2) and (3.3) are illustrated in Figure 1. Z_B^P is discontinuous at the values λ where n_r changes. The jumps are "up" for $\lambda < \mu$ and "down" for $\lambda > \mu$. The functions coincide whenever $n_r = 1$, since in this case the firm's revenue coincides with the social welfare, while customers have zero surplus. As λ increases the functions approach $\mu R - c$, the net rate of benefit to the customer who is currently served.

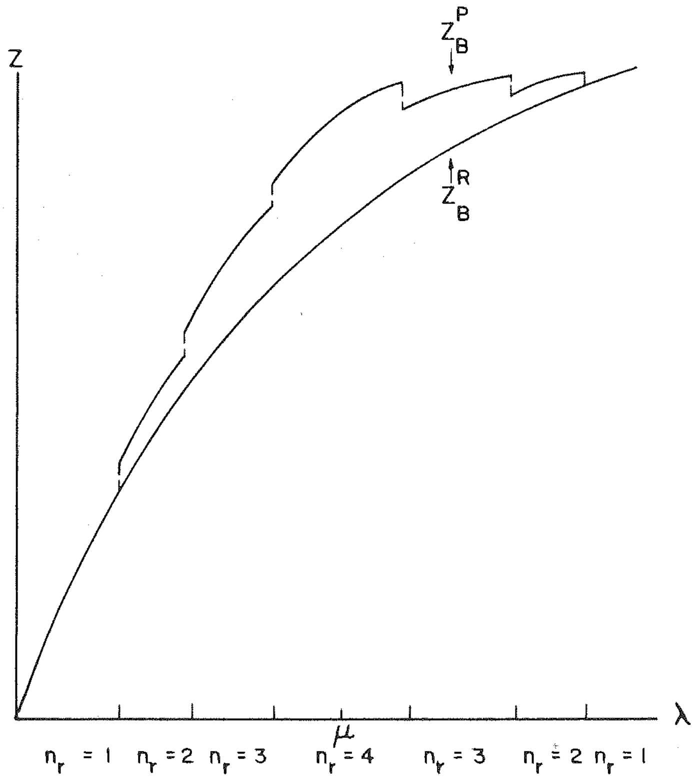

FIGURE 1.

4. QUEUES WITHOUT BALKING

Consider a facility with admission toll θ. If the arrival rate is λ then it is a well known result in elementary queueing theory that the customer's expected waiting time (in queue and in service) is $1/(\mu-\lambda)$. Thus a customer's net benefit is $R-\theta-c/(\mu-\lambda)$. If this value is negative then the actual arrival rate will reach equilibrium at $\lambda(\theta)$ such that

$$\theta+\frac{c}{\mu-\lambda(\theta)}=R. \tag{4.1}$$

Thus, the actual arrival rate will be min $(\lambda, \lambda(\theta))$. The firm's objective is to choose a toll θ maximizing $\theta\cdot\min(\lambda, \lambda(\theta))$. It can be easily shown that this product is maximized by setting

$$\theta=R\left(1-\left(\frac{1}{\nu_s}\right)^{1/2}\right) \tag{4.2}$$

in which case the actual arrival rate is λ_m, where

$$\lambda_m=\mu\left(1-\left(\frac{1}{\nu_s}\right)^{1/2}\right), \tag{4.3}$$

provided that $\lambda\geq\lambda_m$.

If, however, $\lambda<\lambda_m$, then the firm will set the maximum toll that customers are ready to pay when all of them choose to arrive, so that $\theta=R-c/(\mu-\lambda)$. The firm's revenue, Z^R_{NB}, satisfies

$$Z^R_{NB}=\begin{cases}\lambda R\left(1-\dfrac{1}{\nu_s(1-\rho)}\right), & \lambda\leq\lambda_m,\\[2ex] \mu R\left(1-\left(\dfrac{1}{\nu_s}\right)^{1/2}\right)^2, & \lambda\geq\lambda_m.\end{cases} \tag{4.4}$$

It is important to observe that, under the model's assumptions, no customer will balk after joining the queue! It can be shown that the posterior distribution of the number of customers ahead of a customer who has already waited t units of time is identical for all values of t.

Edelson and Hildebrand (1975) have shown that in a queueing system without balking, the enterpreneur's objective function is identical to its social counterpart. Thus the tolls imposed by the profit maximizing firm are Pareto optimal, and the social welfare, Z^P_{NB}, satisfies

$$Z^P_{NB}=Z^R_{NB}. \tag{4.5}$$

5. MAXIMIZATION OF REVENUE

We turn now to the revenue maximizer's problem. Which of the systems the firm will find more profitable depends on the potential arrival rate λ (assuming R, c, and μ to be fixed). Under both options the firms's revenue increases as λ

increases from zero. In the NB case, this increase stops at λ_m and Z^R_{NB} is constant for $\lambda > \lambda_m$. In the balking case the firm's revenue always increases with λ, and as λ becomes large the firm imposes the maximum toll which will not prevent customers from entering the system even when it is empty, i.e., $\theta = R - c/\mu$. As λ approaches ∞, the firm's revenue approaches $\mu R - c$ per unit time. It is clear that for large values of λ the firm prefers the B option, and will choose to inform its customers about the queue size upon their arrival. No trivial answer exists however for small values of λ.

As shown in Appendix I, Z^R_B and Z^R_{NB} intersect at most once. If $\nu_s \equiv R\mu/c \leq 2$, then $Z^R_{NB} < Z^R_B$ for all $\lambda > 0$ and the firm will always want to reveal the queue length. Note that this condition implies that even when no toll is imposed, a customer will decide to balk whenever he finds the server busy.

If $\nu_s > 2$, then there exists a unique value $\bar{\lambda}^R$ such that $Z^R_{NB} > Z^R_B$ for $\lambda < \bar{\lambda}$ and $Z^R_{NB} < Z^R_B$ for $\lambda > \bar{\lambda}^R$. Thus for $\lambda < \bar{\lambda}^R$ the firm will choose to conceal the queue length from its customers, and for $\lambda > \bar{\lambda}^R$ it will disclose this information.

6. SOCIAL WELFARE

We can use now the results described in the previous sections to solve the second best problem addressed in the introduction. It is shown in Appendix II that Z^P_B and Z^P_{NB} intersect at most once, and that the condition for such an intersection is just as with Z^R_B and Z^R_{NB}, depending on whether ν_s is greater or less than 2. Since $Z^P_B \geqslant Z^R_B$ for every λ, then Z^P_B and Z^P_{NB} intersect in a smaller value of λ than Z^R_B and Z^R_{NB}. This is illustrated in Figure 2.

For arrival rates $\bar{\lambda}^P < \lambda < \bar{\lambda}^R$, social welfare would increase if the firm could be induced to choose the balking option. This is clearly not the case if $\lambda < \bar{\lambda}^P$. Also, it is never worthwhile to induce the firm to choose the no balking option when it does not voluntarily choose it.

7. OPTIMAL NUMBER OF FACILITIES

When the demand rate, λ, is large, the firm may find it profitable to put to work several facilities rather than just one. To simplify the analysis we assume that λ is very large so that we can find the optimal arrival rate to each facility without taking into account possible problems of indivisibility. We also assume that the operation costs are q per unit of time per facility, independent of the rate of customers that are actually served.

A necessary condition for any profit is that $q \leqslant \mu R - C$. If also $q > \mu R[1-(\nu_s)^{-1/2}]^2$, then [cf., equation (4.4)] no profit can be gained if customers are not informed about the queue size. For lower values of q, both B and NB may be profitable.

If the gain from a single facility is $Z(\lambda)$, then the optimal arrival rate per facility is that which maximizes $[Z(\lambda)-q]/\lambda$; that is, maximum profit per unit of arrival rate. A graphic solution can be obtained by looking at the line which

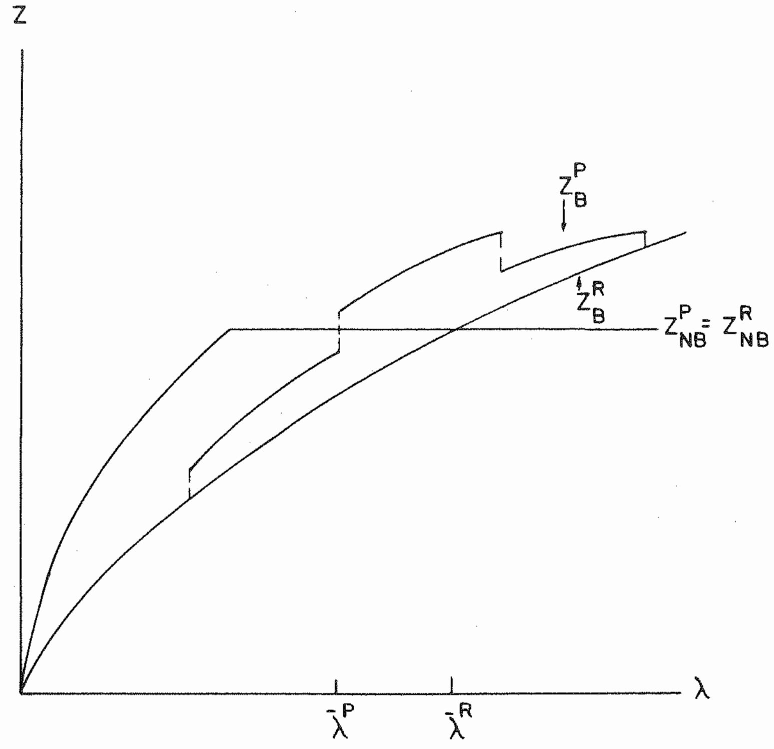

FIGURE 2.

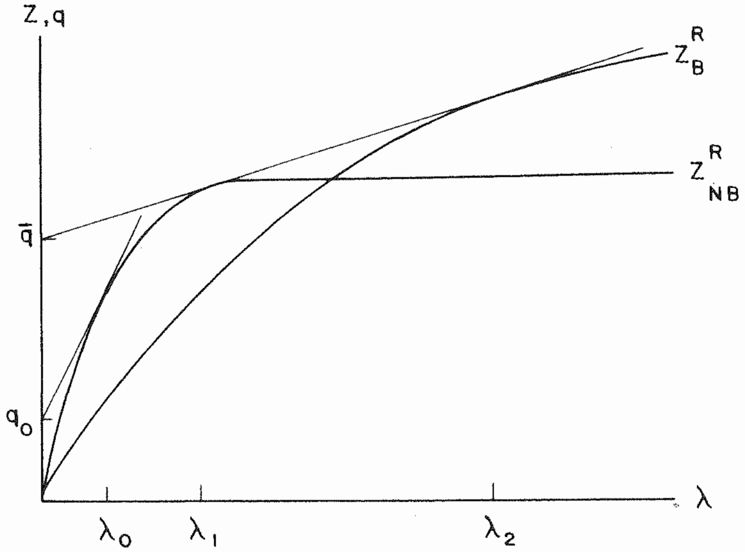

FIGURE 3.

passes through $(0, q)$ and is tangent to $Z(\lambda)$. Thus, in Figure 3, for $q = q_0$ the optimal rate per facility in the NB case is λ_0.

In Figure 3, for $q = \bar{q}$ both λ_1 with NB and λ_2 with B are revenue maximizing. For $q < \bar{q}$ the optimal λ is smaller than λ_1 and for $q > \bar{q}$ it is greater than λ_2. Thus the optimal congestion at the facility is not a continuous function of the operating costs. By imposing taxes or subsidies on the operating costs the solution of the revenue maximizer may be changed. However, no such tax or subsidy alone will induce the firm to choose λ in the interval (λ_1, λ_2).

Figure 4 depicts the Z functions for $\nu_s = 5.05$. Here $n_r = 2$ for $0.8\mu < \lambda < 1.25\mu$ and $n_r = 1$ elsewhere. It can be seen that if $q = q_1$, then social welfare can be increased by forcing the firm to choose the balking option, and with this the firm will maximize its revenue by operating a smaller number of more congested facilities. On the other hand, if $q = q_2$ then social welfare can be increased if the firm is induced (e.g., by proper subsidy on its operating costs) to operate a larger number of less congested facilities, while sticking to the balking option.

It is easy to see that it never pays, from a social point of view, to induce the firm to choose the NB option when it prefers balking. Let q be given, such that the firm prefers balking. Let λ_1 be the rate it chooses if it can adopt balking, and let λ_2 be the rate it chooses if it must adopt no balking. Then

$$\frac{Z_B^P - q}{\lambda_1} \geqslant \frac{Z_B^R - q}{\lambda_1} > \frac{Z_{NB}^R - q}{\lambda_2} = \frac{Z_{NB}^P - q}{\lambda_2}.$$

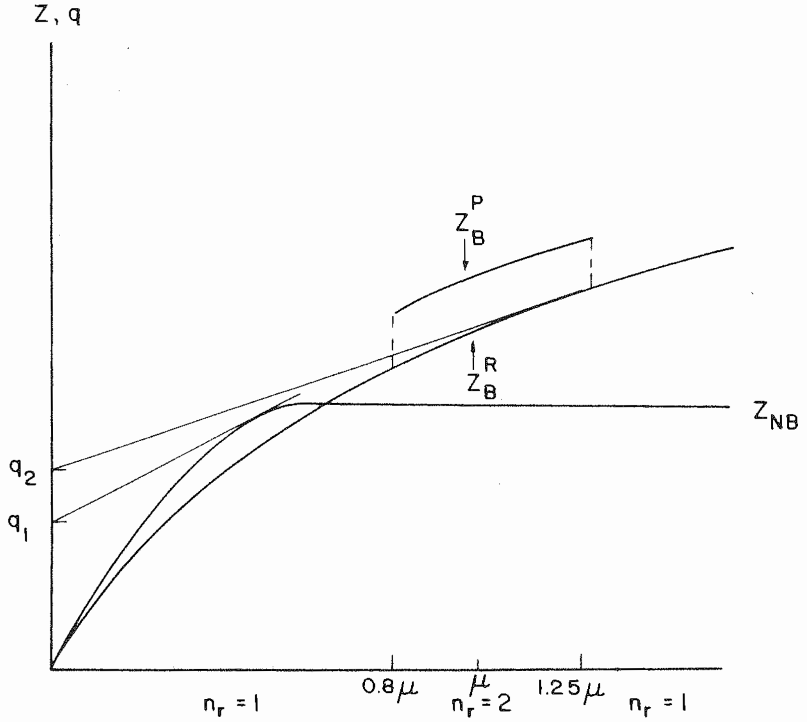

FIGURE 4.

The first inequality holds since $Z_B^P \geqslant Z_B^R$, the second since the firm prefers balking. Thus social welfare will decrease if the firm will choose λ_2 and no balking. We conclude that, as in the case of a single facility, if it is socially desirable to motivate the firm to change the option it chooses, this change is from concealing the queue length to revealing it.

8. CONCLUDING REMARKS

Some topics related to ours have already been discussed in the literature. Allocation of resources to information dissemination through advertising has been analyzed by Kotowitz and Mathewson (1979a). The same authors (1979b) have also shown that the monopoly's optimal policy in the short run may contain misleading information. Another related topic considers the firm's policy where it can control the quality of the product on which its reputation depends. The assumption here is that consumers are willing to pay more for the products of firms which have supplied high quality in the past (see Klein and Leffler (1981), Rogerson (1982), Shapiro (1982)). A more specific example of a problem of this type can be found in Glazer and Hassin (1982) where it is shown that by concealing the exact content of a good sold on a subscription, the firm's revenue and the social welfare may increase, simultaneously.

We did not examine some interesting extensions of the queueing model, and they may serve for future research. Instead of installing several single-server facilities, the firm may be able to operate one multi-server facility. The analysis will become much more complicated in this case, where Naor's simple results that we used must be replaced by those of Knudsen (1972). In another extension, the firm may choose to reveal partial information. For example it may announce when the queue length is greater than some value. By this the actual arrival rate to the facility will increase (relative to the case with suppressed information), while customers will balk only when the queue is very long. Yet another possible extension is a short run model, similar to the one discussed by Kotowitz and Mathewson (1979b), where the firm may choose to submit misleading information. Finally, it would be interesting to extend the analysis to systems such as public transportation and recreation facilities, where service quality depends on the system's congestion.

Statistics Dept., Tel Aviv University, Tel Aviv 69978, Israel

Manuscript received April, 1984; final revision received September, 1985.

APPENDIX I

We show here that if $\nu_s > 2$ then Z_B^R and Z_{NB}^R intersect exactly once. Otherwise, if $\nu_s \geqslant 2$, $Z_{NB}^R < Z_B^R$ for all values of λ. We first show that the functions intersect at most once, by proving that if they intersect at $\bar{\lambda}$ then $Z_B^R > Z_{NB}^R$ for all $\lambda > \bar{\lambda}$. This is clearly the case if $\lambda \geqslant \lambda_m$ (where λ_m is defined by equation (4.3)) since Z_{NB}^R is constant there while Z_B^R increases. Suppose $\lambda < \lambda_m$; then since $\lambda_m \leqslant \mu$, it follows that $\rho < 1$. By equations (3.2) and (4.4) $Z_{NB}^R \lesseqgtr Z_B^R$ is equivalent to

$$1 - \frac{1}{\nu_s(1-\rho)} \lesseqgtr \frac{1-\rho^n}{1-\rho^{n+1}}\left(1 - \frac{n}{\nu_s}\right),$$

where $n \equiv n_r$. This can be shown to be equivalent to

$$\nu_s \leq \frac{(1-\rho^{n+1})-(1-\rho)(1-\rho^n)n}{\rho^n(1-\rho)^2} \equiv g.$$

Note that although n_r is a step function of ρ, Z_B^R and Z_{NB}^R and thus also g are continuous functions of ρ. It can be shown that g is an increasing function of ρ and therefore equality can hold for at most a single value of ρ. For small values of λ, $n_r = 1$ and $g = 2/(1-\rho) > 2$. Since g is an increasing function of ρ (and thus, of λ) then $g > 2$ for all $\lambda < \mu$. Therefore, if $\nu_s \leq 2$, then $Z_{NB}^R < Z_B^R$ for all $\lambda < \mu$ and thus for all λ. If however $\nu_s > 2$, then for sufficiently small values of λ, $Z_{NB}^R > Z_B^R$ and there is a unique intersection of these functions.

APPENDIX II

We show here that Z_B^R and Z_{NB}^P intersect exactly once if $\nu_s > 2$, and that $Z_{NB}^P < Z_B^P$ for all λ if $\nu_s \leq 2$. We note that Z_B^P is not a continuous function of λ, and by "intersection" we mean a value $\bar{\lambda}$ such that $Z_{NB}^P > Z_B^P$ for all $\lambda < \bar{\lambda}$ and $Z_{NB}^P < Z_B^P$ for all $\lambda > \bar{\lambda}$, where Z_B^P may be undefined at $\bar{\lambda}$. We use the same method of proof as in Appendix I. For $\rho > 1$, Z_B^P is not monotone, but satisfies $Z_B^P \geq Z_B^R > Z_{NB}^R = Z_{NB}^P$, so that the proof for $\lambda > \lambda_m$ is again trivial. Consider $\lambda < \lambda_m$, so that $\rho < 1$. Using equations (3.3), (4.4), and (4.5), the condition $Z_{NB}^P \lesseqgtr Z_B^P$ is equivalent to

$$1 \lesseqgtr \frac{1-\rho^n}{1-\rho^{n+1}} + \frac{1}{\nu}\frac{(n+1)\rho^n}{1-\rho^{n+1}},$$

where $n \equiv n_r$. After some simplifications, this condition reduces to

$$\nu_s \lesseqgtr \frac{n+1}{1-\rho}.$$

For $\rho < 1$, n_r increases with ρ and therefore the right-hand side of this condition is monotone increasing. Hence, there can be at most one value of λ where $Z_{NB}^P - Z_B^P$ changes sign.

Since $n_1 \geq 1$, the right-hand side of the condition is greater than 2 for $\rho < 1$. Thus if $\nu_s \leq 2$, $Z_{NB}^P < Z_B^P$ for all values of λ. If $\nu_s > 2$, then for a sufficiently small λ, $n_r = 1$ and $Z_{NB}^P > Z_B^P$, so that a unique intersection of these functions exists.

REFERENCES

BALACHANDRAN, K. R., AND M. E. SCHAEFER (1980): "Public and Private Optimization at a Service Facility with Approximate Information on Congestion," *European Journal of Operational Research*, 4, 195-202.

EDELSON, N. M., AND K. HILDEBRAND (1975): "Congestion Tolls for Poisson Queueing Processes," *Econometrica*, 43, 81-92.

GLAZER, A., AND R. HASSIN (1982): "On the Economics of Subscriptions," *European Economic Review*, 19, 343-356.

——— (1983): "Search Among Queues," IMSSS Technical Report No. 406, Stanford University.

HASSIN, R. (1985): "On the Optimality of First-Come Last-Served Queues," *Econometrica*, 53, 201-202.

KNUDSEN, N. C. (1972): "Individual and Social Optimization in a Multi-server Queue with a General Cost Benefit Structure," *Econometrica*, 40, 515-528.

KLEIN, B, AND K. B. LEFFLER (1981): "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, 89, 615-641.

KOTOWITZ, Y., AND F. MATHEWSON (1979a): "Informative Advertising and Welfare," *American Economic Review*, 69, 284-294.

——— (1979b): "Advertising, Consumer Information, and Product Quality," *Bell Journal of Economics*, 10, 566-588.

LITTLECHILD, S. C. (1979): "Optimal Arrival Rate in a Simple Queueing System," *International Journal of Production Research*, 12, 391-397.

NAOR, P. (1969): "The Regulation of Queue Size by Levying Tolls," *Econometrica*, 37, 15-23.