APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind. 2008; **24**:307–323
Published online 22 January 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asmb.708

Coordination of staffing and pricing decisions in a service firm

Doğan A. Serel*, † and Erdal Erel

Faculty of Business Administration, Bilkent University, 06800 Bilkent, Ankara, Turkey

SUMMARY

Customer demand is sensitive to the price paid for the service in many service environments. Using queueing theory framework, we develop profit maximization models for jointly determining the price and the staffing level in a service company. The models include constraints on the average waiting time and the blocking probability. We show convexity of the single-variable subproblem under certain plausible assumptions on the demand and staffing cost functions. Using numerical examples, we investigate the sensitivity of the price and the staffing level to changes in the marginal service cost and the user-specified constraint on the congestion measure. Copyright © 2008 John Wiley & Sons, Ltd.

Received 20 February 2007; Accepted 28 November 2007

KEY WORDS: queue; optimal staffing level; waiting lines; nonlinear programming; Erlang loss; pricing

1. INTRODUCTION

Determining the best level of production capacity to install is not an easy decision since the customer demand faced by a company in many cases is unpredictable and varies over time. Making frequent changes in capacity in accordance with fluctuations in the demand level usually is not an acceptable strategy since this may entail high costs or firms may not possess the resource flexibility needed to closely match demand and supply over time. As a result of insufficient service capacity, arrival of a large number of customers over a short time interval creates congestion in the system. While these service delays undoubtedly cause customer dissatisfaction, nonetheless companies consider these delays necessary in order to keep operating costs in control. Thus, given randomly arriving customers at a service facility, the service providing firm has to build an appropriate level

*Correspondence to: Doğan A. Serel, Faculty of Business Administration, Bilkent University, 06800 Bilkent, Ankara, Turkey.
†E-mail: serel@bilkent.edu.tr

of service capacity in advance by considering the trade-off between the cost of capacity and the waiting time experienced by the customers. Various researchers have used queueing theory as a tool to analyze the performance of stochastic manufacturing and service systems. In a well-known model in the literature for determining the optimal capacity level (see, e.g. [1]), the objective is defined as the minimization of the sum of the service capacity cost and the cost of waiting (measured based on the customers' average waiting time).

In this paper, we explore the issue of coordinating the staffing and pricing decisions in a service facility using mathematical models based on queueing theory. Since in many cases arrival rate to the system depends on the price of the service charged to customers (e.g. [2]), we focus on the joint optimization of price and service capacity. Thus, rather than cost minimization, we adopt the objective of profit maximization in our study. To determine the optimal price and capacity, the joint impact of these decisions on the profit margin, the congestion level, and the cost of servers should be taken into account.

To broaden the scope of application, we investigate three different types of multiserver queueing systems, categorized according to the maximum length of the waiting line allowed: (1) infinite queue capacity (M/M/s), (2) loss model with no waiting in line (M/G/s/s), and (3) finite queue capacity (M/M/s/K). In the M/M/s system, all customers will be eventually served regardless of their arrival times. In the no queue and finite queue systems, the occurrence of a high number of arrivals within a short time interval may cause some customers to be blocked when they arrive, resulting in lost business for the firm.

The particular queueing systems (M/M/s, M/G/s/s, and M/M/s/K) underlying the optimization models are fairly suitable for representing a wide range of real-world settings. Single-stage, multiple server queueing models have been applied in areas including tele-marketing, emergency calls for police and ambulances, fast food restaurants, consumer banking, supermarkets, and call centers (e.g. [3, 4]).

To incorporate the service providing firm's concern with customer dissatisfaction into the model, we include constraints defining maximum allowable limits on the average waiting time or the probability of blockage by the system. In many service environments, delays beyond a certain threshold evoke negative reactions by customers. For the over the counter service in a bank, it is possible to talk about a typical patience threshold of 3 min for the customers. Customers waiting more than 3 min show various signs of impatience such as checking their watches, angrily watching the tellers, and discussing the wait with others in the queue [5]. The delay threshold for airline departure times is around 30 min [6]. Hui and Tse [7] observed that the users did not show signs of disapproval when the wait was 5 min or less in a computerized course registration service at a Canadian university. Some studies have found that up to 27% of customers who cannot get through on the telephone will either purchase elsewhere or not recall again [8].

The remainder of the paper is organized as follows. After reviewing the related literature, we discuss key modeling assumptions in Section 3. The mathematical formulations of the M/M/s system and its variants are presented in Section 4. Under mild conditions on the demand and server cost functions, for each type of queueing system, we show the convexity of the problem in the single-variable maximization case and analyze properties of the optimal solution. We show that when the elasticity of arrivals is increasing in price and the number of servers is kept fixed, the optimal price is higher than the price that maximizes the revenues. In Section 5, we provide numerical examples to illustrate our methodology. Concluding remarks are given in Section 6.

DOI: 10.1002/asmb

2. LITERATURE REVIEW

The problem of optimally determining the service capacity in a queueing system has been studied extensively in the literature, e.g. [9, 10]. Since our work primarily concerns the use of price to influence the arrival rate of customers, we concentrate on this part of the literature.

Various researchers have studied the problem of choosing the optimal price in a service facility. Assuming users with delay costs and a fixed level of service capacity, it has been shown that the socially optimal level of congestion (or equivalently, the optimal level of arrivals) in a queueing system can be attained by imposing fees on the users [11]. Mendelson [12] proposes an economic model where the service price is determined optimally to minimize the sum of service capacity and user waiting costs, assuming that the user delay cost accrues at a constant rate over waiting time. Dewan and Mendelson [13] extend that model to the case where a general delay cost function is allowed. Using a similar model, Stidham [14] investigates the properties of the optimal solution when there is an upper bound on the arrival rate. Ha [15] studies the problem of finding the optimal class-specific pricing schemes that can coordinate a multiclass system where service requirements are chosen by the customers. Several other papers in this research stream incorporate the effect of guarantees on the maximum waiting time. Palaka *et al.* [16] treat demand as being linear in price and quoted lead time and employ an M/M/1 queueing model to study the firm's price and quoted lead time choices. So and Song [17] study a similar model with a demand function log-linear in price and quoted lead time. Ray and Jewkes [18] consider delivery time-dependent price and also allow economies of scale by assuming that the unit operating cost is a decreasing function of the mean demand rate. Larsen [19] develops a model where the value placed upon service by a potential customer is a random variable, and the customer enters the system if this value exceeds the sum of the price charged for his job plus the expected waiting costs.

Our research is closer to several papers that assume that the mean arrival rate is inversely related to price and/or the average waiting time of customers. Ittig [20] investigates the problem of finding the optimal number of servers in an M/M/s system when the arrival rate is negatively related to the average waiting time. Jahnke *et al.* [21] study an M/M/1 system with a kinked demand curve. Up to a threshold capacity utilization level, demand is a decreasing linear function of price only; if the capacity utilization is greater than the threshold level, decreasing service level (caused by higher capacity utilization) also negatively affects demand.

While in general a single-channel (server) delay system is assumed in the papers cited above, there are also studies on Erlang loss systems with multiple channels. In a loss system, no queue is allowed, and customers finding all servers busy at their arrival are not served and rejected from the system. Carrizosa *et al.* [22] study the optimal admission policy when the arrivals at a loss system can be classified into different groups. Caro and Simchi-Levi [2] explore the pricing problem faced by a network service provider that has a fixed capacity and different classes of customers. Ziya *et al.* [23] look into a similar problem in which there is only one class of customers, and the waiting line has a finite capacity. Each arriving customer has her own reservation price and enters the system only when this reservation price is higher than or equal to the price charged by the firm.

We note that there are numerous other economic models for the multiserver systems proposed in the literature assuming a price-independent arrival rate. For example, Borst *et al.* [24] minimize the sum of staffing and waiting costs in an M/M/s system in which the waiting cost is an increasing function of the waiting time experienced by a customer. Kochel [25] explores the problem of

DOI: 10.1002/asmb

optimally choosing the number of servers and number of waiting places in an M/M/s/K finite queue system. For a recent literature review, see Tadj and Choudhury [9]. In general, in the previous literature either the service capacity or price is assumed given, and the optimization is carried out on only one of these two variables. The models that consider the joint selection of capacity and price have generally used the single-server M/G/1 framework. Hence, previous studies have not fully explored the issue of jointly determining the optimal price, service capacity, and queue capacity in a single-stage, multiple server queueing system subject to price-dependent customer demand.

3. MODELING FRAMEWORK

We address the problem of optimally choosing the price p, and the number of servers s, to maximize the expected profit per hour in a single-stage system consisting of identical server stations working in parallel with a mean service rate of μ per server per hour. Throughout the paper, we assume μ is fixed and given. In a single-stage system, the customer receives service from only one station and then leaves the system. Customers arrive at the system following a Poisson process with a price-dependant mean arrival rate of λ per hour; they join a single waiting line and are served by the first available server. In Sections 4.1–4.3, we separately consider three cases: (1) an Erlang delay system (M/M/s) in which the waiting line capacity is infinite, and service times are exponentially distributed; (2) an Erlang loss system (M/G/s/s) where no distributional assumption is made for the service times, and there are no waiting places; and (3) an M/M/s/K system with s servers, finite waiting line capacity $m = K - s$, and exponentially distributed service times. Thus, because some customers are blocked from entering the system when they arrive, the average number of customers served per hour will be less than λ in the second and third cases.

The average arrival rate λ is inversely related to the price p. To simplify the analysis, we will work with the inverse demand function $p(\lambda)$ and assume that $p(\lambda)$ is concave and nonincreasing in λ. The hourly cost of staffing $g(s)$ is assumed to be an increasing convex function of the number of servers s. This assumption implies that the marginal cost of using an additional server does not decrease as the number of servers increases. For instance, in tight labor markets, the hourly wages increase with the demand for labor, resulting in a convex staffing cost function [24]. In a tight labor market, the workers are hired at relatively low wage initially; as the available supply of potential workers decreases, higher wages should be offered in order to attract additional workers who will not accept lower wages. We remark that a linear relationship between λ and p, and the linear capacity cost $c_s s$ (with c_s as the server cost per hour) belong to the set of functions satisfying our assumptions regarding $p(\lambda)$ and $g(s)$. We also assume that the marginal cost of each service to the firm is c, such that each served customer contributes $(p - c)$ to the profit. Thus, the total hourly cost of the firm is a function of both the cost of service c and the staffing cost $g(s)$.

As noted previously, there are situations where the quality of a service perceived by a customer rapidly deteriorates if the waiting time experienced by the customer exceeds a certain threshold level. To incorporate this factor into our model, we include a constraint setting an upper limit for the average waiting time in our optimization model of the M/M/s system. A similar constraint on the Erlang loss probability is imposed on the M/G/s/s loss system and the M/M/s/K finite queue system, based on the idea that limiting the probability of rejection at arrival helps increase customer satisfaction. We note that models with this kind of constraints on congestion measures

DOI: 10.1002/asmb

have also been investigated by other researchers, e.g. Berman and Larson [26] and Jahnke *et al.* [21]. In the facility location literature, the constraint on the congestion measure has been specified as an upper bound on the probability of encountering more than a certain number of users in queue at arrival [27–29]. In Section 4.4, we study the extension of the M/M/s infinite queue model to the case where the total hourly cost also explicitly includes the estimated cost of customer dissatisfaction due to waiting.

4. DETERMINATION OF THE OPTIMAL PRICE AND STAFFING LEVEL

In this section we present optimization models for queueing systems under three different scenarios regarding the maximum allowable queue length. We first consider an M/M/s system in which customers finding all servers busy at their arrival wait in line until a server becomes available.

4.1. Profit maximization in the M/M/s model

Let $a=\lambda/\mu$ (the offered load) and $\rho=\lambda/s\mu$ (the traffic intensity). To maximize the expected profit per hour, $\Pi_D(s,\lambda)$, the service provider should solve the following nonlinear programming model.

$$\begin{aligned} \text{(P1) Max} \quad & \Pi_D(s,\lambda)=[p(\lambda)-c]\lambda-g(s) \\ \text{s.t.} \quad & w(s,a)\leqslant w_{\max} \\ & \lambda<s\mu \end{aligned}$$

where $w(s,a)$ is the average waiting time in the M/M/s system and $w_{\max}$ is the prespecified upper bound on the average waiting time. The stability condition $\lambda<s\mu$ ensures that the overall system capacity is greater than demand. The average waiting time in the system (including the service time), $w(s,a)$, is given by

$$w(s,a)=C(s,a)/[\mu(s-a)]+(1/\mu) \tag{1}$$

where $C(s,a)$ is the Erlang-C probability of delay defined as

$$C(s,a)=\frac{\dfrac{a^s}{s!(1-\rho)}}{\sum_{i=0}^{s-1}a^i/i!+a^s/s!(1-\rho)} \tag{2}$$

$C(s,a)$ gives the probability that an arriving customer has to wait in line.

We show in Lemma 1 that the objective function in (P1) is jointly concave in s and λ.

Lemma 1
$\Pi_D(s,\lambda)$ in (P1) is jointly concave in s and λ.

Proof
The elements of the Hessian matrix of $\Pi_D(s,\lambda)$, $\nabla^2\Pi_D=H(s,\lambda)=[h_{ij}]$ are

$$h_{11}=p''(\lambda)\lambda+2p'(\lambda), \quad h_{12}=h_{21}=0 \quad \text{and} \quad h_{22}=-g''(s)$$

DOI: 10.1002/asmb

By assumption $p''(\lambda_p) \leqslant 0$, $p'(\lambda_p) \leqslant 0$, and $g''(s) \geqslant 0$, implying that h_{11} and h_{22} are nonpositive. Thus, $H(s, \lambda)$ is negative semidefinite and $\Pi_D(s, \lambda)$ is jointly concave in s and λ. □

The constraint function $\lambda - s\mu$ is jointly convex in s and λ. If the other constraint function $w(s, a)$ is also jointly convex in s and λ, (P1) would be a convex program, guaranteeing a unique local maximum. Unfortunately, we are not able to show joint convexity of $w(s, a)$. Nonetheless, we can prove that (P1) has a unique local maximum if one of the variables is kept fixed. Using Lemma 1, we prove in Proposition 1 that (P1) is a convex program if one of the variables is fixed.

Proposition 1
When (P1) is solved for a fixed value of λ or s, the local maximum point will also be a global maximum.

Proof
First we consider a fixed s. From Lemma 1, $\Pi_D(s, \lambda)$ is concave in λ. The average waiting time $w(s, \lambda)$ is convex in λ in an M/M/s system [30]. The constraint $\lambda < s\mu$ also defines a convex region because it is linear in λ. Hence, the feasible set of constraints is convex when we treat s as fixed. Thus, the local maximum of (P1) will be the global maximum when s is fixed. Now we fix λ. The negative of a convex function is concave; hence, $-g(s)$, and as a result of that, the objective function $\Pi_D(s, \lambda)$ is concave in s. The convexity of $w(s, \lambda)$ in s has been shown in Dyer and Proll [31]. The constraint $\lambda < s\mu$ is linear in s. Hence, there exists only one local maximum of (P1) when λ is fixed. □

To solve (P1), we use a sequential search method in which we find the optimal arrival rate and profit keeping the number of servers fixed. We increase the number of servers by one at each iteration and continue to iterate until the expected profit starts to decrease. This method converges to the optimal solution if $w(s, a)$ is jointly convex in s and λ, but it may be trapped in a local optimum if there are multiple local optima in the search space. Note that, using an off-the-shelf nonlinear optimization software, we can also attempt to find the optimal solution by searching over s and λ simultaneously. As discussed in Section 5, in our numerical study, we have observed that the solutions obtained from our method are consistent with the global optimal solutions found via grid search, suggesting that our approach can be successfully used at least in a certain set of problems.

In Proposition 2, we show that, for a given s, the price maximizing the (partial) profit $(p - c)\lambda(p)$ is larger than the price maximizing the revenue $p\lambda(p)$.

Proposition 2
For a fixed number of servers s, the optimal price in (P1) increases in the marginal service cost c.

Proof
To show that $p^*(c)$ is increasing in c, it is sufficient to show that $\Pi_D(s, \lambda)$ is submodular in (λ, c) [32], which is true if $f_D(\lambda, c) \equiv -\lambda c$ is submodular in (λ, c). The lower bound for price is the marginal service cost c, and this lower bound increases as c increases. Since the mixed partial derivative $\partial^2 f_D / \partial\lambda \partial c = -1 < 0$, $f_D(\lambda, c)$ is submodular and $\lambda^*(c)$ is decreasing in c. Hence, $p^*(c)$ is increasing in c. □

DOI: 10.1002/asmb

4.2. *Profit maximization in the M/G/s/s loss model*

In the loss system, the expected number of customers served per hour will be $\lambda[1-B(s,a)]$, where $B(s,a)$ is the Erlang loss probability, i.e. the fraction of customers finding all servers busy at their arrival. $B(s,a)$ can be computed from (see, e.g. [33])

$$B(s,a)=\frac{(a^s/s!)}{\sum_{i=0}^{s}(a^i/i!)} \tag{3}$$

The optimization problem in the loss system can be expressed as

$$\text{(P2)}\ \text{Max} \quad \Pi_L(s,\lambda)=[p(\lambda)-c]\lambda[1-B(s,a)]-g(s)$$

$$\text{s.t.} \quad B(s,a)\leqslant b_{\max}$$

where $b_{\max}$ is the maximum loss probability allowed and $\Pi_L(s,\lambda)$ is the service provider's expected profit per hour. Although we are not able to show that (P2) is a convex program, in Proposition 3, we show that problem (P2) has properties similar to (P1). To that end, we first present Lemma 2.

Lemma 2
If f is a nonnegative, increasing, and concave function of a single variable x, and g is a nonnegative, nonincreasing, and concave function of x, then the multiplication of f and g is concave in x.

Proposition 3
When (P2) is solved for a fixed value of λ or s, the local maximum point will also be a global maximum.

Proof
We first treat s as fixed. The term $\lambda[1-B(s,a)]$, expected number of customers served per hour, has been shown to be concave in λ [33, 34]. It is also increasing in λ [35]. The function $[p(\lambda)-c]$ is nonnegative, nonincreasing, and concave in λ. Hence, from Lemma 2, $\Pi_L(s,\lambda)$ is concave in λ. Although $B(s,a)$ is not convex in λ in general [33], we can show that the feasible region is convex. We can rewrite the constraint on the loss probability as $1/[1-B(s,a)]\leqslant 1/(1-b_{\max})$. Since $1/[1-B(s,a)]$ is convex in λ [22], (P2) has a single maximum when s is fixed. When λ is fixed, it is easy to show that $\Pi_L(s,\lambda)$ is concave in s due to the concavity of $\lambda[1-B(s,a)]$ and $-g(s)$. It is known that $B(s,a)$ is convex in s [36]. Thus, the local maximum of (P2) is also the global maximum when λ is fixed. □

To determine the optimal solution to problem (P2), we use the sequential search approach described for problem (P1). If the arrival rate function $\lambda(p)$ exhibits increasing price elasticity, we can describe the optimal arrival rate for a given s in more detail. The coefficient of demand elasticity, $e(p)$, is defined as

$$e(p)=-(p/\lambda)(\mathrm{d}\lambda/\mathrm{d}p) \tag{4}$$

The price elasticity of demand measures the responsiveness of demand to a change in price. If the percentage of change in demand is larger than the percentage of change in price, $e(p)>1$, and the demand is described as elastic. If the demand is elastic, a price increase results in a lower revenue. Conversely, if the demand is inelastic ($e(p)<1$), a price increase leads to a higher revenue. If

DOI: 10.1002/asmb

$e(p)$ is increasing in p, a 1% change in price results in a higher percentage change in demand at higher prices, and $p\lambda(p)$ is maximized for the unitary elasticity $e=1$. From Proposition 2, if $e(p)$ is increasing in p, $(p-c)\lambda(p)$ is maximized in the region $e(p)>1$.

Let λ_p be the arrival rate maximizing $[p(\lambda)-c]\lambda$, and $e_l>1$ be the price elasticity of demand when $\lambda=\lambda_p$. If the elasticity of arrival rate e is increasing in price, then we can establish an upper bound on the optimal arrival rate when (P2) is solved for a given s. This bound is described in Proposition 4.

Proposition 4
Suppose that the elasticity of demand is increasing in p, that is, $e'(p)>0$. When (P2) is solved for a fixed value of s, the optimal arrival rate λ^* is less than or equal to λ_p.

Proof
Let $p_{\min}$ be the minimum feasible price and p_m be the price satisfying $e=e_l$. The loss probability $B(s,a)$ is increasing in λ. Hence, $1-B(s,a)$ increases as price increases. The term $(p-c)\lambda(p)$ is increasing in p in the region $e<e_l$. Since $e(p)$ is increasing in p, $(p-c)\lambda(p)[1-B(s,a)]$ is increasing in p for $e\leqslant e_l$. The constraint $B(s,a)\leqslant b_{\max}$ is satisfied more easily as p increases. Hence, the optimal price cannot be in the interval between $p_{\min}$ and p_m. Correspondingly, the optimal arrival rate cannot exceed λ_p. □

As an example for the case of increasing price elasticity, assume that the arrival rate depends linearly on price:

$$\lambda(p)=\alpha-\beta p \tag{5}$$

where α and β are positive parameters. Using (4), the coefficient of demand elasticity is

$$e(p)=p\beta/(\alpha-\beta p) \tag{6}$$

which is increasing in p.

It can also be shown that the optimal price increases in c when the number of servers s is kept fixed. Analogously to Proposition 2, define $f_L(\lambda,c)=-c\lambda[1-B(s,a)]$. Then $\partial^2 f_L/\partial\lambda\partial c=-\partial[\lambda-\lambda B(s,a)]/\partial\lambda<0$ since the throughput $[\lambda-\lambda B(s,a)]$ is increasing in λ [35]. Thus, $\lambda^*(c)$ is decreasing in c and $p^*(c)$ is increasing in c.

We also note that the optimal unconstrained price in problem (P2) for a given s can be obtained by using the first-order condition:

$$\frac{\partial \Pi_{\mathrm{L}}}{\partial p}=\left[\lambda(p)+(p-c)\frac{\partial\lambda}{\partial p}\right][1-B(s,a)]-(p-c)\lambda(p)\frac{\partial B(s,a)}{\partial p}=0 \tag{7}$$

The first partial derivative of the Erlang loss probability with respect to price is

$$\partial B(s,a)/\partial p=[\partial B(s,a)/\partial\lambda](\partial\lambda/\partial p) \tag{8}$$

where $\partial B(s,a)/\partial\lambda=B(s,a)[B(s,a)+(1/\rho)-1]/\mu$ [30]. Thus, substituting (8) into (7), the optimal unconstrained p can be determined by finding the zero of a nonlinear equation.

4.3. Profit maximization in the M/M/s/K finite queue model

In this subsection we consider the case where the maximum queue length is finite. If the waiting line capacity is finite, new arrivals will be blocked when there are already K customers in the

DOI: 10.1002/asmb

system (including those at server stations). The Erlang loss probability in the M/M/s/K system with line capacity $m = K - s$, $B(s, a, m)$ is given by

$$B(s,a,m) = \frac{(a^s/s!)(a/s)^m}{\sum_{i=0}^{s} a^i/i! + (a^s/s!)\sum_{i=1}^{m}(a/s)^i} \tag{9}$$

Thus, $B(s, a, m)$ is the probability that an arriving customer will not enter the system. In this scenario, we can incorporate the cost of waiting line capacity into our model. Let $h(m)$ be the hourly cost of maintaining m units of queue capacity. We assume $h(m)$ is an increasing convex function of m. To make the model more general, we allow the possibility that the queue capacity m is a decision variable. Since the throughput rate (the fraction of arrivals served) will be $\lambda[1 - B(s, a, m)]$, the relevant optimization problem now is

$$\begin{aligned} \text{(P3) Max} \quad & \Pi_{FQ}(s,\lambda,m) = [p(\lambda) - c]\lambda[1 - B(s,a,m)] - g(s) - h(m) \\ \text{s.t.} \quad & B(s,a,m) \leqslant b_{\max} \end{aligned}$$

where $b_{\max}$ is the maximum limit on the proportion of customers rejected and $\Pi_{FQ}(s, \lambda, m)$ is the service provider's expected profit per hour. Lemma 3 presented below indicates that the term $[1 - B(s, a, m)]^{-1}$ in the M/M/s/K system is convex in λ.

Lemma 3
The inverse of the nonblocking probability in the M/M/s/K finite queue system is a convex function of the arrival rate λ.

Proof
The nonblocking probability for a new arrival is

$$1 - B(s,a,m) = \frac{\sum_{i=0}^{s} a^i/i! + (a^s/s!)\sum_{i=1}^{m-1}(a/s)^i}{\sum_{i=0}^{s} a^i/i! + (a^s/s!)\sum_{i=1}^{m}(a/s)^i} \tag{10}$$

Then, we have

$$\frac{1}{1 - B(s,a,m)} = 1 + \frac{(a^s/s!)(a/s)^m}{\sum_{i=0}^{s} a^i/i! + (a^s/s!)\sum_{i=1}^{m-1}(a/s)^i} \tag{11}$$

or, equivalently,

$$[1 - B(s,a,m)]^{-1} = 1 + (a/s)B(s,a,m-1) \tag{12}$$

The loss rate $\lambda B(s, a, m-1)$ is convex in λ [37]. Hence, it follows that $[1 - B(s, a, m)]^{-1}$ is convex in λ [22, cf. Lemma 7.1]. □

In Proposition 5, we show that when we fix two of the three decision variables, problem (P3) has a unique optimal value for the remaining variable.

Proposition 5
When (P3) is solved by treating two of the three variables (s, λ, and m) as fixed, the maximum expected profit is unimodal in the remaining third variable.

DOI: 10.1002/asmb

Proof
We first treat s and m as fixed. The throughput $\lambda[1-B(s,a,m)]$ is increasing concave in λ [37]. Using Lemma 2, $\Pi_{\mathrm{FQ}}(s,\lambda,m)$ is concave in λ. As $B(s,a)$, $B(s,a,m)$ is not convex in λ for all values of a [37]. However, similar to Proposition 3, we rewrite the constraint on the loss probability as $1/[1-B(s,a,m)] \leqslant 1/(1-b_{\max})$. From Lemma 3, $1/[1-B(s,a,m)]$ is convex in λ, and thus, (P3) is unimodal in λ when s and m are fixed. Now fix λ and m. The term $\lambda[1-B(s,a,m)]$ is increasing concave in s [25, 38], and $g(s)$ is convex by assumption; hence, $\Pi_{\mathrm{FQ}}(s,\lambda,m)$ is concave in s. Since $B(s,a,m)$ is decreasing convex in s [25, 38], (P3) is unimodal in s when λ and m are fixed. Finally, we consider the behavior of (P3) with respect to line capacity m. The throughput $\lambda[1-B(s,a,m)]$ is increasing concave in m [37, 39]. Combining this result with convexity of $h(m)$, $\Pi_{\mathrm{FQ}}(s,\lambda,m)$ is concave in m. The Erlang loss probability $B(s,a,m)$ is convex in m [37, 39]. Thus, (P3) is unimodal in m. □

For solving problem (P3), we extend the sequential search method described in Section 4.1. We first find the optimal number of waiting places and arrival rate keeping the number of servers s fixed, and then conduct a search over s in order to arrive at the optimal solution.

For the case of increasing price elasticity, the arrival rate $\lambda=\lambda_p$ leading to $e=e_l$ defines an upper bound on the optimal arrival rate when (P3) is solved for fixed s and m. This can be shown in a similar manner to Proposition 4.

4.4. M/M/s model with customer waiting cost

We now consider a different economic model for the M/M/s system in which the objective function also explicitly includes the cost of waiting by customers. As noted earlier, economic models of this kind are well established in the literature. Let $L(s,a)$ be the average number of customers in the system, and c_w be the cost of waiting per customer per hour. It can be thought that this cost is caused by customer dissatisfaction and lost goodwill. Higher waiting times typically lead to a decrease in customer loyalty, and consequently lower future purchases. By Little's Law, $L(s,a)=\lambda w(s,a)$. Thus, the expected customer waiting cost is $L(s,a)c_w$ per hour. We can express the optimization problem as

$$\begin{aligned} \text{(P4) Max} \quad & \Pi_{\mathrm{D2}}(s,\lambda(p)) = (p-c)\lambda(p) - g(s) - L(s,\lambda(p))c_w \\ \text{s.t.} \quad & w(s,\lambda(p)) \leqslant w_{\max} \\ & \lambda(p) < s\mu \end{aligned}$$

Instead of the arrival rate, we consider price p as the decision variable in problem (P4). In addition to revenues from service, price now also has an impact on the customer waiting cost through the function $\lambda(p)$. An increase in price results in a decrease in waiting cost. Lemma 4, stated without proof, will be helpful in showing the uniqueness of the local optimal solution in the single-variable maximization case.

Lemma 4
Suppose $f(y)$ is a nondecreasing convex function of y and $y(x)$ is convex, then $f(y(x))$ is convex in x.

If $\lambda(p)$ is convex in p and $p\ \lambda(p)$ is concave in p, then (P4) is a convex program when one of the variables is treated as fixed; consequently, Proposition 6 holds. The second condition

DOI: 10.1002/asmb

is satisfied when $\lambda'(p) < -0.5p\lambda''(p)$. For example, a negative linear relationship between price and arrival rate satisfies these two conditions. Another example is that the double-log model

$$\ln(\lambda) = \alpha - \beta \ln(p)$$

meets these conditions when $\alpha > 0$ and $0 < \beta < 1$. The exponential demand function $\lambda(p) = \alpha \exp(-\beta p)$ satisfies the conditions if $\beta < 2/p$ or, equivalently, if the price elasticity of demand is less than 2. The convexity of $\lambda(p)$ implies that as price decreases, the arrival rate increases at an increasing rate. Various convex demand functions such as linear and log-linear have been employed in the literature and used in empirical studies [16–18, 20]. The linear, log-linear, and exponential demand functions have successfully fit the optical scanner data for a number of nondurable household and grocery store items sold in supermarkets [40]. The concavity of the revenue function $p\lambda(p)$ means that as the price increases the revenue will increase at a decreasing rate.

Proposition 6
Suppose $\lambda(p)$ is convex and $p\lambda(p)$ is concave. When (P4) is solved for a fixed value of p or s, the local maximum point will also be a global maximum.

Proof
The proof is similar to Proposition 1. The average number of customers in the system L is nondecreasing and convex in λ [41, 42]. Then, for fixed s, by Lemma 4, $L(s, \lambda(p))$ is convex in p. By convexity of $\lambda(p)$, the constraints define a convex region; hence, we have a unique maximum for a given s. For fixed p, L is convex in s [31]. Using same arguments as in Proposition 1, it follows that there is only one local maximum of (P4) when p is fixed. □

To solve problem (P4), the sequential search approach described in Section 4.1 can be applied in conjunction with the objective function $\Pi_{D2}(s, \lambda(p))$. A possible extension is to consider a convex waiting cost rather than a linear waiting cost c_w. If the waiting cost is convex, as the waiting time increases, the waiting cost will increase at an increasing rate. To respond to increasing waiting cost, the price and/or the number of servers should be increased.

Mandelbaum and Shimkin [43] argue that the waiting cost can be divided into two parts: a linear *alternative* waiting cost related to the actual value of time and a convex *psychological* waiting cost caused by the impatience that develops during waiting. The psychological cost is affected by the feeling of waste of invested time as well as the stress related to uncertainty associated with the remaining waiting time [44]. In a study of a fast food chain's customers, it has been found that after the actual waiting time in queue exceeds 5 min, customer perception of waiting time increases exponentially and differs from the actual time spent in line [45]. Nonlinear waiting cost functions may be observed in industries such as securities trading, food processing, banking and communication systems, and airline reservation systems [13, 46].

Let $D(t)$ be the waiting cost incurred by a customer who waits t units of time before the service begins. Assume $D(t)$ is increasing in t. Without loss of generality, assume $D(0) = 0$. The expected waiting cost per hour is

$$\lambda(p)E[D(\text{wait})] = \lambda(p)C(s, a)G(s, \lambda(p)) \tag{13}$$

DOI: 10.1002/asmb

where $G(s,\lambda(p))=E[D(\text{wait})|\text{wait}>0]=(s\mu-\lambda)\int_0^\infty D(t)\mathrm{e}^{-(s\mu-\lambda)t}\,\mathrm{d}t$. Hence, we can rewrite the objective function of the service provider as (cf. [24])

$$\text{Max } \Pi_{\mathrm{D3}}(s,\lambda(p))=(p-c)\lambda(p)-g(s)-\lambda(p)C(s,a)G(s,\lambda(p)) \tag{14}$$

Proposition 6 will also hold under the objective function $\Pi_{\mathrm{D3}}(s,\lambda(p))$ if we show that the expected total waiting cost per hour $\lambda(p)C(s,a)G(s,\lambda(p))$ is component-wise convex in s and p. First consider that p is fixed. Borst *et al.* [24] have shown that $\lambda(p)C(s,a)G(s,\lambda(p))$ is convex in s for a fixed p. This follows from the fact that $G(s,\lambda)$ is convex decreasing in s [24, Lemma C.1], the Erlang-C probability $C(s,a)$ is convex decreasing in s [47], and the multiplication of two nonnegative convex decreasing functions is convex. We now show convexity of the total waiting cost function in p. $C(s,a)$ is convex increasing in λ [42]. Then $\lambda C(s,a)$, a product of two nonnegative convex increasing functions, is convex increasing in λ. To show the convexity of $G(s,\lambda)$, let $\delta(\lambda)=s\mu-\lambda$, and consider the function:

$$\kappa(\delta)=\delta\int_0^\infty D(t)\mathrm{e}^{-\delta t}\,\mathrm{d}t$$

Borst *et al.* [24] have shown that $\kappa(\delta)$ is decreasing and convex in δ; hence, $\kappa(\delta)$ is increasing in λ. Since $\delta(\lambda)$ is linear in λ, $\kappa(\delta(\lambda))$ is convex in λ. Thus, we have shown that $\kappa(\delta(\lambda))$ or, equivalently, $G(s,\lambda)$ is convex and increasing in λ. Consequently, multiplication of $\lambda C(s,a)$ and $G(s,\lambda)$ is convex increasing in λ. Since $\lambda(p)$ is convex in p, by Lemma 4, the expected total waiting cost per hour, $\lambda(p)C(s,a)G(s,\lambda(p))$, is convex in p when s is kept fixed. In sum, Proposition 6 also holds when the objective function is given by (14).

5. NUMERICAL EXAMPLES

In this section we present some numerical examples to illustrate the models developed in earlier sections. We consider the linear demand function $\lambda(p)=100-6p, 0<p<\frac{100}{6}$. The service rate $\mu=5$ per hour and the marginal cost of serving a customer $c\in\{6,10\}$. We also assume linear hourly staffing cost function $g(s)=\$c_s s, c_s\in\{3,10\}$ and linear waiting line maintenance cost $h(m)=\$1m$ per hour. We also consider several different values for $b_{\max}$ and $w_{\max}$. We remark that in all numerical examples we have identified the optimal solution by both the sequential search method and the exhaustive grid search method. The results have turned out to be the same in both approaches.

We first assume that infinite queue space is available, i.e. the M/M/s system. The results are given in Table I. For example, given $w_{\max}=0.5$, $c_s=10$, and $c=10$, we obtain $\lambda^*=12.62$, $p^*=14.56$, and $s^*=3$. The optimal expected profit $\Pi_{\mathrm{D}}(s,\lambda)$ is \$27.58 per hour. Note that the arrival rate maximizing the partial profit $[p(\lambda)-c]\lambda$ is $\lambda_p=20$, and the corresponding unconstrained maximum profit (with $s=3$) is \$36.67 per hour. As the results in Table I indicate, the optimal arrival rate is not smoothly related to the upper limit on the average waiting time. For $c_s=3$ and $c=6$, λ^* decreases as $w_{\max}$ increases from 0.25 to 0.3, but λ^* increases when $w_{\max}$ increases from 0.3 to 0.5.

Next, we look into the Erlang loss system for which the results are reported in Table II. When the maximum loss probability $b_{\max}$ is 0.2, $c_s=10$, and $c=10$, the optimal solution is $\lambda^*=14.73$, $p^*=14.21$, and $s^*=4$. The corresponding optimal profit is $\Pi_{\mathrm{L}}(s,\lambda)=\$9.62$ per hour. Note that for

DOI: 10.1002/asmb

Table I. Optimal solution for the M/M/s model.

$w_{\max}$	c_s	c	λ^*	p^*	s^*	Π_D
0.25	3	6	31.44	11.43	8	146.61
		10	17.48	13.75	5	50.60
	10	6	26.74	12.21	7	96.05
		10	12.96	14.51	4	18.41
0.3	3	6	29.32	11.78	7	148.47
		10	19.69	13.38	5	51.65
	10	6	24.49	12.59	6	101.26
		10	14.95	14.18	4	22.41
0.5	3	6	32.00	11.33	7	149.67
		10	17.53	13.75	4	53.65
	10	6	27.42	12.10	6	107.17
		10	12.62	14.56	3	27.58
0.7	10	6	28.30	11.95	6	108.39
		10	13.39	14.44	3	29.39

Table II. Optimal solution for the M/G/s/s model.

$b_{\max}$	c_s	c	λ^*	p^*	s^*	Π_L
0.02	3	6	29.21	11.80	11	132.98
		10	18.14	13.64	8	40.77
	10	6	25.42	12.43	10	60.18
		10	11.38	14.77	6	−6.80
0.1	3	6	29.96	11.67	11	133.09
		10	17.24	13.79	6	42.22
	10	6	23.33	12.78	7	72.33
		10	10.23	14.96	4	5.67
0.2	3	6	29.96	11.67	11	133.09
		10	17.24	13.79	6	42.22
	10	6	25.03	12.49	7	72.92
		10	14.73	14.21	4	9.62
0.3	10	6	25.03	12.49	7	72.92
		10	13.17	14.47	3	11.22

$b_{\max}=0.02$, $c_s=10$, and $c=10$, the best solution satisfying the constraint on the loss probability results in a negative expected profit, implying that the optimal decision in this case is to drop the offering of the service. The effect of $b_{\max}$ on λ^* is not predictable since the loss probability $B(s,a)$ also depends on the number of servers. For $c_s=10$ and $c=6$, λ^* decreases when $b_{\max}$ changes from 0.02 to 0.1, but λ^* is higher when $b_{\max}=0.2$ than when $b_{\max}=0.1$.

The results for the finite queue problem are shown in Table III. For $b_{\max}=0.2$, $c_s=10$, and $c=10$, we obtain $\lambda^*=14.28$, $p^*=14.29$, $s^*=3$, and $m^*=5$. The optimal expected profit is $\Pi_{FQ}(s,\lambda,m)=$ \$19.72 per hour. The probability of blocking $B(s,a,m)=0.106<0.2$, indicating that the constraint is not binding at the optimal solution. Moving from an M/G/s/s to an M/M/s/K system, λ^* may increase or decrease. The optimal number of servers s^* appears to decrease when waiting is

DOI: 10.1002/asmb

Table III. Optimal solution for the M/M/s/K model.

$b_{\max}$	c_s	c	λ^*	p^*	s^*	m^*	Π_{FQ}
0.02	3	6	29.45	11.76	8	5	137.19
		10	17.51	13.75	5	5	44.32
	10	6	25.42	12.43	6	10	90.18
		10	12.11	14.65	3	9	16.15
0.1	3	6	29.58	11.74	8	5	137.20
		10	17.75	13.71	5	3	44.83
	10	6	26.58	12.24	6	8	91.12
		10	14.06	14.32	3	5	19.70
0.2	3	6	29.58	11.74	8	5	137.20
		10	17.75	13.71	5	3	44.83
	10	6	26.58	12.24	6	8	91.12
		10	14.28	14.29	3	5	19.72

Table IV. Optimal solution for the M/M/s model with waiting cost.

$w_{\max}$	c_s	c	λ^*	p^*	s^*	Π_{D2}
0.25	3	6	28.49	11.92	8	125.36
		10	16.32	13.95	5	37.89
	10	6	22.07	12.99	6	77.69
		10	12.96	14.51	4	8.69
0.3	3	6	28.49	11.92	8	125.36
		10	16.32	13.95	5	37.89
	10	6	23.97	12.67	6	79.39
		10	10.29	14.95	3	11.68
0.5	10	6	23.97	12.67	6	79.39
		10	11.02	14.83	3	12.09

allowed. As expected, under similar conditions, the optimal profit in the M/M/s/K system is not lower than that in the M/G/s/s system.

Finally, we consider the M/M/s model with linear customer waiting cost. Assuming $c_w = \$3$ per customer per hour, $w_{\max} = 0.5$ h, $c_s = 10$, and $c = 10$, after solving (P4) we find $\lambda^* = 11.02$, $p^* = 14.83$, and $s^* = 3$. The average number of customers in the system L is 3.71 at this optimal point. The maximum expected profit $\Pi_{D2}(s, \lambda(p)) = \$12.09$ per hour. The constraint on the average waiting time is nonbinding since the resulting w is 0.34 h. Other computational results for this model are listed in Table IV.

In Tables I–IV we observe some similar patterns. Analogously to our earlier analytical results, as the cost of service per customer c increases from 6 to 10, the optimal price p^* is observed to increase. The optimal number of servers s^* is nonincreasing with respect to the server cost c_s. Similarly, s^* does not increase when $w_{\max}$ or $b_{\max}$ increases; we observe no change in s^* as $b_{\max}$ changes in the numerical examples for the M/M/s/K finite queue. As expected, the optimal expected profit is always nondecreasing in $w_{\max}$ and $b_{\max}$.

DOI: 10.1002/asmb

6. CONCLUSION

In this paper, we have explored the pricing and capacity decisions in a service organization using an analytical framework built upon standard queueing models in the literature. On the basis of the steady-state performance measures, we have developed practical optimization models for coordinating the staffing and pricing decisions. It can be expected that rather than selecting the price and staffing level independently of each other, simultaneous consideration of these decisions will improve the profitability of a business organization.

We have considered service systems ranging from those with no waiting space to those with infinite waiting space. Constraints on the average waiting time and the blocking probability have been included to limit the level of customer dissatisfaction caused by excessive delays or busy servers. Under certain conditions of the parameters, we have shown the concavity of the objective function and convexity of the feasible region for each model in the single-variable optimization case and subsequently proposed solution procedures. We have also investigated structural properties of the optimal solution. When the number of servers is fixed and the elasticity of arrivals is increasing in price, the price maximizing the revenues is a lower bound on the optimal price. In our numerical study, using a linear demand curve and a linear staffing cost function, we have observed that the optimal price is nondecreasing in the service cost per customer, and the optimal number of servers is nonincreasing in the maximum allowable average waiting time and the maximum allowable blocking probability.

Future research may study the impact of different forms of demand and staffing cost functions. Another possibility is to extend the current single-class setting to the case where multiple customer classes with different service time requirements exist.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the valuable comments by an anonymous reviewer, which have greatly improved the paper.

REFERENCES

1. Wolff RW. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall: NJ, U.S.A., 1989.
2. Caro F, Simchi-Levi D. Static pricing for a network service provider. *Working Paper*, University of California, Los Angeles, Anderson School of Management, 2006. Available at: http://repositories.cdlib.org/anderson/dotm/FC05.
3. Kolesar PJ, Green LV. Insights on service system design from a normal approximation to Erlang's delay formula. *Production and Operations Management* 1998; **7**:282–293.
4. Jongbloed G, Koole G. Managing uncertainty in call centres using Poisson mixtures. *Applied Stochastic Models in Business and Industry* 2001; **17**:307–318.
5. Rossiter M. State-based management—a process for reducing customer waiting in over the counter service operations. *International Journal of Service Industry Management* 2003; **14**:458–470.
6. Taylor S. Waiting for service: the relationship between delays and evaluations of service. *Journal of Marketing* 1994; **58**(2):56–69.
7. Hui MK, Tse DK. What to tell consumers in waits of different lengths: an integrative model of service evaluation. *Journal of Marketing* 1996; **60**(2):81–90.
8. Fitzsimmons JA, Fitzsimmons MJ. *Service Management*: *Operations*, *Strategy*, *and Information Technology* (4th edn). Irwin/McGraw-Hill: NY, U.S.A., 2004.
9. Tadj L, Choudhury G. Optimal design and control of queues. *Top* (Spanish Statistical and Operations Research Society) 2005; **13**:359–414.

DOI: 10.1002/asmb

10. Artalejo JR, Orlovsky DS, Dudin AN. Multi-server retrial model with variable number of active servers. *Computers and Industrial Engineering* 2005; **48**:273–288.
11. Naor P. The regulation of queue size by levying tolls. *Econometrica* 1969; **37**:15–24.
12. Mendelson H. Pricing computer services: queueing effects. *Communications of the ACM* 1985; **28**:312–321.
13. Dewan S, Mendelson H. User delay costs and internal pricing for a service facility. *Management Science* 1990; **36**:1502–1517.
14. Stidham S. Pricing and capacity decisions for a service facility: stability and multiple local optima. *Management Science* 1992; **38**:1121–1139.
15. Ha AY. Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Science* 2001; **47**:915–930.
16. Palaka K, Erlebacher S, Kropp DH. Lead-time setting, capacity utilization and pricing decisions under lead-time dependent demand. *IIE Transactions* 1998; **30**:151–163.
17. So KC, Song J-S. Price, delivery time guarantees and capacity selection. *European Journal of Operational Research* 1998; **111**:28–49.
18. Ray S, Jewkes EM. Customer lead time management when both demand and price are lead time sensitive. *European Journal of Operational Research* 2004; **153**:769–781.
19. Larsen C. Investigating sensitivity and the impact of information on pricing decisions in an $M/M/1/\infty$ queueing model. *International Journal of Production Economics* 1998; **56–57**:365–377.
20. Ittig PT. Planning service capacity when demand is sensitive to delay. *Decision Sciences* 1994; **25**:541–559.
21. Jahnke H, Chwolka A, Simons D. Coordinating service-sensitive demand and capacity by adaptive decision making: an application of case-based decision theory. *Decision Sciences* 2005; **36**:1–32.
22. Carrizosa E, Conde E, Munoz-Marquez M. Admission policies in loss queueing models with heterogenous arrivals. *Management Science* 1998; **44**:311–320.
23. Ziya S, Ayhan H, Foley RD. Optimal prices for finite capacity queueing systems. *Operations Research Letters* 2006; **34**:214–218.
24. Borst S, Mandelbaum A, Reiman MI. Dimensioning large call centers. *Operations Research* 2004; **52**:17–34.
25. Kochel P. Finite queueing systems—structural investigations and optimal design. *International Journal of Production Economics* 2004; **88**:157–171.
26. Berman O, Larson RC. A queueing control model for retail services having back room operations and cross-trained workers. *Computers and Operations Research* 2004; **31**:201–222.
27. Marianov V, Serra D. Probabilistic maximal covering location-allocation models for congested systems. *Journal of Regional Science* 1998; **38**(3):401–424.
28. Marianov V, Serra D. Location-allocation of multiple-server service centers with constrained queues or waiting times. *Annals of Operations Research* 2002; **111**:35–50.
29. Marianov V, Serra D. Location models for airline hubs behaving as M/D/c queues. *Computers and Operations Research* 2003; **30**:983–1003.
30. Harel A, Zipkin PH. Strong convexity results for queueing systems. *Operations Research* 1987; **35**:405–418.
31. Dyer ME, Proll LG. On the validity of marginal analysis for allocating servers in M/M/c queues. *Management Science* 1977; **23**:1019–1022.
32. Topkis DM. Minimizing a submodular function on a lattice. *Operations Research* 1978; **26**:305–321.
33. Harel A. Convexity properties of the Erlang loss formula. *Operations Research* 1990; **38**:499–505.
34. Yao DD, Shanthikumar JG. The optimal input rates to a system of manufacturing cells. *INFOR* 1987; **25**:57–65.
35. Krishnan KR. The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates. *IEEE Transactions on Communications* 1990; **38**:1314–1316.
36. Messerli EJ. Proof of a convexity property of the Erlang B formula. *Bell System Technical Journal* 1972; **51**:951–953.
37. Pacheco A. Second-order properties of the loss probability in M/M/s/s+c systems. *Queueing Systems* 1994; **15**:289–308.
38. Chang C-S, Chao X, Pinedo M, Shanthikumar JG. Stochastic convexity for multidimensional processes and its applications. *IEEE Transactions on Automatic Control* 1991; **36**:1347–1355.
39. Shanthikumar JG, Yao DD. Second-order properties of the throughput of a closed queueing network. *Mathematics of Operations Research* 1988; **13**:524–534.
40. Bolton RN. The robustness of retail-level price elasticity estimates. *Journal of Retailing* 1989; **65**:193–219.
41. Grassman W. The convexity of the mean queue size of the M/M/c queue with respect to the traffic intensity. *Journal of Applied Probability* 1983; **20**:916–919.

DOI: 10.1002/asmb

42. Lee HL, Cohen MA. A note on the convexity of performance measures of M/M/c queueing systems. *Journal of Applied Probability* 1983; **20**:920–923.
43. Mandelbaum A, Shimkin N. A model for rational abandonments from invisible queues. *Queueing Systems* 2000; **36**:141–173.
44. Zohar E, Mandelbaum A, Shimkin N. Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science* 2002; **48**:566–583.
45. Hueter J, Swart W. An integrated labor management system for Taco Bell. *Interfaces* 1998; **28**:75–91.
46. van Mieghem JA. Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *Annals of Applied Probability* 1995; **5**:809–833.
47. Jagers AA, van Doorn EA. Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Review* 1991; **33**:281–282.

DOI: 10.1002/asmb