M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 11, No. 4, Fall 2009, pp. 674–693
ISSN 1523-4614 | EISSN 1526-5498 | 09 | 1104 | 0674

informs

DOI 10.1287/msom.1080.0246

Priority Assignment Under Imperfect Information on Customer Type Identities

Nilay Tanık Argon, Serhan Ziya
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599 {nilay@unc.edu, ziya@unc.edu}

In many service systems, customers are not served in the order they arrive, but according to a priority scheme that ranks them with respect to their relative "importance." However, it may not be an easy task to determine the importance level of customers, especially when decisions need to be made under limited information. A typical example is from health care: When triage nurses classify patients into different priority groups, they must promptly determine each patient's criticality levels with only partial information on their conditions.

We consider such a service system where customers are from one of two possible types. The service time and waiting cost for a customer depends on the customer's type. Customers' type identities are not directly available to the service provider; however, each customer provides a signal, which is an imperfect indicator of the customer's identity. The service provider uses these signals to determine priority levels for the customers with the objective of minimizing the long-run average waiting cost. In most of the paper, each customer's signal equals the probability that the customer belongs to the type that should have a higher priority and customers incur waiting costs that are linear in time. We first show that increasing the number of priority classes decreases costs, and the policy that gives the highest priority to the customer with the highest signal outperforms any finite class priority policy. We then focus on two-class priority policies and investigate how the optimal policy changes with the system load. We also investigate the properties of "good" signals and find that signals that are larger in convex ordering are more preferable. In a simulation study, we find that when the waiting cost functions are nondecreasing, quadratic, and convex, the policy that assigns the highest priority to the customer with the highest signal performs poorly while the two-class priority policy and an extension of the generalized $c\mu$ rule perform well.

Key words: priority queues; partially observable customer types; service differentiation; triage
History: Received: June 28, 2007; accepted: October 3, 2008. Published online in *Articles in Advance* February 10, 2009.

1. Introduction

There are many service systems where customers are not served in the order they arrive, but according to some priority scheme. Such systems typically aim to give priority to their "important" customers, who are more sensitive to delays than the others. However, in some cases, it is not possible to identify these customers perfectly, and thus service providers try to determine the relative importance of the customers based on the limited information about them.

There are various settings where priority decisions have to be made under imperfect information. Patients visiting emergency rooms or those who are injured in mass-casualty incidents go through a triage process where a triage nurse assigns them a priority level within a very short time based on limited information, mainly by checking basic signs and mostly relying on his or her visual impression of the patients (Frykberg 2002). Similarly, 911 operators have the difficult job of determining the relative importance of different calls based on the answers that the callers provide to some standard questions (Palumbo et al. 1996 and Reilly 2006). In some countries, similar classification problems arise in managing waiting lists for public-sector services under budget limitations. For example, in Australia, because of budget restrictions, patients cannot be given access to psychotherapy services as the need arises, and therefore they are put on waiting lists and are given priority according to their conditions (Walton and Grenyer 2002). The practice of giving priorities as opposed to following the first-come-first-served (FCFS) policy is not exclusive to service systems in the public sector. Many companies give preferential treatment to their "valuable"

customers or to those who show signs of dissatisfaction. For example, Charles Schwab Corporation's historically most profitable customers wait significantly less than the other customers for their calls to get answered (Brady 2000). On the other hand, for call centers, companies have been developing new tools, which monitor customers' conversations with customer representatives, detect deviations from the customers' regular speech and alert supervisors accordingly with the objective of identifying frustrated customers (or at least those who show signs of anger) and possibly give them preferential treatment (Shin 2006).

All these examples share a common characteristic: decision makers involved have to determine priorities under less-than-perfect information. Triage nurses do not know exactly who the critical patients are, but they simply make their best educated guesses. Some scoring systems have been proposed to help make triage decisions (e.g., the Trauma Score, the Circulation, Respiration, Abdomen, Motor, Speech (CRAMS) scale, and the Prehospital Index), but research has shown that these scoring systems do not work well in practice (Baxt et al. 1989). Similarly, companies do not actually know which customers are going to be more profitable in the future. Various metrics have been developed and adopted by the companies to help make that decision. Some examples are the recency frequency monetary value, share of wallet, past customer value, and customer lifetime value (CLV). Among these metrics, CLV has recently become more popular because it is being regarded as a more forward-looking metric as it tries to predict customers' future behavior, and takes this prediction into account together with any past information about the customers when determining customers' value to the company (Kumar 2008). Clearly, however, neither CLV nor any of the other metrics can perfectly identify the profitable customers. Foregoing priorities altogether is always an option, but in some cases, is not "affordable." When resources are severely limited, they need to be rationed in some way even when classification errors are inevitable. In many cases, for example, there is no alternative but to triage patients even when there is a high risk of making mistakes. For companies in the service sector, giving priorities may seem to be a choice rather than a necessity, but it appears that some are willing to take the risk of alienating some of their customers to keep the seemingly more valuable ones happy.

The objective of this paper is to provide some insights into this problem of assigning priorities under imperfect information. There are several questions of interest. For example, if we do not have perfect information on the customers, what kind of information about the customers can we use to classify them and how should we use this information to at least improve on the standard FCFS policy? What kind of information is more useful than others? How do the "optimal" prioritization policies change with certain system parameters such as the customer load?

To investigate these questions, we consider a queueing model where arriving customers belong to one of two different types: (1) type 1 customers are those who deserve to get the higher priority, and (2) type 2 customers are the ones who are supposed to be given lower priority. Both customer types have different service requirements and different delay sensitivities. Differing from most of the literature, we assume that customers' types are not directly observable. However, each customer provides a *signal*, which is a quantified summary of the relevant information about the customer. In most of the paper, the signal from each customer is the probability that the customer is of type 1, which can be computed using the information available. However, we also consider the possibility that even that information may not be available, and the service provider may have to use less informative signals that do not reveal the type probabilities, but a tendency to be of type 1 or 2 in some stochastic sense. Depending on the application, these signals can differ. For example, in patient triage, the signal can be a criticality score computed based on the patients' condition (similar to the way the CRAMS scale and other existing triage scores are computed). In call centers, it can be any past information about the customers (e.g., total purchases within the last year) that is an indicator of how valuable the customer is to the firm.

When each customer's signal represents the probability of that customer being of type 1 and the waiting costs are linear in time, we show that increasing the number of priority classes decreases the long-run average cost for the whole system and that the *highest-signal-first* (HSF) policy, which gives higher priority to those with higher probabilities of being type 1,

outperforms any priority policy with a finite number of priority classes. (We show in the paper that the HSF policy is, in fact, a generalization of the well-known $c\mu$ rule for priority queues.) Despite this result, however, policies with finite number of priority classes are still of interest because they are commonly used (e.g., in patient triage) and might be more practical in some cases. Furthermore, our numerical analysis suggests that switching from a two-class policy to HSF may not bring much benefits. Therefore we specifically analyze two-class priority policies, which are easy to implement and whose simple structure makes it easier to generate useful insights on the relationships between the optimal priority policies and certain system parameters and also on the characteristics of "good" signals. We find that the search for an optimal two-class policy can be reduced to a search for the optimal threshold value on the customer signal, which separates high-priority customers from the low-priority ones. Interestingly, it turns out that the optimal threshold value gets smaller as the system load increases and converges to zero as the load converges to 1. This means that as the load gets larger, more customers need to be classified as high priority, and for very high levels of customer load, only those customers with a very small probability of being type 1 are given lower priority.

The service provider can have alternative signals at her disposal, each computed by using different pieces of the available information or processing the same information differently. The question then arises as to which one of these signals to use. Are there any characteristics that "better" signals possess? We find that high variability is a desirable feature for a signal. Although it is not true that signals with higher variance necessarily lead to lower costs, we show that for two-class priority policies, the long-run average cost is lower under signals that are larger in convex ordering (which implies higher variance). Our numerical analysis suggests that the same insight holds for the HSF policy as well.

For systems where signals do not reveal customers' type probabilities, but are "weaker" indicators of type identities, we find conditions under which, signals can still be used to determine priorities. First, we find that if signals coming from type 1 customers are larger than signals coming from type 2 customers in likelihood ratio ordering, then higher signals imply higher probabilities of being type 1. Thus, even if the service provider does not know type probabilities of the customers, she can still order them according to their type 1 probabilities by simply ordering them according to their signals. However, if the ordering between the signals of the two types is weaker (e.g., only the hazard rate ordering or usual stochastic ordering holds between signals), this is no longer true. Nevertheless, as long as signals of type 1 customers dominate the signals of type 2 customers in the usual stochastic sense, we can show that by giving priority to the customer with the highest signal, the service provider cannot do worse than using FCFS.

For all of the results discussed above, we assume that customers' waiting cost functions are linear in time. Using simulation, we test how some of our findings for the linear cost case change when customers incur nonlinear costs. Our analysis suggests that when waiting costs are nondecreasing, quadratic, and convex with respect to time, HSF performs very poorly mainly because under HSF customers with low signals end up waiting for a very long time, which is penalized significantly under the convex cost structure. On the other hand, the optimal two-class priority policy performs surprisingly well under the nonlinear cost structure considered. This shows that the optimal two-class priority policy might be more robust than HSF because its performance under the linear cost structure is only slightly worse than that of HSF. However, the "best" policy (among all considered) turns out to be an extension of the *generalized $c\mu$ rule* proposed by Van Mieghem (1995), which outperforms all the other policies under the convex structure and reduces to HSF when costs are linear.

The rest of the paper is organized as follows. We first provide a review of the relevant literature in §2 and continue with the model description in §3. In §4, we characterize the structure of the optimal priority policies with a finite number of classes and show that the HSF policy outperforms any finite class policy. We provide our results on two-class priority policies in §5. We compare the performances of HSF and optimal two-class priority policies in §6. Section 7 deals with the comparison of different signals, i.e., identifying characteristics of useful signals, and §8 considers an alternative signal formulation

that assumes that signals do not reveal customers' type probabilities and investigates conditions under which such signals can be used to determine customer priorities. In §9, we report our findings based on a simulation study of a system where customers' waiting costs are convex in time. Finally, we provide our concluding remarks in §10. Proofs of all our results are presented in the appendix (online).

2. Review of Relevant Literature

Starting with Cobham (1954, 1955), priority queues have received significant attention. For a single-server queue with Poisson arrivals, where customers are classified into a finite number of priority classes and given nonpreemptive priority accordingly, Cobham (1954, 1955) derived expressions for the expected waiting times for each priority class. Many researchers followed by analyzing priority queues in various settings, and Cobham's (1954, 1955) results have been widely used in the literature, as we also do in this paper. Jaiswal (1968) is a good source for a review of early work on priority queues.

The optimality of the so-called "$c\mu$ rule" for multiclass queues with Poisson arrivals appears to have been first established by Cox and Smith (1961). According to the $c\mu$ rule, each customer class i has a priority index calculated by $c_i\mu_i$, where c_i is the per unit time cost of keeping class i customers wait and $1/\mu_i$ is the mean service time for class i customers, and higher priority is given to customer classes with higher priority indices. Other researchers have established the optimality of the $c\mu$ rule under various conditions and analyzed its extensions for more complex systems. Van Mieghem (1995) provides a review of this work and also establishes the asymptotic optimality of a generalized version of the $c\mu$ rule in a model where waiting costs are not necessarily linear but are convex in time. A number of papers have considered models where customers possibly differing in their delay sensitivities and utility functions purchase priorities. For examples of such work, see Kleinrock (1967), Balachandran (1972), Mendelson and Whang (1990), Rao and Petersen (1998), and Afèche (2007), and for an extended review, see Hassin and Haviv (2003).

The implicit assumption underlying the work that deals with the $c\mu$ rule and its various extensions is that customers' type identities are perfectly observable. Not much has appeared on priority systems when customers' identities are not available. In fact, to our knowledge, van der Zee and Theil (1961) is the only such work. In their paper, van der Zee and Theil (1961) consider a standard single-server queue with two priority classes with the additional feature that customers are possibly misclassified. They first assume that the arrival rate of customers who are supposed to be in class 1 but end up in class 2 and the arrival rate of customers who are supposed to be in class 2 but end up in class 1 are known, and then determine a condition under which prioritizing one of the classes is superior than the standard FCFS policy. Then, they carry out an approximate analysis (assuming very small misclassification rates) and propose a classification policy. The authors also consider a more general model where there are three priority classes, carry out another approximate analysis for small values of misclassification rates, and based on their analysis, propose that customers be classified to classes 1, 2, or 3, depending on the probability that an arriving customer should be classified as class 1. Our paper is fundamentally different from the work of van der Zee and Theil (1961). The main difference is that in our model, the service provider receives a signal from each arriving customer (where the signal indicates the probability distribution of the customer's true class identity or more generally it represents the information available about the customer) and assigns the customer a priority level, depending on this signal. In the model of van der Zee and Theil (1961), on the other hand, this priority assignment process is not modeled at all. They implicitly assume that the service provider classifies the customers in some unspecified way and knows the associated misclassification probabilities. Explicit modeling of the information about the customers allows us to generate various insights on the relationships between the available information, different classification policies, and improvements that would be obtained over the standard FCFS policy. Furthermore, unlike our analysis, the approximate analysis of van der Zee and Theil (1961) assumes very small misclassification rates and ignores the fact that the arrival rate of misclassified customers depends on the classification procedure used. Although van der Zee and Theil (1961)

suggest that similar analysis can be carried out if misclassification rates are large, they do not discuss how the dependence between the classification policy and error rates would be captured.

Within the general area of service operations, several authors have investigated systems where customers are classified into different groups, depending on the information available about the customers and their service requirements. The questions that these authors pose and investigate are different from ours, but our classification model shares some similarities with theirs. Shumsky and Pinker (2003) are interested in systems where a gatekeeper makes an initial diagnosis of each incoming customer and decides whether to serve the customer herself or send him to a specialist. If the gatekeeper chooses to serve the customer herself, she takes the risk of not serving the customer satisfactorily, and as a result, incurring a cost. The gatekeeper can rank the customers according to the complexity of the service they require and she knows that she can successfully serve a customer with a complexity level k with probability $f(k)$. (The complexity information in this model is in some sense similar to customer signals in our model, where each signal corresponds to a certain probability that the customer carrying the signal belongs to type 1.) Shumsky and Pinker (2003) are interested in developing incentive mechanisms that will induce the gatekeeper to act in a way that is optimal for the overall system when there is information asymmetry between the gatekeeper and the firm. Although Shumsky and Pinker (2003) do not model queueing effects explicitly, in a related paper, Hasija et al. (2005) consider queueing effects in a similar model, and investigate how optimal referral rates change with system parameters.

Güneş and Akşin (2004) use queueing-based models to investigate value creation/service-delivery design questions under congestion effects. In one of their models, each customer is classified into one of two types (high and low), depending on whether the probability that the customer will generate revenue when offered a high-level service is above a certain level θ or not. The server then further decides whether to offer an extended or regular service to the customers, depending on their type identities. Güneş and Akşin (2004) investigate how the server should determine the type of service each customer type receives, and how the manager should determine the value of θ and what kind of incentive mechanisms she should offer to the server so as to maximize long-run average profits.

Service level differentiation in various forms have received significant attention most recently within the context of call centers. For some examples, other than the above, see Gans and Zhou (2003, 2007) and Gurvich et al. (2005). Also, see Akşin et al. (2007) for a recent survey of work on call centers.

Finally, we would like to note that priority queues have been previously used with the objective of prioritizing emergency calls, but with models that do not consider the possibility of misclassification. For examples of such work, see Green (1984) and Schaack and Larson (1986, 1989).

3. Model Description

We consider a service system, which can be modeled as a single-server queue. Customers arrive according to a Poisson process with rate λ, and each customer is either of type 1 with probability p_1 or type 2 with probability $p_2 = 1 - p_1$ independently of other customers' types, the arrival and service processes, and system state. Service times of type $i \in \{1, 2\}$ customers are independent and identically distributed (i.i.d.) with finite first and second moments given by a_i and e_i, respectively. We define $\rho = \lambda(p_1 a_1 + p_2 a_2)$ to be the system load (or customer load) and we assume that $\rho < 1$, so that the system is stable. We use h_i to denote the per unit time cost of having a type i customer wait, and without loss of generality, we assume that $h_1/a_1 > h_2/a_2$. We also assume that the service is nonpreemptive, i.e., once the service of a customer begins, it cannot be interrupted. The performance measure of interest is the long-run average expected waiting cost.

The $c\mu$ rule assumes that customer types are perfectly observable, and for the system described above, the rule says that the optimal policy is to give priority to type 1 customers whenever there is at least one such customer in the system at the end of a service. In our model, however, types of customers are not directly observable. Instead, the service provider has some partial information about each customer, and uses this information to determine the *probability* that the customer belongs to type 1 (or equivalently the probability that the customer belongs to type 2). We

refer to this probability as the customer *signal*. Thus the signal is an imperfect measure of the customer's "importance," i.e., a probabilistic indicator of whether the customer is of type 1 or 2, and can be computed using historical data as well as specific information about the customer. For example, call centers can identify the purchasing habits of their premium ("valuable") customers based on historical data, and can use this information along with the available data on each individual customer to determine the probability that the customer is a premium customer or not. One common method that can be used to estimate the probability of a customer's type identity is logistic regression (see, e.g., Hosmer and Lemeshow 2000 and Hastie et al. 2001).

For each arriving customer, the signal, i.e., the probability that the customer belongs to type 1, is a random variable and we assume that it is i.i.d. for all customers with a probability density function $b(\cdot)$ and a strictly increasing cumulative distribution function $B(\cdot)$. Note that $B(\cdot)$ is the probability distribution of the probability of an arbitrary customer belonging to type 1, and thus $B(x)$ is the probability that an arriving customer has less than $100x$ percent chance of being type 1. By definition, we also have

$$p_1 = \int_0^1 xb(x)\,dx \quad \text{and} \quad p_2 = \int_0^1 (1-x)b(x)\,dx.$$

The signal distribution $B(\cdot)$ can be estimated by taking a random sample of customers and fitting a distribution to the estimated type 1 probabilities of these customers. Alternatively, one can first estimate the probability distributions of the covariates in the logistic regression (which is used to estimate type 1 probabilities), which can, in turn, be used to estimate the probability distribution of the signal.

In §§4–7, we use the model described above to obtain insights on "good" priority policies and also signals that yield the smallest long-run average cost. Later in §§8 and 9, we relax the assumption of availability of the distribution of probability of being a type 1 customer and the assumption of linear waiting costs, respectively.

4. Grouping Customers into Priority Classes and the HSF Policy

Consider a policy that prioritizes customers so that customers with higher signals have priority over customers with lower signals. In other words, whenever the server is available, the server picks the customer with the highest signal among those waiting for service. We call this policy the HSF policy. The HSF policy is in some sense a generalization of the $c\mu$ rule because giving higher priority to customers with higher signals is actually the same as giving higher priority to customers with higher "expected $c\mu$" values, because using our notation, the expected $c\mu$ value for a customer with signal x equals $x(h_1/a_1) + (1-x)(h_2/a_2)$, which increases with x.[1] It is thus reasonable to expect HSF to perform well. Indeed, as we demonstrate in this section, HSF outperforms all finite class priority policies. Toward that end, we first give a precise description of finite class policies, prove some structural properties of optimal finite-class policies, and investigate how the performances of these policies change as the number of classes increases.

An N-class priority policy π (where $1 \le N < \infty$) can be characterized by a partitioning of the interval $[0, 1]$ into N subintervals $I_{j,\pi}$, $j \in \{1, \ldots, N\}$. (The case with $N = 1$ corresponds to the FCFS policy.) To be more precise, an N-class priority policy π can be described by a set of signal sets $\{I_{1,\pi}, I_{2,\pi}, \ldots, I_{N,\pi}\}$, where $I_{j,\pi} \subset [0, 1]$ for $j \in \{1, 2, \ldots, N\}$; $I_{j,\pi} \cap I_{k,\pi} = \varnothing$ for any $j, k \in \{1, 2, \ldots, N\}$, $j \ne k$; $\bigcup_{j=1}^{N} I_{j,\pi} = [0, 1]$; and where customers whose signals belong to set $I_{j,\pi}$ are put into priority class j and have higher priority than customers from any class $k > j$. Let Π_N denote the class of all such policies for fixed N. Let also $W_{j,\pi}$ denote the steady-state expected queueing time of customers in priority class j under policy $\pi \in \Pi_N$. Then, using results on nonpreemptive priority queues (see, e.g., Cobham 1954 or Wolff 1989, §10.2), we have

$$W_{j,\pi} = \frac{\lambda(p_1 e_1 + p_2 e_2)}{2\left(1 - \lambda \sum_{k=1}^{j-1} \Gamma_{k,\pi}\right)\left(1 - \lambda \sum_{k=1}^{j} \Gamma_{k,\pi}\right)}, \quad \text{for } j = 1, 2, \ldots, N, \quad (1)$$

[1] The reader can check that HSF is equivalent to implementing the expected $c\mu$ rule also by computing the expected $c\mu$ values after computing the expected cost and expected service time individually. This gives $(xh_1 + (1-x)h_2)/(xa_1 + (1-x)a_2)$ as the index value of a customer with a signal of x.

where $\Gamma_{k,\pi} = \int_{I_{k,\pi}} (xa_1 + (1-x)a_2)b(x)\,dx$ for $k = 1, \ldots, N$. Now, define C_π to be the long-run average cost under policy $\pi \in \Pi_N$. Then

$$C_\pi = \lambda \sum_{k=1}^{N} W_{k,\pi} \int_{I_{k,\pi}} (h_1 x + h_2(1-x))b(x)\,dx. \quad (2)$$

Using (1) and (2), we next obtain a result that partially characterizes an optimal policy within Π_N. (The proofs of all our results are provided in the appendix online.) First, define $\widetilde{\Pi}_N$ to be the set of policies for which there exists a sequence of $N+1$ real numbers $t_0 > t_1 > t_2 > \cdots > t_{N-1} > t_N$ such that $I_{1,\pi} = [t_1, t_0]$, $I_{j,\pi} = [t_j, t_{j-1})$ for $j = 2, \ldots, N$, $t_0 = 1$, and $t_N = 0$. Note that $\widetilde{\Pi}_N \subset \Pi_N$ and each policy in $\widetilde{\Pi}_N$ is completely characterized by $N-1$ "thresholds," i.e., $t_1, t_2, \ldots, t_{N-1}$.

Theorem 1. *For fixed $N \geq 2$, there exists a policy π^* in $\widetilde{\Pi}_N$ that provides the smallest long-run average cost among all policies in Π_N. In other words, there exists an optimal policy in Π_N for which signal interval $[0, 1]$ is partitioned into N disjoint intervals by $N-1$ threshold values, so that customers whose signals fall into intervals with higher signals have higher priority than customers whose signals fall into intervals with lower signals.*

Theorem 1 simplifies the search for an optimal policy because it gives a characterization of a class of policies (i.e., $\widetilde{\Pi}_N$), which contains at least one optimal policy (if not the only one) and is much smaller than Π_N. The theorem says that it is sufficient to search for optimal values of $N-1$ thresholds that will separate the signal interval for one priority level from that of the next priority level.

Next, we investigate the effects of increasing the number of classes on the long-run average cost.

Theorem 2. *Let π be any N-class policy in $\widetilde{\Pi}_N$ characterized by thresholds $t_1, t_2, \ldots, t_{N-1}$. Also, let $\hat{\pi}$ be an $(N+1)$-class policy in $\widetilde{\Pi}_{N+1}$ obtained from π by partitioning one of its class intervals $[t_m, t_{m-1})$ into two subintervals $[t_m, \bar{t})$ and $[\bar{t}, t_{m-1})$, where $\bar{t} \in (t_m, t_{m-1})$ and $m = 1, 2, \ldots, N$ such that customers with signals in $[\bar{t}, t_{m-1})$ receive a higher priority than those with signals in $[t_m, \bar{t})$, while all other priority relations remain the same as in π. Then, the long-run average cost under $\hat{\pi}$ is at most the same as that under π.*

Corollary 1. *For every N-class priority policy in Π_N (where $N \geq 1$), there exists an $(N+1)$-class priority policy in Π_{N+1} under which the long-run average cost is at most the same.*

Corollary 2. *The long-run average cost under any N-class priority policy in $\widetilde{\Pi}_N$ for $N \geq 2$ is at most the same as that under the* FCFS *policy.*

Theorem 2 says that any N-class priority policy in $\widetilde{\Pi}_N$ can be improved by adding a new class by arbitrarily partitioning the signal interval corresponding to any one of the classes into two, and thereby creating an $(N+1)$-class policy (while still enforcing that customers whose signals fall into intervals with higher signals have higher priorities). Note that the theorem does not say that any $(N+1)$-class policy is better than any N-class policy, which is not correct even if the policies are restricted to be in policy sets $\widetilde{\Pi}_{N+1}$ and $\widetilde{\Pi}_N$, respectively. However, as Corollary 1 directly implies, the optimal $(N+1)$-class policy is better than the optimal N-class policy. We also know that any N-class policy in $\widetilde{\Pi}_N$ is better than FCFS as stated in Corollary 2.

Now, consider the HSF policy. This policy can, in fact, also be considered as a priority policy with infinite number of classes because each customer can be seen as belonging to a different class. Corollary 1 says that any finite class priority policy can be improved by adding another class, and Theorem 1 says that customers with lower signals should not have higher priorities than those with higher signals. These two findings lead to the following result, which is formally proven in the appendix online.

Theorem 3. *The long-run average cost under the nonpreemptive policy that gives higher priority to customers with higher signals (i.e., the HSF policy) is at most as large as the long-run average cost under any finite class priority policy.*

Findings of this section suggest the following: When waiting costs change linearly in time, use HSF preferably, if not, offer as many priority classes as possible. Note, however, that these findings ignore the "cost" associated with implementing different priority policies. One can imagine that implementation of a priority policy can become more difficult as the number of priority classes increases, and thus the

benefits of having additional priority classes might not be sufficient to overcome the additional "costs." (Our numerical analysis in §6 also strongly supports this claim.) Thus, in many service systems, the number of priority classes is not too large. For example, in patient triage, while different triage systems have different numbers of priority classes, this number typically does not exceed six. In the following section, we focus on two-class priority policies. Relatively simpler structure of this class of policies allows us to give a complete characterization of the optimal policy and derive insights regarding the relationships between the optimal policy and certain system parameters.

5. Priority Policies with Two Classes

Suppose that the service provider is employing a two-class priority policy. Then, we know from Theorem 1 that there exists an optimal policy that is completely characterized by a single-threshold value $t \in [0, 1]$. Those with signals t or above are classified as class 1, while those with signals below t are classified as class 2 and class 1 customers have nonpreemptive priority over class 2 customers.

Define $\mathcal{W}_i(t)$ to be the expected waiting time of class i customers under threshold t. Then, from (1), we have

$$\mathcal{W}_1(t) = \frac{\lambda(p_1 e_1 + p_2 e_2)}{2\left(1 - \lambda \int_t^1 (a_1 x + a_2(1-x)) b(x)\, dx\right)} \tag{3}$$

and

$$\mathcal{W}_2(t) = \frac{\mathcal{W}_1(t)}{1-\rho}. \tag{4}$$

Let $C(t)$ denote the long-run average cost when the threshold level is set to $t \in [0, 1]$. Then, we have

$$C(t) = \lambda \mathcal{W}_1(t) \int_t^1 (h_1 x + h_2(1-x)) b(x)\, dx + \lambda \mathcal{W}_2(t) \int_0^t (h_1 x + h_2(1-x)) b(x)\, dx. \tag{5}$$

Our objective is to minimize $C(t)$ with respect to the threshold t. (Note that when $t = 1$, there are no class 1 customers, and therefore no customer experiences (3) as the expected waiting time and similarly when $t = 0$, there are no class 2 customers and as a result no customer experiences (4) as the expected waiting time.)

We next show that $C(t)$ is a unimodal function of t and provide an expression for the optimal threshold value using the first-order condition. We first define

$$E(t) = \int_t^1 z b(z)\, dz \tag{6}$$

and

$$\bar{E}(t) = \int_t^1 (1-z) b(z)\, dz \quad \text{for } t \in [0, 1]. \tag{7}$$

Proposition 1. *Let t^* be the unique solution to $g(\rho, t^*) = 0$, where*

$$g(\rho, t) = t(p_2 - \rho \bar{E}(t)) - (1-t)(p_1 - \rho E(t)). \tag{8}$$

Then, t^* is the unique optimal threshold that minimizes $C(\cdot)$.

Proposition 1 provides several insights on the optimal threshold t^*. Given the assumption that $h_1/a_1 > h_2/a_2$, t^* is independent of h_1 and h_2. In other words, the optimal threshold does not depend on how much more "costly" it is to keep higher priority customers waiting. On the other hand, if $h_1/a_1 < h_2/a_2$, this implies that priority is not given to the right customers. In that case, t^* as defined in Proposition 1 maximizes (5). Another interesting observation is that the optimal threshold depends on customer arrival rate λ and mean service times a_1 and a_2 through the system load ρ only. Furthermore, higher moments of service times have absolutely no effect on the optimal threshold t^*.

We next investigate the effect of the system load on t^*.

Proposition 2. *Fix $p_1 \in [0, 1]$. Then, the optimal threshold t^* decreases as the system load ρ increases. Furthermore, t^* converges to p_1 as the system load ρ approaches to 0, and t^* converges to 0 as the system load ρ approaches to 1.*

Proposition 2 states that as the system load increases, the optimal decision calls for classifying a higher percentage of customers as class 1. Also, in the limit, as ρ converges to 1, t^* converges to 0. Therefore, under heavy traffic, policies that classify a small percentage of customers as class 2 are more desirable. On the other hand, as ρ converges to 0, t^* converges to p_1 meaning that $t^* < p_1$ (or equivalently $1 - t^* > p_2$) for any $\rho > 0$. Thus, for any level of customer load, for

a customer to be classified as class 2 under the optimal classification policy, the probability of that customer being of type 2 has to be larger than the fraction of type 2 customers in the whole customer population.

Although Proposition 2 may seem surprising at first, it has an intuitive explanation. As the customer load increases, all customers experience longer waiting times, but this is especially the case for those customers who are in class 2. With increasing load, the "cost" of a misclassified type 1 customer becomes even higher. However, if the threshold is set at a very low level, and thus the probability of a customer being classified as class 2 is very small, even though customers will still experience longer waiting times with increasing load, it is very unlikely that a type 1 customer will have to bear a significantly longer waiting time because of misclassification, as only those customers who are very likely to be of type 2 are classified as class 2. Thus, with increasing customer load, the policies that classify fewer customers as class 2 become increasingly more preferable.

Next, we provide a lower bound on the magnitude of improvement brought by the optimal two-class policy over the FCFS policy. We then use this lower bound to identify conditions under which the optimal two-class priority policy does not bring significant benefits.

Proposition 3. *We have*

$$\frac{p_2(t^*/(1-t^*))h_1 + p_2h_2}{p_1h_1 + p_2h_2} \le \frac{C(t^*)}{C_{\mathrm{FCFS}}} \le 1,$$

where C_{FCFS} *is the long-run average cost under the FCFS policy and equals* $C(0)$.

Corollary 3. *When either* ρ *converges to* 0 *while* p_1 *is fixed or* p_1 *converges to* 0 *while* ρ *is fixed, the fraction* $C(t^*)/C_{\mathrm{FCFS}}$ *approaches to* 1.

Corollary 3 implies that for low values of the system load, benefits from the priority policy will be very small. This is intuitive because when the system load is low, customers' queueing time will be low as well regardless of their priority level. Hence there is not much to gain from giving priorities. Corollary 3 also suggests that the benefit of the priority system is insignificant when the fraction of type 2 customers is much higher than that of type 1 customers. In the following section, we give a more detailed discussion on the benefits of using priority policies by means of a numerical analysis.

6. A Numerical Study on the Performance of HSF and Two-Class Priority Policies

As we have established in §4, increasing the number of priority classes increases the performance of the optimal priority policy. We do not know, however, how much improvement is possible and how it depends on system characteristics such as the customer load and ratio of types 1 and 2 customers. The objective of this section is to shed some light on these questions by reporting our findings based on a numerical study.

We first obtain an expression for C_{HSF}, which denotes the long-run average cost under the HSF policy.

$$C_{\mathrm{HSF}} = \lambda \int_0^1 (h_1x + h_2(1-x))W(x)b(x)\,dx,$$

where $W(x)$ is the steady-state expected queueing time of a customer with signal x under the HSF rule. In the following proposition, we obtain an expression for $W(x)$. Proposition 4 generalizes Theorem 1 of Kleinrock (1967), which provides an expression for $W(x)$ when there is a single customer type.[2]

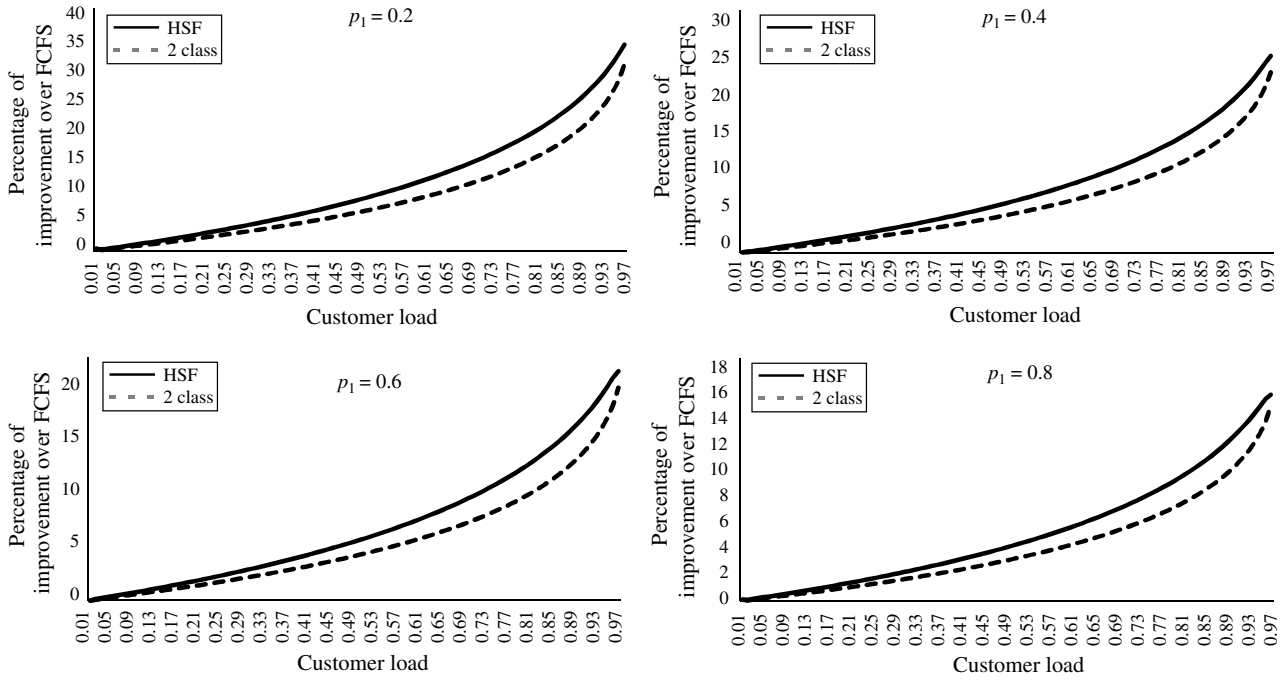
Proposition 4. *Under the HSF rule, the expected waiting time for a customer with signal* $x \in [0,1]$ *is given by*

$$W(x) = \frac{\lambda(p_1e_1 + p_2e_2)}{2\left(1 - \lambda\int_x^1 (ya_1 + (1-y)a_2)b(y)\,dy\right)^2}. \quad (9)$$

In our numerical study, we considered various scenarios by assigning different values for system parameters such as h_1, h_2, a_1, a_2, and λ. We observed some common characteristics that these different scenarios exhibited. Here, we report these observations over a representative example. For this example, the holding cost rates are given by $h_1 = 4$ and $h_2 = 1$, service times are exponentially distributed for both types with the

[2] Kleinrock (1967) is interested in optimal bribing for queueing position in a single-server queue. He assumes that there is a probability distribution for customer bribes, and he derives an expression for the steady-state expected waiting time of a customer with a bribe of x. Bribes in his model can be seen as signals in ours, however, he assumes that service times are independent of customer bribes, which does not hold for the signals in our model.

Figure 1 Percentage Improvement Over FCFS Under HSF and Optimal Two-Class Priority Policies for Different Levels of Customer Load



same mean of one unit, and signals are uniformly distributed over the interval $[p_1 - 0.2, p_1 + 0.2]$.

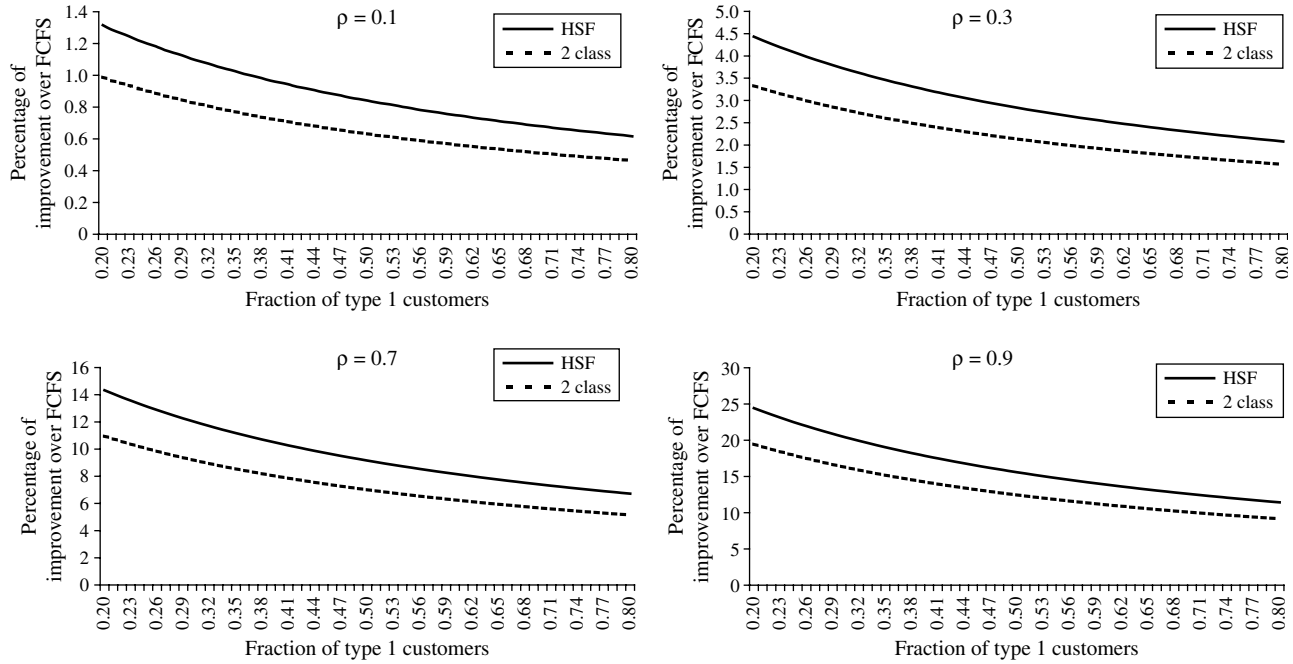
We first observe the effect of changes in the customer load. To isolate the effects of the customer load, we fix p_1, the ratio of type 1 customers to all customers, to one of $\{0.2, 0.4, 0.6, 0.8\}$, in turn, and for each fixed ratio, we let λ (and equivalently traffic load ρ) change from 0.01 to 0.99. Figure 1 shows the percentage improvement in costs that would be obtained using HSF and the optimal two-class priority policies over the standard FCFS policy.

We observe that under both priority policies, improvements in cost increase with customer load. In particular, regardless of the value of p_1, the percentage improvement appears to be an increasing convex function of customer load ρ, which takes substantially large values under heavy customer load. This finding may sound contradictory to Proposition 2, which states that as the system load ρ approaches to 1, t^* converges to 0, implying that the optimal two-class priority policy becomes similar to the FCFS policy. Note, however, that even under a small threshold level, the priority policy will classify those customers who have a very high chance of being type 2 as class 2, and thereby reduce the waiting time of type 1 customers, which will result in substantial savings in a setting where the waiting times are already very large because of heavy customer load.

Figure 1 is also insightful about how the performance of the optimal two-class priority policy compares to that of the HSF policy. Because we know that HSF outperforms any finite class priority policy, its performance constitutes an upper bound on the performance of any finite class priority policy for a given signal distribution. Interestingly, for almost all levels of customer load, we observe that the optimal two-class priority policy captures most of the potential improvement in costs. This suggests that classifying customers into two priority classes can be a quite satisfactory alternative if keeping a strict ordering of the customers with respect to their signals or using a finite class priority policy with more than two classes is impractical or costly. The performances of HSF and the optimal two-class priority policies appear to be especially close when the customer load is either light or very heavy. Relatively speaking, the advantage of using HSF over the two-class priority policy is more significant for mid to high levels of customer load.

Next, we investigate the effect of the customer mix (more specifically the fraction of type 1 customers, p_1)

Figure 2 Percentage Improvement Over FCFS Under HSF and Optimal Two-Class Priority Policies for Different Levels of Type 1 Customer Fraction



on the percentage improvement over the FCFS policy. This time, to isolate the effects of changes in the customer mix, we fix the arrival rate (and thus customer load ρ) to one of $\{0.1, 0.3, 0.7, 0.9\}$, in turn, and for each fixed load, we let p_1 change from 0.2 to 0.8. Note that because we assume that the signal distribution is uniform over $[p_1 - 0.2, p_1 + 0.2]$, the variance of the signal does not change as we change p_1. This is important to isolate the effect of change in the fraction of type 1 customers, because as we later discuss in §7, the effectiveness of a signal is closely related with its variability.

Figure 2 shows the percentage improvement in costs that would be obtained using HSF and the optimal two-class priority policies over the standard FCFS policy. For all values of customer load ρ, the behavior of percentage improvement with respect to the fraction of type 1 customers appears to have a convex decreasing shape with more substantial improvements when the fraction of type 1 customers is smaller. This observation makes intuitive sense. When type 1 customers constitute a significantly large percentage of all the customers, because type 1 customers are mostly inconvenienced by other type 1 customers anyway, the improvements brought by the priority policies are relatively small. However, if type 1 customers are in minority, they are mostly kept waiting by type 2 customers under the FCFS policy, and thus identifying these type 1 customers (to the extent that is possible) and giving them higher priority brings more significant benefits. Note that there are examples that show that this insight is no longer true if the variance of the signal changes with the fraction of type 1 customers (unlike in our experimental setting where the variance is fixed at 1/75). Indeed, Corollary 1 states that the improvements brought by priority policies become insignificant as the fraction of type 1 customers approaches zero. This is not contradicting the numerical results that we present here, because the fact that p_1 approaches zero implies that the variance also approaches zero (because the second moment of the signal is always less than or equal to p_1) unlike the case in our numerical results.

7. Comparison of Signals

By using different pieces of the customer data available or processing the same data differently, the service provider can develop different ways of coming up with a signal, i.e., the probability that a given customer is of type 1. In other words, there can be more than one type of signal available for determining priorities of

customers. Then, an interesting question is which one of these signals would lead to the smallest long-run average cost.

To be more precise, suppose that we would like to compare two signals, signal Y and signal Z, with cumulative distribution functions $B_Y(\cdot)$ and $B_Z(\cdot)$, and probability density functions $b_Y(\cdot)$ and $b_Z(\cdot)$, respectively. These two signal distributions are related by the fact that they have the common mean p_1; i.e.,

$$p_1 = \int_0^1 x b_Y(x)\,dx = \int_0^1 x b_Z(x)\,dx.$$

Thus we assume that both signals correctly estimate the true proportion of type 1 customers, and therefore it is fair to determine which signal is better by comparing the associated long-run average costs.

We first make this comparison for a service provider that employs a two-class priority policy in $\widetilde{\Pi}_2$. Define $E_Y(\cdot)$ and $E_Z(\cdot)$ as in (6), but for signals Y and Z, respectively. Similarly, define $\bar{E}_Y(\cdot)$ and $\bar{E}_Z(\cdot)$ as in (7), but for signals Y and Z, respectively. Let $C_i(t_i)$ denote the long-run average cost when threshold t_i is used for signal i, where $i \in \{Y, Z\}$. Then, we can show that (see Appendix C online)

$$C_Y(t_Y) - C_Z(t_Z) = \frac{\lambda^3(p_1e_1+p_2e_2)(h_1a_2-h_2a_1)\nu(t_Y,t_Z)}{2(1-\rho)(1-\lambda(a_1E_Y(t_Y)+a_2\bar{E}_Y(t_Y)))(1-\lambda(a_1E_Z(t_Z)+a_2\bar{E}_Z(t_Z)))}, \tag{10}$$

where

$$\nu(t_Y,t_Z) = (p_2-\rho\bar{E}_Y(t_Y))(E_Z(t_Z)-E_Y(t_Y)) - (p_1-\rho E_Y(t_Y))(\bar{E}_Z(t_Z)-\bar{E}_Y(t_Y)). \tag{11}$$

Since $h_1/a_1 > h_2/a_2$, the sign of $C_Y(t_Y) - C_Z(t_Z)$ is determined by the sign of $\nu(t_Y, t_Z)$. Hence, signal Y (when applied with threshold t_Y) is better than signal Z (when applied with threshold t_Z) if and only if $\nu(t_Y, t_Z) < 0$. Now, suppose that we would like to compare the performances when threshold values t_Y and t_Z are set optimally. Let t_Y^* and t_Z^* denote the corresponding optimal values. Then, from (11) and Proposition 1, after some algebra, we obtain

$$\nu(t_Y^*, t_Z^*) = \frac{(p_1-\rho E_Y(t_Y^*))(p_1-\rho E_Z(t_Z^*))(t_Y^*-t_Z^*)}{\rho t_Y^* t_Z^*},$$

which implies that $\nu(t_Y^*, t_Z^*) < 0$ (and thus signal Y is better than signal Z) if and only if $t_Y^* < t_Z^*$. We can thus conclude the following.

Theorem 4. *Suppose that the service provider decides to implement the optimal two-class priority policy $\pi \in \widetilde{\Pi}_2$, which is characterized by a single-threshold value. Then, among all available signals, the one under which the optimal value of the threshold is the smallest provides the smallest long-run average cost.*

Theorem 4 provides an easy-to-use procedure for determining which signal to pick among a finite set of alternatives: Calculate the optimal threshold value for each alternative first and then use the signal under which the optimal threshold is the smallest. Theorem 4 is also useful in obtaining insights on the characteristics of "good" signals. In particular, we can identify conditions on signals under which the corresponding optimal threshold values are guaranteed to be ordered. The following proposition gives one such condition. (See Appendix A online for a definition of convex ordering.)

Proposition 5. *Suppose that signal Y is larger than signal Z in the convex order; i.e., $B_Y \geq_{cx} B_Z$. Then, the optimal threshold for signal Y is smaller than the optimal threshold for signal Z, i.e., $t_Y^* \leq t_Z^*$. Consequently, the minimum long-run average cost under signal Y is smaller than the minimum long-run average cost under signal Z when a two-class priority policy is applied.*

Proposition 5 says that the long-run average cost is smaller for signals that are larger in the sense of convex ordering. Note that because we are also assuming that the expected values of the signals are the same, assuming convex ordering is, in fact, equivalent to assuming increasing convex ordering (see Theorem 1.5.3 of Müller and Stoyan 2002), which is easier to verify than convex ordering.

Proposition 5 does not only give us a technical condition to compare signals, but also provides some general insights because convex ordering is, in fact, closely related to the "variability" or "dispersion" of random variables. In particular, random variables that are larger in convex order have larger variances (see Corollary 1.5.4 of Müller and Stoyan 2002). Thus, Proposition 5 points to a close relationship between the variability of the signal and its usefulness, which makes intuitive sense. If a signal distribution is not spread out, then that signal may not be very helpful. For example, consider the extreme case where the signal is deterministic, which means that all arriving

customers have the same probability of being type 1. In such a case, there is no basis to differentiate the customers, and thus the signal is not helpful. On the other hand, if the signal distribution is well spread, then customer signals provide useful information. In the extreme case, for example, type 1 customers will give signal 1, and type 2 customers will give signal 0, which leads to a priority policy in which customers are perfectly classified according to their types.

Note that convex ordering is one of the weakest univariate variability orders. In particular, it is weaker than the excess wealth ordering and dispersive ordering (see Equations (3.C.8) and (3.C.9) in Shaked and Shanthikumar 2007). However, it is stronger than the ordering of variances, and therefore it is of interest to investigate whether a variance order without the convex ordering would be sufficient to rank alternative signals. Although, in general, it might be reasonable to believe that signals with higher variances will tend to be more useful, there are examples that show that this is not always the case (see Appendix C online).

Based on Theorem 4 and Proposition 5, it is reasonable to conjecture that when the service provider uses signals that are larger in the convex order, the long-run average cost will be smaller not only for two-class priority systems but also for systems with more than two priority classes. While our numerical analysis on the HSF policy supported this conjecture, proving it appears to be quite challenging. Similarly, as in the case of two-class priority policies, we found that higher variance does not necessarily imply that long-run average cost under the HSF policy will be smaller, but for many cases, we observed that long-run average cost is smaller with signals that have larger variance. We next discuss this numerical analysis in more detail.

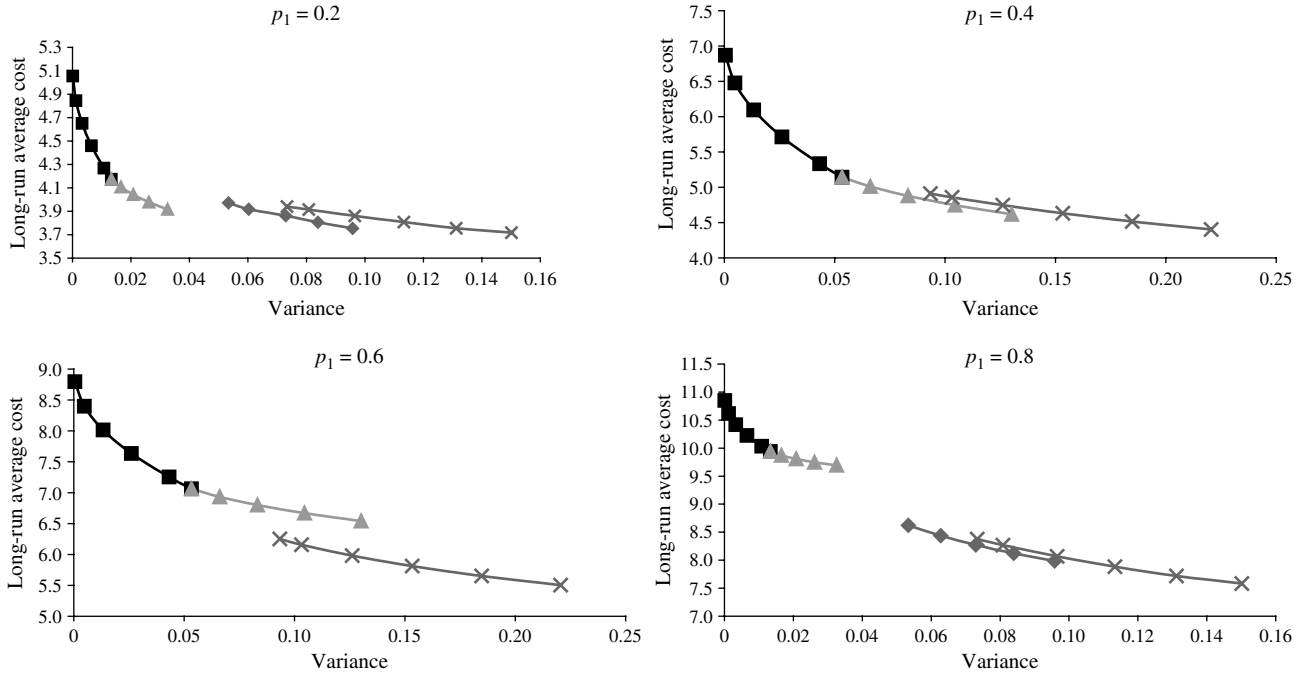
For our numerical analysis, we assumed that $h_1 = 4$, $h_2 = 1$, $\lambda = 0.8$, $a_1 = a_2 = 1$, and $e_1 = e_2 = 2$. We set p_1 to one of the four different values from the set $\{0.2, 0.4, 0.6, 0.8\}$. For each fixed value of p_1, we identified a number of probability distributions, each defined over a subset of the interval $[0, 1]$ with a mean of p_1 but with a different variance. We then computed the long-run average cost under each distribution. To cover a wide range of variance values for each fixed value of p_1, we considered four different families of distributions. For example, Family 1 is the set of distributions that are uniform over the interval $[p_1 - s, p_1 + s]$, where $0 < s < \min(p_1, 1 - p_1)$ and distributions from this family are obtained by choosing different values of s. Other families of distributions are obtained by modifying Family 1 in various ways so as to obtain different variance values for each fixed value of p_1. All these distributions are uniform restricted to certain intervals but with possibly different uniform rates for different intervals. (A complete description of these different families of distributions are given in Appendix C online.) For each fixed value of p_1, Figure 3 gives a scatter plot of the long-run average cost under the HSF policy with respect to the signal variance, where each variance value corresponds to a different distribution. (Note that, in Figure 3, there are four types of markers each corresponding to a different family. Observations made for signal distributions from the same family are indicated by the same type of marker.) As it can be seen from Figure 3, overall the long-run average cost tends to decrease as variance increases, but there are some exceptions. For example, in the lower left plot, where $p_1 = 0.6$, one can see that the distribution corresponding to the left-most cross ($\times$) marker has a smaller variance than the distribution corresponding to the right-most triangle marker, but the long-run average cost under the former distribution is lower than that under the latter. One can check to see that convex order does not hold for any of such exceptions, and therefore none of them disproves our conjecture.

8. An Alternative Signal Formulation: When Signals Do Not Reveal Type Probabilities

So far in this paper, we have assumed that each customer's signal is equal to that customer's probability of being type 1. In this section, we consider the possibility that customer signals are less informative. Our objective is to identify conditions under which these less informative signals are still useful and describe how they can be used to assign priorities.

We first change our signal formulation in a way that will allow us to "weaken" the signals that we have taken to be type 1 probabilities. Suppose now that each customer's signal is a quantified summary of the relevant information that the service provider has about the customer. More precisely, it is an imperfect measure of the customers' "importance," i.e., an imperfect

Figure 3 Scatter Plot of the Long-Run Average Cost with Respect to the Signal Variance



indicator of whether the customer is of type 1 or 2. For example, in patient triage, the signal can be a criticality score computed based on the patients' condition (similar to the way the CRAMS scale and other existing triage scores are computed), while for call centers, it can be any information about the customers that is believed to be a good indicator of the customer's value to the firm (e.g., total purchases within the last year).

Specifically, we assume that signals from type i customers are i.i.d. with a cumulative distribution function $F_i(\cdot)$ that has a finite mean and a differentiable probability density function $f_i(\cdot)$. Distributions $F_1(\cdot)$ and $F_2(\cdot)$ have the same support $\mathcal{S} = [c, d]$ (or $[c, d)$ if $d = \infty$), where $c < d$. Therefore, given the signal of a customer, the service provider does not know with certainty whether the customer is type 1 or 2.

Suppose for now that the service provider knows the distributions F_1 and F_2. Then, given the customer signal, the service provider can actually compute the probability that the customer is of type 1. Let $p_i(x)$ denote the probability that a customer belongs to type i given that the customer's signal is x. Then, we have

$$p_i(x) = \frac{p_i f_i(x)}{p_1 f_1(x) + p_2 f_2(x)} \quad \text{for } i = 1, 2, \qquad (12)$$

where p_i is the unconditional probability that a random customer belongs to type i. So, if the signal distributions are known, the service provider can simply transform the existing signal x via (12), obtain $p_1(x)$, and then use the policies prescribed in the previous sections.

Now, what if the service provider does not have complete information on the signal distributions F_1 and F_2? What if the service provider only knows (or at least conjectures) that there is some stochastic relationship between the signals of type 1 customers and signals of type 2 customers? For example, it is known that past expenditures of a customer at a firm might be a good predictor of the customer's future value (see, e.g., Reinartz and Kumar 2003). Specifically, suppose that there is a firm, which has a reason to believe that its valuable customers of the future spend more than the other customers in the near past in some stochastic sense, and the firm wants to use the past spending amounts of the customers (e.g., the amounts spent in the last three months) to determine whether each customer should receive a premium service or not. The firm may not be able to compute the probability that each customer is of type 1 (given his or her spending amount within the last three months)

because that requires the complete distribution of the customer expenditure for each type. But, as we discuss below, even without such detailed information, the firm can still use past expenditures to assign priorities as long as certain stochastic ordering relations hold.

First, suppose that F_1 is larger than F_2 in likelihood ratio ordering. (See Appendix A online for definitions of stochastic orders that are used in this section.) The likelihood ratio ordering is a strong stochastic order that implies hazard rate ordering, which, in turn, implies usual stochastic ordering (see, e.g., Müller and Stoyan 2002). In the following proposition, we show that higher signals imply higher probabilities of being type 1 under the assumption of likelihood ratio orders.

Proposition 6. *$p_1(x)$ is nondecreasing in x if and only if signals from type* 1 *customers are larger than signals from type* 2 *customers in likelihood ratio ordering; i.e., $F_1 \geq_{lr} F_2$.*

Proposition 6 implies that when there is likelihood ratio ordering between the signal distributions, ordering customers according to their signals is equivalent to ordering them according to their probabilities of being type 1. Therefore, results of §4 continue to hold for this alternative signal formulation. For example, the HSF policy, i.e., the policy that orders customers according to their signals, still outperforms all finite class priority policies.

Proposition 6 also implies that if there is no likelihood ratio ordering between signals but a weaker stochastic order, such as hazard rate ordering, then ordering customers according to their signals is not necessarily equivalent to ordering them according to their probabilities of being type 1. Hence, in that case, the HSF policy (where signals are as defined in this section) is no longer guaranteed to be better than all finite class priority policies. However, the fact that the HSF policy is no longer better than other policies does not necessarily mean that the service provider should avoid using it. If the firm does not know which policy outperforms HSF (as it is the case in the absence of complete information on F_1 and F_2), it can still be perfectly acceptable as long as it at least improves on FCFS. To see whether that would be the case, we first obtain expressions for C_{HSF} and C_{FCFS}, which denote the long-run average costs under the HSF and FCFS policies, respectively. If customers are served in an FCFS fashion, the resulting queueing system is a standard *M/G/1* queue, for which the expected waiting time W_{FCFS} is known to be

$$W_{\text{FCFS}} = \frac{\lambda(p_1 e_1 + p_2 e_2)}{2(1-\rho)}. \tag{13}$$

Then, it follows that

$$C_{\text{FCFS}} = \lambda(p_1 h_1 + p_2 h_2) W_{\text{FCFS}}. \tag{14}$$

For the HSF policy, on the other hand, we obtain

$$C_{\text{HSF}} = \lambda p_1 h_1 \int_c^d W(x) f_1(x)\,dx + \lambda p_2 h_2 \int_c^d W(x) f_2(x)\,dx, \tag{15}$$

where $W(x)$ denotes the steady-state expected queueing time of a customer with signal x under the HSF rule and as in the proof of Proposition 4, we can show that (see Appendix B online for the derivation)

$$W(x) = \frac{\lambda(p_1 e_1 + p_2 e_2)}{2(1-\rho+\lambda p_1 a_1 F_1(x) + \lambda p_2 a_2 F_2(x))^2}. \tag{16}$$

Now that we have expressions for the long-run average cost of HSF and FCFS policies, we can compare these two costs to determine conditions under which the service provider should give priorities to the customers with higher signals rather than employing the FCFS discipline. It turns out that the usual stochastic ordering between signal distributions is, in fact, sufficient.

Proposition 7. *Suppose that signals from type* 1 *customers are stochastically larger than signals from type* 2 *customers; i.e., $F_1 \geq_{\text{st}} F_2$. Then, the long-run average cost under a policy that gives higher priority to customers with higher signals is at most the same as the long-run average cost under the FCFS policy; i.e., $C_{\text{HSF}} \leq C_{\text{FCFS}}$.*

In a way, Proposition 7 describes what would be an acceptable signal to use when determining priorities. The service provider might use different portions of the information about the customers or process the information differently and come up with alternative ways of obtaining customer signals that can be used in determining priorities. Proposition 7 suggests that no matter what piece of information is used

or how the signal is obtained from this information, in the end, if signals coming from type 1 customers dominate signals coming from type 2 customers in the sense of usual stochastic orders, then this will ensure that HSF outperforms FCFS. Note that (16) can be generalized to the case where there are more than two customer types. Also, Proposition 7 continues to hold, as long as customer types $1, 2, \ldots, M$ can be numbered such that $F_1 \geq_{\text{st}} F_2 \geq_{\text{st}} \cdots \geq_{\text{st}} F_M$ and $h_1/a_1 \geq h_2/a_2 \geq \cdots \geq h_M/a_M$, where M denotes the number of customer types.

If the service provider prefers using a finite class priority policy with $N \geq 2$ classes (as in §5), then it turns out that under the usual stochastic ordering condition, it is also easy to determine a policy that performs better than FCFS.

Proposition 8. *Consider an N-class priority policy π_T under which the signal support $[c, d]$ is divided into N nonoverlapping and exhaustive intervals by a set of $N-1$ thresholds $T = \{t_1, t_2, \ldots, t_{N-1}\}$ such that $d = t_0 > t_1 > t_2 > \cdots > t_{N-1} > c = t_N$, and customers whose signals fall into interval $[t_i, t_{i-1}]$ are assigned to priority class i, where class i jobs have higher priority than class j jobs for all $j > i$. Let Θ_N be the class of all such policies for $N \geq 2$ and let C_{π_T} be the long-run average cost under policy $\pi_T \in \Theta_N$. If $F_1 \geq_{\text{st}} F_2$, then $C_{\pi_T} \leq C_{\text{FCFS}}$ for any policy $\pi_T \in \Theta_N$.*

According to Proposition 8, if there is usual stochastic ordering between signal distributions, the service provider simply needs to pick $N-1$ threshold values, which will determine the N signal intervals for N priority classes, and give priority to customers whose signals fall into intervals with higher signals. No matter how these threshold values are picked, this policy improves on FCFS.

9. Nonlinear Waiting Costs

So far, we assumed that customers' waiting costs are linear in time. However, approximating customers' delay sensitivities with a linear function may not always be reasonable, and thus it is of interest to investigate whether or not (and how) our main findings change if customers experience nonlinear waiting costs. When customers incur nonlinear waiting costs, the analysis becomes significantly more difficult. For finite class priority policies, expressions for any moment of steady-state waiting times are known (see, e.g., Lu and Squillante 2004), and thus it is possible to come up with an expression for the long-run average cost if the waiting cost function is a polynomial. However, even if the cost function is simply $h_i z^2$ for customer type i (where h_i is a constant and z is the time spent in the queue) and we are interested in two-class priority policies, the long-run average cost is not a unimodal function of the threshold and it is not possible to obtain an expression for the optimal threshold. More research is needed to develop a better understanding of how the optimal threshold changes with various system parameters, but it is not difficult to come up with examples where the optimal threshold is not monotone in the traffic load ρ, and thus Proposition 2 does not generalize to systems where customers' waiting costs are nonlinear. Under HSF, the analysis is much more difficult mainly because it appears to be a significant challenge to derive expressions for higher moments of customers' steady-state waiting times. Nevertheless, simulation is viable, and therefore we carried out a simulation study to get insights into the performances of various policies when customers incur nonlinear waiting costs.

In our simulation study, we assumed that waiting cost for a type i customer with a waiting time of z is given by $H_i(z) = h_i z^2$, where h_i is a constant and z is the time the customer spends in the queue. We studied a total of 18 scenarios generated by all combinations of $\lambda \in \{0.3, 0.7, 0.9\}$, $p_1 \in \{0.1, 0.4, 0.7\}$, and $h_1 \in \{4, 50\}$. We set $h_2 = 1$ and $a_1 = a_2 = 1$, so that the priority relations between the two types are clear (i.e., if type identities of customers were available, among two customers of each type who have been in the system for the same amount of time, the service provider would choose to serve the type 1 customer earlier). Arrivals were assumed to be Poisson and service times were assumed to be exponentially distributed. Finally, we assumed that the signal distribution is uniformly distributed over $[p_1 - 0.1, p_1 + 0.1]$.

We investigated the performances of four different policies: the FCFS policy, the HSF policy, the optimal two-class priority policy,[3] and a new policy that we call the *generalized expected $c\mu$* (GE-$c\mu$) policy and

[3] The optimal threshold under each scenario is obtained numerically.

is an extension of the *generalized* $c\mu$ (G-$c\mu$) rule first studied and proposed by Van Mieghem (1995). The G-$c\mu$ rule assumes that customers' type identities are observable and whenever the server completes a service, it computes an index $H'_{i(n)}(z_n)/a_{i(n)}$ for each customer n, where $H'_i(\cdot)$ is the first derivative of $H_i(\cdot)$, $i(n)$ is the type of customer n, and z_n is the time that the customer spent in the system so far, and picks the customer with the largest index value to service next. Van Mieghem (1995) proves that G-$c\mu$ rule is asymptotically optimal for large traffic intensities and under nondecreasing convex waiting cost functions. The GE-$c\mu$ policy carries the same idea to our setting. More specifically, the policy works exactly the same as the G-$c\mu$ rule except that the index for customer n is computed by $x_n H'_1(z_n)/a_1 + (1-x_n)H'_2(z_n)/a_2$, where x_n is the signal of customer n.

We used Arena 10.0 simulation software and constructed 95% confidence intervals on the long-run average waiting cost under each policy and scenario. To obtain these confidence intervals, we used the batch means output analysis method with 30 batches, each having 32,000, 80,000, and 180,000 customers per batch for the scenarios with $\lambda = 0.3$, $\lambda = 0.7$, and $\lambda = 0.9$, respectively. We have also deleted 100,000 initial observations from the runs with $\lambda \in \{0.3, 0.7\}$ and 200,000 initial observations from the runs with $\lambda = 0.9$ based on a warm-up period analysis. Our results are presented in Tables 1 and 2. When comparing confidence intervals for two policies in a row, if the intervals overlapped, we conducted a paired-t test and confirmed that there is a statistical difference between any such policies at a significance level of 95% and the difference is in favor of the policy with the smaller mean performance. Note, however, that for the scenarios where we report the exact same confidence interval for the FCFS and optimal two-class policies, the two policies are exactly the same because the optimal threshold for the two-class policy turns out to be zero.

From Tables 1 and 2, we observe that HSF is no longer the best policy. In fact, its performance is the worst in most cases with significant margins particularly when system load is high. This poor performance of HSF is not surprising. The convex waiting cost function considered for this simulation study punishes long customer waits severely. Under the HSF policy, however, especially when the system load

Table 1 95% Confidence Intervals on the Long-Run Average Holding Costs When $h_1 = 4$

λ	p_1	GE-$c\mu$	HSF	Two-class priority	FCFS
0.3	0.1	0.483 ± 0.012	0.567 ± 0.015	0.501 ± 0.012	0.501 ± 0.012
0.3	0.4	0.825 ± 0.021	0.987 ± 0.027	0.828 ± 0.021	0.828 ± 0.021
0.3	0.7	1.158 ± 0.027	1.407 ± 0.036	1.161 ± 0.027	1.161 ± 0.027
0.7	0.1	14.21 ± 0.336	29.61 ± 1.225	14.42 ± 0.343	14.42 ± 0.343
0.7	0.4	24.29 ± 0.574	54.11 ± 2.163	24.43 ± 0.574	24.43 ± 0.574
0.7	0.7	34.37 ± 0.805	78.40 ± 3.227	34.51 ± 0.805	34.51 ± 0.805
0.9	0.1	204.3 ± 8.712	$1{,}098 \pm 72.18$	207.9 ± 8.811	207.9 ± 8.811
0.9	0.4	350.1 ± 14.94	$2{,}052 \pm 131.4$	351.9 ± 14.94	351.9 ± 14.94
0.9	0.7	495.0 ± 21.06	$3{,}015 \pm 195.3$	496.8 ± 21.15	496.8 ± 21.15

Table 2 95% Confidence Intervals on the Long-Run Average Holding Costs When $h_1 = 50$

λ	p_1	GE-$c\mu$	HSF	Two-class priority	FCFS
0.3	0.1	1.992 ± 0.078	2.136 ± 0.087	2.055 ± 0.078	2.214 ± 0.078
0.3	0.4	7.710 ± 0.198	9.000 ± 0.288	8.040 ± 0.207	7.770 ± 0.204
0.3	0.7	13.17 ± 0.312	15.84 ± 0.429	13.23 ± 0.315	13.23 ± 0.315
0.7	0.1	49.91 ± 1.204	67.97 ± 2.653	59.43 ± 2.247	65.38 ± 1.673
0.7	0.4	224.7 ± 5.236	469.0 ± 18.06	228.9 ± 5.341	228.9 ± 5.341
0.7	0.7	390.6 ± 9.030	868.0 ± 35.28	392.7 ± 9.100	392.7 ± 9.100
0.9	0.1	666.9 ± 29.16	$1{,}584 \pm 89.82$	948.6 ± 40.32	945.0 ± 39.33
0.9	0.4	$3{,}231 \pm 139.5$	$17{,}190 \pm 1{,}089$	$3{,}294 \pm 140.4$	$3{,}294 \pm 140.4$
0.9	0.7	$5{,}616 \pm 239.4$	$32{,}850 \pm 2{,}106$	$5{,}652 \pm 240.3$	$5{,}652 \pm 240.3$

is high, customers with low signals end up waiting for a long time. The average waiting time across all customers under HSF is the same as that under the other policies but its variance (or its second moment) is higher. In comparison, under FCFS and optimal two-class policies, customers experience much more homogeneous waiting times. In the two-class policy, customers in class 2 will experience longer waiting times but because the customers are ordered in a FCFS fashion within each class, no customer experiences as long waits as some under HSF. Therefore, optimal two-class and FCFS policies perform much better than HSF in almost all scenarios. Note that the optimal two-class policy and FCFS policy perform very similarly (or exactly the same), except when λ and p_1 are small.

The GE-$c\mu$ policy is the best policy in all the scenarios considered. This policy helps avoid long customer waits because customers are increasingly more likely to be picked up by the server as they keep waiting. Among the four policies we tested, GE-$c\mu$ is the only policy that takes into account both the times that the customers have already spent in the system and their signals, and therefore its good performance is not surprising. It is, in fact, interesting that the performances of the optimal two-class and FCFS policies are quite close to that of GE-$c\mu$ under many scenarios. However, GE-$c\mu$ performs significantly better than the optimal two-class and FCFS policies when there is a significant difference between the waiting costs of types 1 and 2 customers (h_1 and h_2), a fraction of type 1 customers p_1 is small, and system load ($\rho = \lambda/\mu$) is high. Note that this is consistent with our numerical analysis for the linear cost case discussed in §6, where we found that priority policies are most beneficial when system load is high and a fraction of type 1 customers is small.

10. Conclusions

Priority assignment decisions under imperfect information on customer type identities have received almost no attention in the literature. In particular, to the best of our knowledge, this is the first paper that explicitly considers imperfect customer information and priority assignment decisions within the same model. The paper provides several insights on what kind of information can be used to classify customers and how exactly that information should be used, how the "optimal" classification policies depend on system characteristics, and what kind of information would be more useful than others.

In our formulation, each customer belongs to one of two types. Type identities of the customers are not available to the service provider but each arriving customer provides a *signal*, a numerical score that is an imperfect indicator of the customer's type. In most of the paper, we assume that customers incur waiting costs that are linear in time and the signal coming from a customer equals the probability that the customer is of type 1, which is the type that should have the higher priority. We find that increasing the number of priority classes decreases the long-run average waiting cost for the system and the HSF policy that gives priority to the customer with the highest signal outperforms any finite class priority policy.

Our analysis of two-class policies helped us generate insights on the structural properties of the optimal policies and the comparison of different signals. In particular, we find that if there are two different signals available, the long-run average cost is smaller if the service provider uses the one that is larger in convex ordering. This result suggests that signals with more spread-out distributions are more beneficial. Our numerical analysis suggests that such signals are more preferable under the HSF policy as well.

When signals are less informative in the sense that they do not reveal the type probabilities but the service provider knows (or at least strongly believes) that there is some stochastic ordering relation between the signals coming from the two different types of customers, we find that the HSF policy as well as any finite class priority policy in which higher priority customers have higher signals outperform the standard FCFS policy if type 1 customers' signals are larger than those of type 2 customers in the usual stochastic sense.

Our numerical and simulation analysis provided insights on the performances of different policies not only when waiting costs are linear in time but also when they are convex. Even though HSF outperforms any finite class policy when costs are linear, its performance is surprisingly close to that of the optimal two-class policy. Furthermore, when the waiting cost function is convex, HSF performs very poorly. The optimal two-class policy and FCFS both perform

much better than HSF. This suggests that optimal two-class policies are more robust than HSF with respect to customers' varying delay sensitivities. However, the best choice appears to be the GE-$c\mu$ policy, which is an extension of the generalized $c\mu$ rule of Van Mieghem (1995). When the waiting cost is a convex and quadratic function of time, this policy performs clearly better than all other alternatives. On the other hand, when waiting costs are linear in time, the policy reduces to the HSF policy, which performs better than all finite-class policies. One interesting avenue for future work would be to investigate whether this good performance of the policy could be formalized, perhaps by establishing its asymptotic optimality in the heavy-traffic regime as Van Mieghem did for the generalized $c\mu$ rule when type identities are available. It is also of interest to investigate the performance of the GE-$c\mu$ policy under cost functions that are not convex.

One crucial assumption in our model is that there is a single server. In general, single-server systems can provide useful insights for multiserver systems, which behave like single-server queues under heavy loads. Nevertheless, the analysis of multiple-server systems is of interest at least to see whether or not and how the insights we obtained from the single-server model would change. Unfortunately, however, analysis of multiserver systems is significantly more challenging mainly because of the lack of closed-form expressions for the steady-state expected waiting times for different priority classes. To our knowledge, the only exception is the system with Poisson arrivals and exponential service times, where customers' service times do not depend on their types. In this case, we can show that all of our analytical results hold for the multiserver system operating under these conditions, see Appendix D online.

Future work might concentrate on several ways that our model and results can be extended. One possibility is to investigate a model where customers renege from the system. It would be also of interest to consider game-theoretic models, where customers have the ability to influence their signals and can act strategically to maximize their own objectives. In another direction, future work might investigate how to pick among a number of service improvement alternatives. The service provider can take a variety of actions to improve the quality of service provided to the customers. For example, she can develop policies that make better use of the available signals, she can identify and/or develop alternative customer signals that are more informative, or she can expand the service capacity. Our analysis, in this paper centered around the first two options, assuming that the service capacity is fixed. Future work might consider more complicated decisions, particularly those that consider various service improvement opportunities simultaneously.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

Acknowledgments

The authors thank the associate editor and two anonymous referees for their comments that significantly improved the paper. The work of the first author was supported by the National Science Foundation (NSF) under Grant CMMI-0715020. The work of the second author was supported by the NSF under Grant CMMI-0620737.

References

Afèche, P. 2007. Incentive-compatible revenue management in queueing systems: Optimal strategic delay and other delay tactics. Working paper, Rotman School of Management, University of Toronto, Toronto.

Akşin, O. Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

Balachandran, K. R. 1972. Purchasing priorities in queues. *Management Sci.* **18**(5) 319–326.

Baxt, W. G., C. C. Berry, M. D. Epperson, V. Scalzitti. 1989. The failure of prehospital trauma prediction rules to classify trauma patients accurately. *Ann. Emergency Medicine* **18**(1) 1–8.

Brady, D. 2000. Why service stinks. *BusinessWeek* (October 23) 118–128.

Cobham, A. 1954. Priority assignment in waiting line problems. *J. Oper. Res. Soc. America* **2**(1) 70–76.

Cobham, A. 1955. Priority assignment—A correction. *J. Oper. Res. Soc. America* **3**(4) 547.

Cox, D. R., W. L. Smith. 1961. *Queues*. Methuen & Co., London.

Frykberg, E. R. 2002. Medical management of disasters and mass casualties from terrorist bombings: How can we cope? *J. Trauma* **53**(2) 201–212.

Gans, N., Y.-P. Zhou. 2003. A call-routing problem with service level constraints. *Oper. Res.* **51**(2) 255–271.

Gans, N., Y.-P. Zhou. 2007. Call-routing schemes for call-center outsourcing. *Manufacturing Service Oper. Management* **9**(1) 33–50.

Green, L. 1984. A multiple dispatch queueing model of police patrol operations. *Management Sci.* **30**(6) 653–664.

Güneş, E. D., O. Z. Akşin. 2004. Value creation in service delivery: Relating market segmentation, incentives, and operational performance. *Manufacturing Service Oper. Management* **6**(4) 338–357.

Gurvich, I., M. Armony, A. Mandelbaum. 2005. Service level differentiation in call centers with fully flexible servers. *Management Sci.* **54**(2) 279–294.

Hasija, S., E. J. Pinker, R. A. Shumsky. 2005. Staffing and routing in a two-tier call center. *Internat. J. Oper. Res.* **1**(1/2) 8–29.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue*: *Equilibrium Behavior in Queueing Systems*. Kluwer, Boston.

Hastie, T., R. Tibshirani, J. Friedman. 2001. *The Elements of Statistical Learning*: *Data Mining, Inference, and Prediction*. Springer, New York.

Hosmer, D. W., S. Lemeshow. 2000. *Applied Logistic Regression*. John Wiley & Sons, New York.

Jaiswal, N. K. 1968. *Priority Queues*. Academic Press, New York.

Kleinrock, L. 1967. Optimum bribing for queue position. *Oper. Res.* **15**(2) 304–318.

Kumar, V. 2008. *Managing Customers for Profit*: *Strategies to Increase Profits and Build Loyalty*. Wharton School Publishing, Upper Saddle River, NJ.

Lu, Y., M. S. Squillante. 2004. Scheduling to minimize general functions of the mean and variance of sojourn times in queueing systems. IBM Research Report RC23415 (W0411-053), Yorktown Heights, NY.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38**(5) 870–883.

Müller, A., D. Stoyan. 2002. *Comparison Methods for Stochastic Models and Risks*. John Wiley & Sons, West Sussex, UK.

Palumbo, L., J. Kubincanek, C. Emerman, N. Jouriles, R. Cydulka, B. Shade. 1996. Performance of a system to determine EMS dispatch priorities. *Amer. J. Emergency Medicine* **14**(4) 388–390.

Rao, S., E. R. Petersen. 1998. Optimal pricing of priority services. *Oper. Res.* **46**(1) 46–56.

Reilly, M. J. 2006. Accuracy of a priority medical dispatch system in dispatching cardiac emergencies in a suburban community. *Prehospital Disaster Medicine* **21**(March–April) 77–81.

Reinartz, W. J., V. Kumar. 2003. The impact of customer relationship characteristics on profitable lifetime duration. *J. Marketing* **67**(1) 77–99.

Schaack, C., R. C. Larson. 1986. An *N*-server cutoff priority queue. *Oper. Res.* **34**(2) 257–266.

Schaack, C., R. C. Larson. 1989. An *N* server cutoff priority queue where arriving customers request a random number of servers. *Management Sci.* **35**(5) 614–634.

Shaked, M., J. G. Shanthikumar. 2007. *Stochastic Orders*. Springer, New York.

Shin, A. 2006. What customers say and how they say it. *Washington Post* (October 18) DO1.

Shumsky, R. A., E. J. Pinker. 2003. Gatekeepers and referrals in services. *Management Sci.* **49**(7) 839–856.

van der Zee, S. P., H. Theil. 1961. Priority assignment in waiting-line problems under conditions of misclassification. *Oper. Res.* **9**(6) 875–885.

Van Mieghem, J. 1995. Dynamic scheduling with convex delay costs. *Ann. Appl. Probab.* **5**(3) 809–833.

Walton, C. J., B. F. S. Grenyer. 2002. Prioritizing access to psychotherapy services: The client priority rating scale. *Clinical Psych. Psychotherapy* **9**(6) 418–429.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Upper Saddle River, NJ.