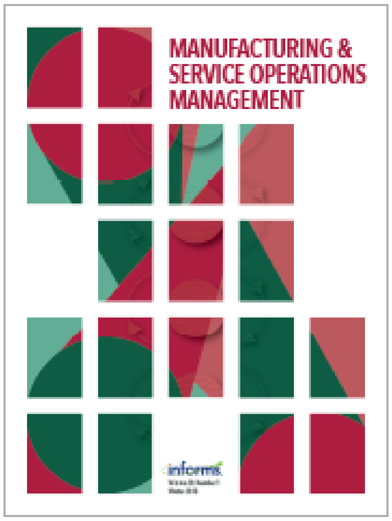


Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Appointment Scheduling Under a Service-Level Constraint

Saif Benjaafar, David Chen, Rowan Wang, Zhenzhen Yan

https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 25, No. 1, January–February 2023, pp. 70–87
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Appointment Scheduling Under a Service-Level Constraint

Saif Benjaafar,[a] David Chen,[b,*] Rowan Wang,[c,*] Zhenzhen Yan[d]

[a] Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota 55455; [b] School of Management and Economics, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China; [c] Department of Information Systems and Management Engineering, Southern University of Science and Technology, Shenzhen, Shenzhen 518055, China; [d] School of Physical and Mathematical Sciences, Nanyang Technological University, 639798 Singapore, Singapore

*Corresponding authors

Contact: saif@umn.edu, https://orcid.org/0000-0002-1949-0105 (SB); davidchen@cuhk.edu.cn, https://orcid.org/0000-0001-6971-6654 (DC); wangy2021@sustech.edu.cn, https://orcid.org/0000-0002-7665-4460 (RW); yanzz@ntu.edu.sg, https://orcid.org/0000-0001-7370-9959 (ZY)

Received: November 6, 2021
Revised: July 9, 2022; September 9, 2022
Accepted: September 22, 2022
Published Online in Articles in Advance: November 28, 2022

https://doi.org/10.1287/msom.2022.1159

Abstract. ***Problem definition*:** This paper studies an appointment system where a finite number of customers are scheduled to arrive in such a way that (1) the expected waiting time of each individual customer cannot exceed a given threshold; and (2) the appointment times are set as early as possible (without breaking the waiting time constraint). ***Methodology/results*:** First, we show that, under the service-level constraint, a prospective schedule can be obtained from a sequential scheduling approach. In particular, we can schedule the appointment time of the next customer based on the scheduled appointment times of the previous customers. Then, we use a transient queueing-analysis approach and apply the theory of majorization to analytically characterize the structure of the optimal appointment schedule. We prove that, to keep the expected waiting time of each customer below a certain threshold, the minimum inter-appointment time required increases with the arrival sequence. We further identify additional properties of the optimal schedule. For example, a later arrival has a higher chance of finding an empty system and is more likely to wait less than the duration of his expected service time. We show the convergence of the service-level-constrained system to the D/M/1 queueing system as the number of arrivals approaches infinity and propose a simple, yet practical, heuristic schedule that is asymptotically optimal. We also develop algorithms that can help system managers determine the number of customers that can be scheduled in a fixed time window. We compare the service-level-constrained appointment system with other widely studied systems (including the equal-space and cost-minimization systems). We show that the service-level-constrained system leads to a lower upper bound on each customer's waiting time; ensures a fair waiting experience among customers; and performs quite well in terms of system overtime. Finally, we investigate various extended settings of our analysis, including customer no-shows; mixed Erlang service times; multiple servers; and probability-based service-level constraints. ***Managerial implications*:** Our results provide guidelines on how to design appointment schedules with individual service-level constraints. Such a design ensures fairness and incorporates the threshold-type waiting perception of customers. It is also free from cost estimation and can be easily applied in practice. In addition, under the service-level-constrained appointment system, customers with later appointment times can have better waiting experiences, in contrast to the situation under other commonly studied systems.

Funding: Z. Yan was partly supported by a Nanyang Technological University startup grant; the Ministry of Education Academic Research Fund Tier 1 [Grant RG17/21] and Tier 2 [Grant MOE2019-T2-1-045]; and Neptune Orient Lines [Fellowship Grant NOL21RP04].

Supplemental Material: The online supplement is available at https://doi.org/10.1287/msom.2022.1159.

Keywords: appointment scheduling • service-level constraint • waiting time • majorization

1. Introduction

Appointment scheduling has been well studied in recent years. Most studies in the existing literature (see, for example, the review papers of Cayirli and Veral 2003 and Gupta and Denton 2008) assume linear waiting cost of customers (i.e., a customer's utility decreases linearly in his waiting time). However, from the survey results in Baron et al. (2017), customers whose waiting time exceeds 20 minutes are substantially less satisfied with their service. This finding indicates the threshold-type waiting perception of customers and brings a new, important, and practical feature to the research on appointment scheduling.

In this paper, we tackle the appointment-scheduling problem from a new angle. We consider an appointment system that schedules a finite number of customers to arrive in such a way that (1) the expected waiting time of each customer cannot exceed a predetermined threshold, and (2) the appointment times are set as early as possible (without breaking the waiting-time constraint). This approach captures the above-mentioned threshold-type waiting perception of customers. Using transient queueing analysis, we analytically characterize the structure and properties of the optimal appointment schedule, as well as its limiting behavior. In the following, we introduce and explain the important features of our work in terms of modeling perspective and analysis method.

1.1. Modeling Perspective

A fundamental issue that arises in appointment scheduling is the trade-off between servers' idling (when appointments are scheduled far from each other) and customers' waiting (when appointments are scheduled close to each other). For decades, appointment scheduling has drawn significant attention from the queueing, optimization, operations management, and healthcare research communities; see Cayirli and Veral (2003), Mondschein and Weintraub (2003), Gupta and Denton (2008), and Keskinocak and Savva (2020) for comprehensive reviews of the literature. As pointed out by Cayirli and Veral (2003), the overwhelming majority of the studies assign unit costs (weights) to servers' idling and customers' waiting and then search for the appointment schedule that minimizes the expected total system cost, which is a linear combination of servers' idling and customers' waiting costs. Mondschein and Weintraub (2003) point out that other objective function forms typically used in the literature (including those with servers' overtime cost) are equivalent to the one described above.

Unlike traditional approaches that minimize the system cost, a few recent papers study appointment systems with service-level targets. Millhiser et al. (2012) and Millhiser and Veral (2015) are among the first to design appointment systems with performance targets. For a given predetermined schedule, Millhiser et al. (2012) use a stochastic model to study the probability that a particular customer will wait more than an established threshold. They find that traditional scheduling systems have serious shortcomings in terms of providing consistent service levels. Millhiser and Veral (2015) use an inverse-simulation approach to design a schedule, under which the probability of waiting longer than a given threshold is uniform across all customers. The authors consider various no-show probabilities and service-time distributions and design heuristic appointment schedules that yield robust system performance. The results and insights obtained in these two papers are mainly based on numerical experiments and simulations.

Along similar lines, in this paper, we study an appointment system with a constraint on the waiting time of each individual customer and focus on analytically characterizing the structure of the optimal schedules. This service-level-based approach has several important features and advantages.

First, it is free from cost estimation. While the majority of existing models minimize the system cost, the resulting schedule depends critically on the relative costs of servers' idling and customers' waiting. Therefore, obtaining accurate cost-parameter values becomes a crucial issue in the application of theoretical results. Although academic literature exists on the estimation of these unobservable costs (see, e.g., Olivares et al. 2008 and Robinson and Chen 2011), practical implementation is still challenging. Fries and Marathe (1981) relate the difficulty in estimating waiting cost to the connection between customers' waiting and the issues of goodwill, as well as the cost to society. Long waiting time can lead to reneging and negative word of mouth, which further complicate the estimation of the cost. In contrast, the service-level-based approach only deals with time. Cost is not involved. Thus, our results and insights can be applied in practice without any cost estimation.

Second, the service-level-based approach incorporates the threshold-type waiting perception of customers. In the more common cost-minimization approaches, customers' waiting cost is typically assumed to be a linear function of waiting time. Yet, from recent empirical studies (see, e.g., Baron et al. 2017), for various service encounters, customers' perception of waiting time reveals a threshold-type behavior. That is, customers are generally satisfied with their waiting experience if they wait no more than a certain time length (e.g., 20 minutes); however, their patience declines rapidly when their waiting time exceeds that threshold. In practice, such a waiting acceptance threshold can usually be obtained from customer-satisfaction surveys (see Baron et al. 2017 for details). Firms often consider this acceptance threshold as their performance target. Compared with the traditional linear cost model, the threshold-type waiting model can better characterize customers' perception of waiting. The service-level-based approach captures this threshold-type feature by bounding an upper limit on each customer's expected waiting time.

Third, the service-level-based approach ensures fairness among customers. It is observed from existing literature that schedules resulting from cost-minimization approaches lead to unequal waiting experiences for individual customers. Hassin and Mendel (2008) provide numerical results showing that for both the equal-space (inter-appointment time stays constant) and the dome-shaped (inter-appointment time initially increases and then decreases) systems, customers who are scheduled to arrive later wait longer than those who are scheduled to arrive earlier. Cayirli and Veral (2003) highlight that the trend of increasing waiting time is observed under most commonly studied appointment systems. Qi (2017) also reports this unfairness issue. Notice that in the service-level-based approach, the service-level constraint sets an upper limit on the expected waiting time of each individual customer. In this sense, the resulting appointment schedule ensures some degree of fairness among customers.

The service-level-based approach raises several important questions. First, how to compute the optimal inter-appointment times under the service-level constraint? Second, what is the structure of the optimal inter-appointment times? Third, do there exist any upper or lower bounds on the optimal inter-appointment times? Last, how are optimal inter-appointment times affected by the value of the waiting threshold?

1.2. Analysis Method

In the past few years, there has been a growing body of appointment-scheduling literature from the optimization community (see, e.g., Ahmadi-Javid et al. 2017 and the references therein). These studies primarily focus on building prospective schedules (where the appointment times of all customers are determined together at once) and deriving tractable computation models to achieve (or approximate) optimality. Few analytical results exist on the structural properties of the optimal schedules.

In this paper, we show that under the service-level constraint, a prospective schedule can be obtained from a sequential scheduling approach. In particular, we can schedule the appointment time of the next customer based on the scheduled appointment times of the previous customers. Such a sequential scheduling procedure is practically relevant and easy to implement. We show that this simple approach is effective and ensures the validity of the service-level constraint.

More importantly, our paper analytically characterizes the structural properties of the optimal schedules. Using the theory of majorization, we prove that to keep the expected waiting time of each customer below a certain threshold, the minimum inter-appointment time required increases with the arrival sequence. That is, the inter-appointment time between customers m and $m+1$ is no less than that between customers $m-1$ and m.

We further identify additional properties of the optimal schedule. For example, a later arrival has a higher chance to find an empty system and is more likely to wait less than the duration of his expected service time. We show the convergence of the service-level-constrained appointment system to the D/M/1 queueing system as the number of arrivals approaches infinity and propose a simple, yet practical, heuristic schedule that is asymptotically optimal. We also develop algorithms that can help system managers determine the number of customers that can be scheduled into a fixed time window. We compare the service-level-constrained appointment system with other widely studied systems (including the equal-space and cost-minimization systems). We show that the service-level-constrained system leads to a lower upper bound on each customer's waiting time; ensures fair waiting experience among customers; and generates shorter system overtime under the overloaded and the balanced settings. Finally, we extend our analysis to study systems with customer no-shows, mixed Erlang service times, multiple servers, and probability-based service-level constraints.

The rest of the paper is organized as follows. In Section 2, we review related literature. In Section 3, we formally describe the problem. In Section 4, we characterize the structure of the optimal appointment schedule. In Section 5, we provide managerial implications of our model. In Section 6, we extend our analysis to consider more generalized settings. In Section 7, we provide concluding remarks.

2. Literature Review

Appointment systems have been studied extensively over decades. One stream in the literature considers the scheduling of customers to minimize the total system cost, which includes customers' waiting and servers' idling costs. Because of the uncertainty in service times, the total cost of a system is stochastic. With different distributional assumptions on service times, stochastic programming (see, e.g., Denton and Gupta 2003, Robinson and Chen 2003, and Chen and Robinson 2014) and robust optimization (see, e.g., Kong et al. 2013, Mak et al. 2015, and Jiang et al. 2020) tools have been used to study appointments. The optimal appointment schedule has been found to be typically dome-shaped (i.e., the inter-appointment time first increases, then stays constant, and finally decreases with the order of arrivals).

Recently, variants of the appointment-scheduling problem have been considered with different types of uncertainties, such as customer no-shows. Hassin and Mendel (2008) apply queueing analysis to study an appointment system with both random service times and no-show behavior. Under the assumption of exponential service times and constant no-show probability, the authors investigate the effect of no-shows on the optimal schedule. Jiang et al. (2017) provide an integer-linear-program formulation for an appointment-scheduling problem with both random service times and no-shows from the distributionally robust optimization perspective. The authors provide a tractable way to calculate the optimal schedule using the uncertainties' support and first-moment information. Kong et al. (2021) further analyze the problem by incorporating the observations that the no-show probability depends on the appointment time. This observation leads to endogenous uncertainty in the optimization model. The authors tackle the challenge by providing an iterative algorithm based on co-positive programs to approximate the optimal schedule. The effect of the schedule-dependent no-show on the scheduling decision is investigated. Besides the effect of no-shows on schedule, other factors have also been studied in the literature. Jouini et al. (2022) study appointment systems with no-shows and nonpunctual arrivals from a queueing perspective. Luo et al. (2012) look at the appointment-scheduling system under exponentially distributed service times; no-shows with a constant probability; and possibilities of service interruption due to emergency requests. Wang et al. (2020) use a data-analytics approach to study appointment systems with walk-ins and heterogeneous, time-dependent no-shows. The authors develop an optimization model to obtain the optimal appointment schedule in the presence of walk-ins.

In some practices, healthcare appointment systems adopt scheduling templates with fixed-length appointment slots. In other words, instead of considering continuous schedules, appointment times are restricted to discrete time periods. A line of research is to design optimal appointment templates with discrete schedules (see, e.g., Begen and Queyranne 2011 and Kuiper et al. 2015). Zacharias and Pinedo (2014) study the optimal design of a discrete appointment schedule for a single service provider, assuming that customers have homogeneous no-show probabilities and the service duration is deterministic. Zacharias and Pinedo (2017) extend the work to an appointment system with multiple identical servers and prove the discrete convexity properties of the optimization problem. Zacharias and Yunes (2020) further incorporate other features (such as stochastic service times, nonpunctuality, and unscheduled walk-ins) and prove the multimodularity structure of the objective function. The authors point out that appointment overbooking is an effective practice, and they observe a significant front-loading pattern in the optimal schedule.

Another stream of related literature studies appointment systems with service-level targets. Millhiser et al. (2012) and Millhiser and Veral (2015), as we previously reviewed, were among the first to design appointment systems with performance targets. Qi (2017) considers appointment systems with customers who generate threshold-type unhappiness from waiting. The objective is to balance the service levels among customers, using the concept of lexicographic min-max fairness. The author provides a sequence of mixed-integer-programming formulations to obtain optimal scheduling policies. Zhou et al. (2021) consider a joint sequencing and scheduling problem under a service-level constraint, such that the expected waiting time of each customer is less than or equal to a certain value. The authors propose approximation methods to derive bounds on the optimal expected time span for each given sequence and then turn the joint sequencing and scheduling problem into a simple optimization problem. Our paper, in contrast, focuses on analyzing the structure of the optimal schedules. We use a transient-queueing-analysis approach and apply the theory of majorization to analytically derive structural properties of the optimal appointment schedule, as well as to characterize its limiting behavior.

3. Problem Description

Consider a single-server system, where M customers are scheduled to arrive over time. Customers are indexed by their appointment times (i.e., customer m refers to the m^{th} arrival, for $m = 1, \ldots, M$). All customers show up punctually. Let A_m denote the appointment time of customer m, for $m = 1, \ldots, M$. The server becomes available (i.e., the system starts) at time 0, and we have $0 \le A_m \le A_{m+1}$. Customer service times are independently, identically, and exponentially distributed with mean $\frac{1}{\mu}$ (systems with mixed Erlang service times will be discussed in Section 6). When a customer arrives, if the server is idle, then the customer starts his service immediately; otherwise, the customer joins the queue and waits. Customers in the queue are served on a first-come-first-served basis (i.e., the same order of their appointment times).

There exists a service-level constraint on the waiting time of each individual customer. Namely, the expected waiting time of each customer must be less than or equal to a predetermined value s (*waiting threshold*). We would like to discover the optimal appointment schedule that minimizes the expected *makespan* (length of the time interval from when the server starts serving the first customer to when the server completes serving the last customer), subject to the service-level constraint.

Let T_m denote the inter-appointment time between customers $m-1$ and m for $m = 2, \ldots, M$—that is, $T_m = A_m - A_{m-1}$. Let W_m denote the random variable describing the waiting time of customer m, and let w_m denote its expectation (i.e., $w_m = E[W_m]$). Notice that because the server becomes available at time 0 and customers are served on a first-come-first-served basis, the makespan corresponds to the departure time of the last customer (or, more precisely, the length of the time between time 0 and the departure time of the last customer). The departure time of the last customer equals the sum of his arrival time, waiting time, and service time; and, further, the arrival time of the last customer equals the sum of all inter-appointment times between consecutive customers. Let D_M denote the departure time of the last customer (i.e., the makespan). Because the expected service time of the last customer equals $\frac{1}{\mu}$, we have $E[D_M] = \sum_{i=2}^{M} T_m + \, w_M + \frac{1}{\mu}$. Thus, we can formulate the optimal appointment-scheduling problem as follows.

$$\begin{aligned} \min \quad & \sum_{i=2}^{M} T_m + w_M + \frac{1}{\mu} \\ \text{subject to} \quad & w_m \le s, \text{ for } m = 1, \ldots, M. \end{aligned} \tag{1}$$

Let A_m^* denote the appointment time of customer m under the optimal schedule, and let $\mathbf{T}^* = (T_2^*, \ldots, T_M^*)$

denote the inter-appointment times under the optimal schedule. We also define $T_1 = T_1^* = 0$. In the following proposition, we show that T_m^* can be computed sequentially.

Proposition 1. *For $m = 2, \dots, M$, T_m^* is the optimal solution to the following optimization problem:*

$$\min \quad T_m$$

$$\text{subject to} \quad E[W_m \mid T_n = T_n^*, \text{ for } n = 1, \dots, m-1] \le s. \qquad (2)$$

From Proposition 1, we see that under the service-level constraint, the system that minimizes expected makespan is equivalent to the system that provides each customer with the earliest possible appointment time. Thus, to obtain the optimal schedule $\mathbf{T}^* = (T_2^*, \dots, T_M^*)$, we can compute T_m^* sequentially from Problem (2). Let $W_m(x)$ denote the random variable describing the waiting time of customer m, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, \dots, m-1$, and let $w_m(x)$ denote its expectation (i.e., $w_m(x) = E[W_m(x)]$). Then, we have $T_m^* = \arg\min_{w_m(x) \le s} x$. We also define $w_m^* = w_m(T_m^*)$.

Now, from Proposition 1, we see an interesting and important fact that the service-level-based approach can be viewed as both prospective scheduling (where the appointment times of all customers are decided together at once) and sequential scheduling (where the appointment time of each customer is set one after another at the time when the customer requests an appointment). The interpretation of the optimization problem in the prospective scheduling setting is to find the earliest possible appointment times for all customers, such that the service-level constraints are fulfilled; and the interpretation in the sequential scheduling setting is, given that all previous inter-appointment times are minimized while keeping the service-level constraint valid, that we need to find the shortest inter-appointment time for the next customer such that the service-level constraint remains valid. For both settings, the complete appointment schedule is determined before the start of the service operating time window. It is easy to see, but worth mentioning, the fact that if there are two systems—one with m customers and the other with n customers ($m < n$)—and both are under the same service-level constraint, then the optimal appointment times of the first m customers in the two systems coincide. The sequential scheduling approach enables us to develop analytical characterizations on the structure and properties of the optimal appointment schedule. It also allows us to develop simple algorithms to compute the optimal schedule.

Throughout the paper, we use "increase/decrease" to denote "nondecrease/nonincrease"; x^+ to denote $\max\{0, x\}$; and $\lfloor x \rfloor$ to denote the largest integer that is less than or equal to x. By convention, an empty sum equals zero. In Table 1, we provide a glossary of notation used in the paper.

Table 1. Glossary of Notation

Notation	Definition
A_m	Appointment time of customer m
A_m^*	Appointment time of customer m, under optimal schedule
T_m	Inter-appointment time between customers $m-1$ and m
T_m^*	Inter-appointment time between customers $m-1$ and m, under optimal schedule
D_M	Makespan
D_M^*	Makespan, under optimal schedule
W_m	Waiting time of customer m
$W_m(x)$	Waiting time of customer m, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, \dots, m-1$
W_m^*	Waiting time of customer m, under optimal schedule
w_m	$E[W_m]$
$w_m(x)$	$E[W_m(x)]$
w_m^*	$E[W_m^*]$
R_m	Number of customers (phases) found in the system by customer m, upon his arrival
$R_m(x)$	Number of customers (phases) found in the system by customer m, upon his arrival, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, \dots, m-1$
R_m^*	Number of customers (phases) found in the system by customer m, upon his arrival, under optimal schedule
r_m	$E[R_m]$
$r_m(x)$	$E[R_m(x)]$
r_m^*	$E[R_m^*]$
$p_{m,i}(x)$	$\Pr\{R_m(x) = i\}$, probability that customer m finds i customers in the system, upon his arrival, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, \dots, m-1$
$p_{m,i}^*$	$\Pr\{R_m^* = i\}$, probability that customer m finds i customers in the system, upon his arrival, under optimal schedule
$P_{m,n}(x)$	$\sum_{i=n}^{m-1} p_{m,i}(x)$, probability that customer m finds at least n customers in the system, upon his arrival, with $T_m = x$ and $T_l = T_l^*$ for $l = 2, \dots, m-1$
$P_{m,n}^*$	$\sum_{i=n}^{m-1} p_{m,i}^*$, probability that customer m finds at least n customers in the system, upon his arrival, under optimal schedule
$c_i(x)$	$\frac{(\mu x)^i}{i!} e^{-\mu x}$
$c_{m,i}^*$	$\frac{(\mu T_m^*)^i}{i!} e^{-\mu T_m^*}$

4. Structure of the Optimal Schedule

In this section, we analytically characterize the structure of the optimal schedule through discovering the properties of the inter-appointment times $(T_2^*, \dots, T_M^*)$. We first use an embedded Markov-chain approach to sequentially compute T_m^* and then apply the theory of majorization to prove structural properties.

4.1. Preliminary Results

We analyze the properties of T_m^* by presenting a sequence of preliminary results. First, it is easy to see that $A_1^* = 0$, and, thus, $w_1^* = 0$.

Lemma 1. *$w_m(x)$ is decreasing in x; T_m^* is decreasing in s.*

Lemma 1 states that, given a fixed schedule of previous customers, for the next one, the longer the inter-appointment time, the less he is expected to wait. In addition, when the waiting threshold gets higher, the optimal inter-appointment time becomes shorter. These results are intuitive.

Next, if customers $m-1$ and m are scheduled to arrive together, then the expected waiting time of customer m equals the expected waiting time of customer $m-1$ plus the expected service time of customer $m-1$. Therefore, if customers $1,\dots,m$ are scheduled to arrive together at time 0, then the expected waiting time of customer m equals $\frac{m-1}{\mu}$.

Lemma 2. $T^*_m = 0$ *for* $m = 2,\dots,\lfloor \mu s\rfloor + 1$; $w^*_m = s$ *for* $m = \lfloor \mu s\rfloor + 2,\dots,M$.

Lemma 2 says that it is optimal to schedule the first $\lfloor \mu s\rfloor + 1$ customers to arrive at time 0 together. The rest have expected waiting time all equal to s. Notice that for each customer m, we search for the smallest x, such that $w_m(x) \le s$. Because $\frac{(\lfloor \mu s\rfloor+1)-1}{\mu} \le s$ and $\frac{\lfloor \mu s\rfloor+1}{\mu} > s$, it is optimal to schedule customers $1,2,\dots,\lfloor \mu s\rfloor+1$ to arrive at time 0 together. From customer $\lfloor \mu s\rfloor+2$, because $w_m(T_m)$ is decreasing in T_m (Lemma 1), T^*_m is such that $w^*_m = s$.

Note that because service times are exponentially distributed, the expected waiting time of a customer depends on the number of customers found in the system (in the queue or in service), upon his arrival. Let R_m denote the random variable describing the number of customers found in the system by customer m, upon his arrival, and let r_m denote its expectation (i.e., $r_m = E[R_m]$). In addition, let $R_m(x)$ denote the random variable describing the number of customers found in the system by customer m, upon his arrival, with $T_m = x$ and $T_n = T^*_n$ for $n = 2,\dots,m-1$, and let $r_m(x)$ denote its expectation (i.e., $r_m(x) = E[R_m(x)]$). Then, the total number of customers in the system immediately after A_m equals $R_m(x)+1$. Now, for $i = 0,\dots,m-1$ and $m = 1,\dots,M$, let $p_{m,i}(x) = \Pr\{R_m(x) = i\}$ denote the probability that customer m finds, upon his arrival, i customers in the system, with $T_m = x$ and $T_n = T^*_n$ for $n = 2,\dots,m-1$. We also define $R^*_m = R_m(T^*_m)$ and $p^*_{m,i} = \Pr\{R^*_m = i\}$.

When a customer finds i customers in the system upon arrival, his expected waiting time equals $\frac{i}{\mu}$. Therefore, we have $w_m(x) = \sum_{i=1}^{m-1} p_{m,i}(x)\frac{i}{\mu} = r_m(x)\frac{1}{\mu}$ and $w^*_m = \sum_{i=1}^{m-1} p^*_{m,i}\frac{i}{\mu} = r^*_m\frac{1}{\mu}$. The service-level constraint on waiting time ($w^*_m \le s$) can then be transformed to $r^*_m \le \mu s$. That is, the expected number of customers found in the system, upon each arrival, is not greater than μs. From Lemma 2, except for the first $\lfloor \mu s\rfloor + 1$ customers, who are scheduled to arrive together at time 0, the expected number of customers found in the system, upon each arrival, equals μs. Suppose now that customers $1,\dots,m-1$ are scheduled optimally, and r^*_{m-1} equals μs. The goal is to find the shortest inter-appointment time T^*_m such that $r_m(T^*_m)$ also equals μs. Notice that the earliest available appointment time of customer m is the appointment time of customer $m-1$ (i.e., $T_m = 0$). If customers $m-1$ and m arrive together, then $r_m(0) = r^*_{m-1} + 1 = \mu s + 1 > \mu s$. The service-level constraint will be broken. Thus, the appointment time of customer m should be later than that of customers $m-1$.

To further analyze the properties of $(T^*_2,\dots,T^*_M)$, we need to find the relationship between T^*_m and T^*_{m-1}. Conditioning on the number of customers found by customer $m-1$, upon his arrival, we obtain

$$p_{m,i}(x) = \sum_{j=i-1}^{m-2} p^*_{m-1,j}\Pr\{R_m(x) = i \mid R^*_{m-1} = j\},$$

for $i = 1,\dots,m-1$ and

$$p_{m,0}(x) = 1 - \sum_{i=1}^{m-1} p_{m,i}(x).$$

Similarly, $p^*_{m,i} = \sum_{j=i-1}^{m-2} p^*_{m-1,j}\Pr\{R^*_m = i \mid R^*_{m-1} = j\}$ for $i = 1,\dots,m-1$ and $p^*_{m,0} = 1 - \sum_{i=1}^{m-1} p^*_{m,i}$.

For customer m to find i customers given that customer $m-1$ finds j, there must be exactly $j-i+1$ service completions during the time interval (A^*_{m-1}, A_m) with length x. Because service time is exponentially distributed with rate μ, the number of service completions during a time interval with length x is Poisson-distributed with rate μx. Define $c_i(x) = \frac{(\mu x)^i}{i!}e^{-\mu x}$ and $c^*_{m,i} = \frac{(\mu T^*_m)^i}{i!}e^{-\mu T^*_m}$. Then,

$$\Pr\{R_m(x) = i \mid R^*_{m-1} = j\} = c_{j-i+1}(x),$$

and $\Pr\{R^*_m = i \mid R^*_{m-1} = j\} = c^*_{m,j-i+1}$.

Define $P_{m,n}(x) = \sum_{i=n}^{m-1} p_{m,i}(x)$ and $P^*_{m,n} = \sum_{i=n}^{m-1} p^*_{m,i}$ for $n = 0,\dots,m-1$. Then, we have

$$w_m(x) = \sum_{i=1}^{m-1} p_{m,i}(x)\frac{i}{\mu} = \frac{1}{\mu}\sum_{i=1}^{m-1}\sum_{j=i}^{m-1} p_{m,j}(x) = \frac{1}{\mu}\sum_{i=1}^{m-1} P_{m,i}(x),$$

where the second equality comes from a reordering of the terms (e.g., $\sum_{i=1}^{2} i p_{m,i} = p_{m,1} + 2p_{m,2} = (p_{m,1} + p_{m,2}) + 2p_{m,2} = \sum_{j=1}^{2} p_{m,j} + \sum_{j=2}^{2} p_{m,j} = \sum_{i=1}^{2}\sum_{j=i}^{2} p_{m,j}$). In addition, $P_{m,0}(x) = 1$, and

$$P_{m,n}(x) = \sum_{i=n}^{m-1} p_{m,i}(x) = \sum_{i=n}^{m-1}\sum_{j=i-1}^{m-2} p^*_{m-1,j}c_{j-i+1}(x)$$
$$= \sum_{i=n-1}^{m-2} P^*_{m-1,i}c_{i-n+1}(x),$$

for $n = 1,\dots,m-1$. Similarly, $w^*_m = \frac{1}{\mu}\sum_{i=1}^{m-1} P^*_{m,i}$; $P^*_{m,0} = 1$; and $P^*_{m,n} = \sum_{i=n-1}^{m-2} P^*_{m-1,i}c^*_{i-n+1}$ for $n = 1,\dots,m-1$.

The above relationship between $P^*_{m,n}$ and $P^*_{m-1,n}$ allows us to develop an algorithm to sequentially compute the optimal appointment schedule $\mathbf{T}^* = (T^*_2,\dots,T^*_M)$. We provide the details in Algorithm 1.

Algorithm 1 (Computing the Optimal Schedule)

```
Input: number of customers $M$; service rate $\mu$; waiting threshold $s$.
Output: optimal appointment schedule $T^*_m$, for $m = 2, \ldots, M$.
Step 1: for $m = 2, \ldots, \lfloor \mu s \rfloor + 1$, set $T^*_m = 0$;
        let $m = \lfloor \mu s \rfloor + 2$;
        set $P^*_{m-1,j} = 1$ for $j = 0, \ldots, m-2$.
Step 2: solve for $x_m$ that satisfies $\sum_{i=1}^{m-1} \sum_{j=i-1}^{m-2} P^*_{m-1,j} c_{j-i+1}(x_m) = \mu s$;
        set $T^*_m = x_m$.
Step 3: if $m = M$, then
            report output $T^*_m$ for $m = 2, \ldots, M$;
            stop;
        otherwise (i.e., if $m < M$),
            for $n = 0, \ldots, m-1$, compute $P^*_{m,n} = \sum_{i=n-1}^{m-2} P^*_{m-1,i} c_{i-n+1}(T^*_m)$;
            let $m = m + 1$;
            go to Step 2.
```

4.2. Majorization

In the following, we apply the theory of majorization to show the relationship between T^*_m and T^*_{m-1}. We first introduce the concept of majorization. For an n-dimensional vector $\mathbf{x} = (x_1, \ldots, x_n)$, let $(x_{(1)}, \ldots, x_{(n)})$ denote the vector with the same components, but sorted in an increasing order (i.e., $x_{(1)} \leq \cdots \leq x_{(n)}$); and let $(x_{[1]}, \ldots, x_{[n]})$ denote the one in an decreasing order (i.e., $x_{[1]} \geq \cdots \geq x_{[n]}$).

Definition 1. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x}$ is majorized by $\mathbf{y}$ ($\mathbf{y}$ majorizes $\mathbf{x}$), denoted by $\mathbf{x} \prec \mathbf{y}$, if

$$\begin{cases} \sum_{i=1}^{j} x_{[i]} \leq \sum_{i=1}^{j} y_{[i]} & \text{for } j = 1, \ldots, n-1 \\ \sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}, \end{cases}$$

or, equivalently,

$$\begin{cases} \sum_{i=1}^{j} x_{(i)} \geq \sum_{i=1}^{j} y_{(i)} & \text{for } j = 1, \ldots, n-1 \\ \sum_{i=1}^{n} x_{(i)} = \sum_{i=1}^{n} y_{(i)}. \end{cases}$$

An important characterization of majorization is given in terms of some matrices D such that $\mathbf{x} = D\mathbf{y}$.

Definition 2. A square matrix $D = (d_{ij})$ is a doubly stochastic matrix if $d_{ij} \geq 0$ and $\sum_i d_{ij} = \sum_j d_{ij} = 1, \ \forall i, j$.

Lemma 3. *For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the following conditions are equivalent:*

1. $\mathbf{x} \prec \mathbf{y}$;
2. $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ *and* $\sum_{i=1}^n (x_i - z)^+ \leq \sum_{i=1}^n (y_i - z)^+$, $\forall z \in \mathbb{R}$;
3. $\sum_{i=1}^n |x_i - z| \leq \sum_{i=1}^n |y_i - z|$, $\forall z \in \mathbb{R}$; *and*
4. $\mathbf{x} = D\mathbf{y}$ *for some doubly stochastic matrices D.*

The proof of Lemma 3 and more properties of majorization can be found in Müller and Stoyan (2002) and Marshall et al. (2011). Majorization is generally used as a tool to measure the dissimilarity within the components of vectors. It has many applications in economics. Wang and Gupta (2014) apply the theory of majorization to evaluate the effect of demand variability and absentee-rate variability on the performance of an assignment in the nurse-staffing setting. In our paper, we use majorization to analytically characterize the structure of the optimal schedule in appointment systems with a service-level constraint.

4.3. Structure of Optimal Inter-Appointment Times

Using the concept of majorization and Lemma 3, we can examine the relationship between $P^*_{m,i}$ and $P^*_{m-1,i}$.

Proposition 2. *$P^*_{m,0} = P^*_{m-1,0} = 1$ and $(P^*_{m,m-1}, P^*_{m,m-2}, \ldots, P^*_{m,0}) \prec (0, P^*_{m-1,m-2}, \ldots, P^*_{m-1,0})$, for $m = \lfloor \mu s \rfloor + 3, \ldots, M$.*

We note that majorization is generally used to make comparisons between vectors. It cannot be directly used to make comparisons between random variables. Thus, in general, it is different from stochastic orders. One relationship between majorization and stochastic order is that, if we associate each n-dimensional vector $\mathbf{x} = (x_1, \ldots, x_n)$ with a random variable X, which assigns probability $\frac{1}{n}$ to each x_i, then majorization is equivalent to convex order. (The definition and properties of convex order can be found in Müller and Stoyan 2002.) Clearly, this relationship cannot be directly applied to our model. However, we can show that the majorization result in Proposition 2 is the same as convex order between R^*_m and R^*_{m-1}. That is, $(P^*_{m,m-1}, P^*_{m,m-2}, \ldots, P^*_{m,0}) \prec (0, P^*_{m-1,m-2}, \ldots, P^*_{m-1,0})$ is equivalent to R^*_m being larger than R^*_{m-1}. Note that $P^*_{m,i}$ can be viewed as the tailed distribution of R^*_m at value $i-1$. Therefore, here, we have established a relationship between majorization and convex order when the vector consists of tailed distributions.

Based on Proposition 2, we are ready to show the main result regarding the structure of the optimal inter-appointment times.

Theorem 1. $T^*_{m+1} \geq T^*_m$.

From Lemma 2 and Theorem 1, we see that the optimal appointment schedule has the structure that (1) the first $\lfloor \mu s \rfloor + 1$ customers are scheduled to come together at time 0; and (2) from customer $\lfloor \mu s \rfloor + 2$, the inter-appointment time increases.

Proposition 3. $T^*_m \geq \frac{1}{\mu}$ *for* $m = \lfloor \mu s \rfloor + 3, \ldots, M$.

Proposition 3 shows that, except for the first $\lfloor \mu s \rfloor + 2$ customers, the optimal inter-appointment times between the following customers have a lower bound equal to $\frac{1}{\mu}$. Notice that the expected number of service completions between two consecutive appointment times equals one (see Lemma OS-E5 in the online supplement). For a system with exponential service rate μ, during a time interval of length $\frac{1}{\mu}$, if the server is always busy, then the expected number of service completions equals one. However, if the server is not always busy, then the expected number of service completions during the time interval of length $\frac{1}{\mu}$ is less than one. Thus, we need a time interval longer than $\frac{1}{\mu}$ to have one service completion in expectation.

Next, we explore some other properties of the optimal schedule. Recall that W^*_m is the random variable describing the waiting time of customer m under the optimal schedule. We have the following important results comparing W^*_m and W^*_{m+1}.

Theorem 2. *For $m = \lfloor \mu s \rfloor + 2, \ldots, M$, the following results hold*:

a. $p^*_{m+1,0} \geq p^*_{m,0}$; *and*

b. $\Pr\left\{W^*_{m+1} \leq \frac{1}{\mu}\right\} \geq \Pr\left\{W^*_m \leq \frac{1}{\mu}\right\}$.

Theorem 2 states that, upon arrival, although seeing an equal number of customers in expectation, a later arrival has a higher chance of finding an empty system and is more likely to wait less than the duration of his expected service time. These results make intuitive sense by noticing that a later arrival has a higher chance of finding a more congested system (e.g., customer 10 could find nine customers in the system, whereas customer 5 could only find four at most). So, to balance the possibilities on two ends and to have an equal average number of customers seen by every arrival, a later arrival should also have a higher chance to find an idler system.

It is worth highlighting here that the fact that $p^*_{m,0}$ increases with m can be viewed as the reason why T^*_m increases with m. As we explained earlier, after Proposition 3, during a time interval with fixed length, the number of service completions depends on the proportion of time while the server is busy (working). When m increases, $p^*_{m,0}$ increases, which means that the proportion of time while the server is busy decreases. Therefore, it takes longer to have one service completion.

The second statement in Theorem 2 implies that the probability of long waiting (defined as the case that a customer's waiting time exceeds his expected service time) decreases in the order of arrivals. In addition, from the first statement, the probability of immediate service (zero waiting) also increases in the order of arrivals. Thus, customers with later appointment times can have better waiting experiences. Notice that in a cost-minimization appointment system, customers with later appointment times generally have worse waiting experiences (see the discussion in the introduction section). This is a key difference between a service-level-constrained appointment system and a cost-minimization system.

4.4. Asymptotic Analysis

To complete the section, we study the limiting behavior of the system. We prove that the system converges to the D/M/1 queueing system as the number of arrivals approaches infinity (the analysis of the D/M/1 queue can be found in Jansson 1966).

First, because $T^*_{m+1} \geq T^*_m$, there are two possibilities. (1) If $\{T_m\}_{m=2,3,\ldots}$ is bounded, then $\lim_{M\to\infty} T^*_M$ exists and is finite. (2) If $\{T_m\}_{m=2,3,\ldots}$ is unbounded, then $\lim_{M\to\infty} T^*_M = \infty$. Let $T^* = \lim_{M\to\infty} T^*_M$ (which can be infinity) and $c^*_i = \lim_{M\to\infty} c^*_{M,i} = \lim_{M\to\infty} c_i(T^*_M) = c_i(\lim_{M\to\infty} T^*_M)$ (the last equality is due to the fact that $c_i(x) = \frac{(\mu x)^i}{i!} e^{-\mu x}$ is a continuous function). Define $p^*_{m,i} = 0$, for $i \geq m$.

Lemma 4. $\lim_{M\to\infty} p^*_{M,i}$ *exists, for* $i = 0, 1, \ldots$.

Recall that $p^*_{M,i} = \sum_{j=i-1}^{M-2} p^*_{M-1,j} c^*_{M,j-i+1}$ for $i = 1, \ldots, M-1$, and $p^*_{M,0} = 1 - \sum_{i=1}^{M-1} p^*_{M,i}$. Because $p^*_{M,i} = 0$ for $i \geq M$, we have

$$p^*_{M,i} = \sum_{j=i-1}^{\infty} p^*_{M-1,j} c^*_{M,j-i+1},$$

for $i = 1, 2, \ldots$, and

$$p^*_{M,0} = 1 - \sum_{i=1}^{\infty} p^*_{M,i}.$$

The existence of $\lim_{M\to\infty} p^*_{M,i}$ allows us to take limits on both sides of the above equations. Define $p^*_i = \lim_{M\to\infty} p^*_{M,i}$. Now, let $M \to \infty$, and we have $p^*_i = \sum_{j=i-1}^{\infty} p^*_j c^*_{j-i+1} = \sum_{j=0}^{\infty} p^*_{j+i-1} c^*_j = \sum_{j=0}^{\infty} p^*_{j+i-1} \frac{(\mu T^*)^j}{j!} e^{-\mu T^*}$ for $i = 1, 2, \ldots$, and $p^*_0 = 1 - \sum_{i=1}^{\infty} p^*_i$. Moreover, for $M \geq \lfloor \mu s \rfloor + 2$, we have $\sum_{i=1}^{\infty} i p^*_{M,i} = \sum_{i=1}^{\infty} P^*_{M,i} = \sum_{i=1}^{m-1} P^*_{M,i} = \mu s$. Again, let $M \to \infty$, and we have $\sum_{i=1}^{\infty} i p^*_i = \mu s$. Altogether, we get

$$\begin{cases} p^*_i = \displaystyle\sum_{j=0}^{\infty} p^*_{j+i-1} \frac{(\mu T^*)^j}{j!} e^{-\mu T^*}, \text{ for } i = 1, 2, \ldots. \\ p^*_0 = 1 - \displaystyle\sum_{i=1}^{\infty} p^*_i \\ \displaystyle\sum_{i=1}^{\infty} i p^*_i = \mu s. \end{cases}$$

T^* and p^*_i can then be obtained by solving the above system of equations.

Theorem 3. $T^* = s\left(1 + \frac{1}{\mu s}\right) \ln\left(1 + \frac{1}{\mu s}\right)$; $p^*_i = \frac{1}{1+\mu s}\left(\frac{\mu s}{1+\mu s}\right)^i$ *for* $i = 0, 1, \ldots$.

Theorem 3 shows that as the number of arrivals approaches infinity, our system converges asymptotically to a D/M/1 queueing system having deterministic interarrival times of length $s\left(1 + \frac{1}{\mu s}\right) \ln\left(1 + \frac{1}{\mu s}\right)$ and exponential service times with rate μ.

From Theorems 1 and 3, we see that, for $m = 1, \ldots, M$, T^*_m has an upper bound that equals $s\left(1 + \frac{1}{\mu s}\right) \ln\left(1 + \frac{1}{\mu s}\right)$. Recall that from Proposition 3, for $m = \lfloor \mu s \rfloor + 3, \ldots, M$, T^*_m has a lower bound that equals $\frac{1}{\mu}$. Hence, we have obtained both upper and lower bounds of T_m^* in explicit forms, for $m = \lfloor \mu s \rfloor + 3, \ldots, M$.

The results from Lemma 2 and Theorems 1 and 3 naturally motivate us to consider a simple, yet practical, heuristic schedule; that is, to schedule the first $\lfloor \mu s \rfloor + 1$ customers to arrive at time 0, and then set a constant inter-appointment time $s\left(1+\frac{1}{\mu s}\right)\ln\left(1+\frac{1}{\mu s}\right)$ for the following customers. It can be shown that this simple schedule is asymptotically optimal.

Proposition 4. *The heuristic schedule of $T_m = 0$ for $m = 2, \ldots, \lfloor \mu s \rfloor + 1$, and $T_m = s\left(1+\frac{1}{\mu s}\right)\ln\left(1+\frac{1}{\mu s}\right)$ for $m = \lfloor \mu s \rfloor + 2, \ldots, M$, is asymptotically optimal i.e., $\left(\frac{D_M^H}{D_M^*} \to 1\right)$ as μ approaches zero or infinity; s approaches zero or infinity; or M approaches infinity, where D_M^* and D_M^H are the makespan under the optimal schedule and that under the heuristic schedule, respectively.*

In addition, when the parameters take median values, the above heuristic schedule is expected to have good performance because our numerical studies show that the optimal inter-appointment time converges quickly to the constant. To check this, we compare D_M^* with D_M^H. Table 2 provides representative results for the relative difference $\left(\Delta D_M = \frac{D_M^H - D_M^*}{D_M^*} \times 100\%\right)$.

As we can see from Table 2, the relative difference is less than 10% under all tested parameter settings. Moreover, it is less than 5% when M is large or when μs is small. This suggests that the heuristic schedule is more effective for systems with a larger number of customers, a higher service rate, or a lower waiting threshold.

5. Managerial Implications

5.1. Scheduling Under a Fixed Time Window

In the analysis up to now, we are constructing the appointment system that minimizes makespan, given the number of customers to be scheduled and the service-level constraint. In practice, system managers may also face the problems of (1) deciding the total number of customers to be scheduled into a fixed time window (for example, eight hours), given a waiting threshold; or (2) determining the waiting threshold, given a fixed time window and the number of customers to be scheduled. In this section, we design algorithms to solve these two problems.

We first construct an algorithm to search for M^* (maximum number of customers to be scheduled), given T (time window) and s (waiting threshold). Because the number of customers is an integer, we can search for its optimal value via enumeration. To improve the computational efficiency, we start from $\left\lfloor \frac{T}{s\left(1+\frac{1}{\mu s}\right)\ln\left(1+\frac{1}{\mu s}\right)} \right\rfloor +$ $\lfloor \mu s \rfloor + 1$. Notice that this number represents the maximum number of customers that can be scheduled into the time window, with the waiting threshold, under our heuristic schedule developed in Section 4.4. Clearly, it serves as a lower bound to the maximum number of customers that can be scheduled under the optimal schedule. Thus, we can start from this number to increase M one by one until the last appointment time exceeds T. We provide the details in Algorithm OS-1 in the online supplement. Using Algorithm OS-1, we can compute the optimal number of customers that can be scheduled into a fixed time window, such that the expected individual waiting time fulfills the service-level constraint.

Now, we construct the algorithm to determine s^* (lowest waiting threshold), given T (time window) and M (number of customers to be scheduled). Notice that given T, there exists a monotonic relationship between M and s (the higher s, the larger M becomes; see Figure OS-1 in the online supplement for an illustration). This motivates us to apply a bisection search. To facilitate the computation, we first develop an algorithm to find lower and upper bounds of s^*, such that the difference between the two bounds is less than or equal to one, and then employ the bisection search. We provide the details in Algorithms OS-2 and OS-3 in the online supplement.

5.2. Comparison with Other Scheduling Systems

In this section, we compare the service-level-constrained (*SL*) appointment system with two other widely studied appointment systems, namely, the equal-space (*ES*) and the cost-minimization (*CM*) systems.

We first compare these systems under given M (number of customers to be scheduled) and T (time window). For the SL system, we can apply Algorithms OS-2 and OS-3 in the online supplement to obtain the optimal waiting threshold. This threshold provides an upper bound on each customer's expected waiting time. The ES system is the appointment system, where arrivals are equally spaced with constant inter-appointment times.

Table 2. Performance of the Heuristic Schedule

	s = 0.5 (%)			s = 1.0 (%)			s = 1.5 (%)		
	$\mu = 0.5$	$\mu = 1.0$	$\mu = 1.5$	$\mu = 0.5$	$\mu = 1.0$	$\mu = 1.5$	$\mu = 0.5$	$\mu = 1.0$	$\mu = 1.5$
M = 15	2.8	5.9	9.2	5.9	4.1	9.8	9.0	9.8	9.4
M = 20	2.0	4.3	6.5	4.3	3.1	7.2	6.5	7.2	7.2
M = 25	1.6	3.4	5.1	3.4	2.4	5.7	5.1	5.7	5.8
M = 30	1.3	2.8	4.2	2.8	2.0	4.7	4.2	4.7	4.9

Clearly, $A_1^{ES} = 0$ and $T_m^{ES} = \frac{T}{M-1}$, for $m = 2, \dots, M$. The CM system comes from the solution of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & E[Z(x,S)] \\ \text{subject to} \quad & \sum_{i=1}^{M+1} x_i = T, \\ & x_1 = 0, \\ & x_i \geq 0, \ \forall i = 2, \dots, M+1, \end{aligned} \tag{3}$$

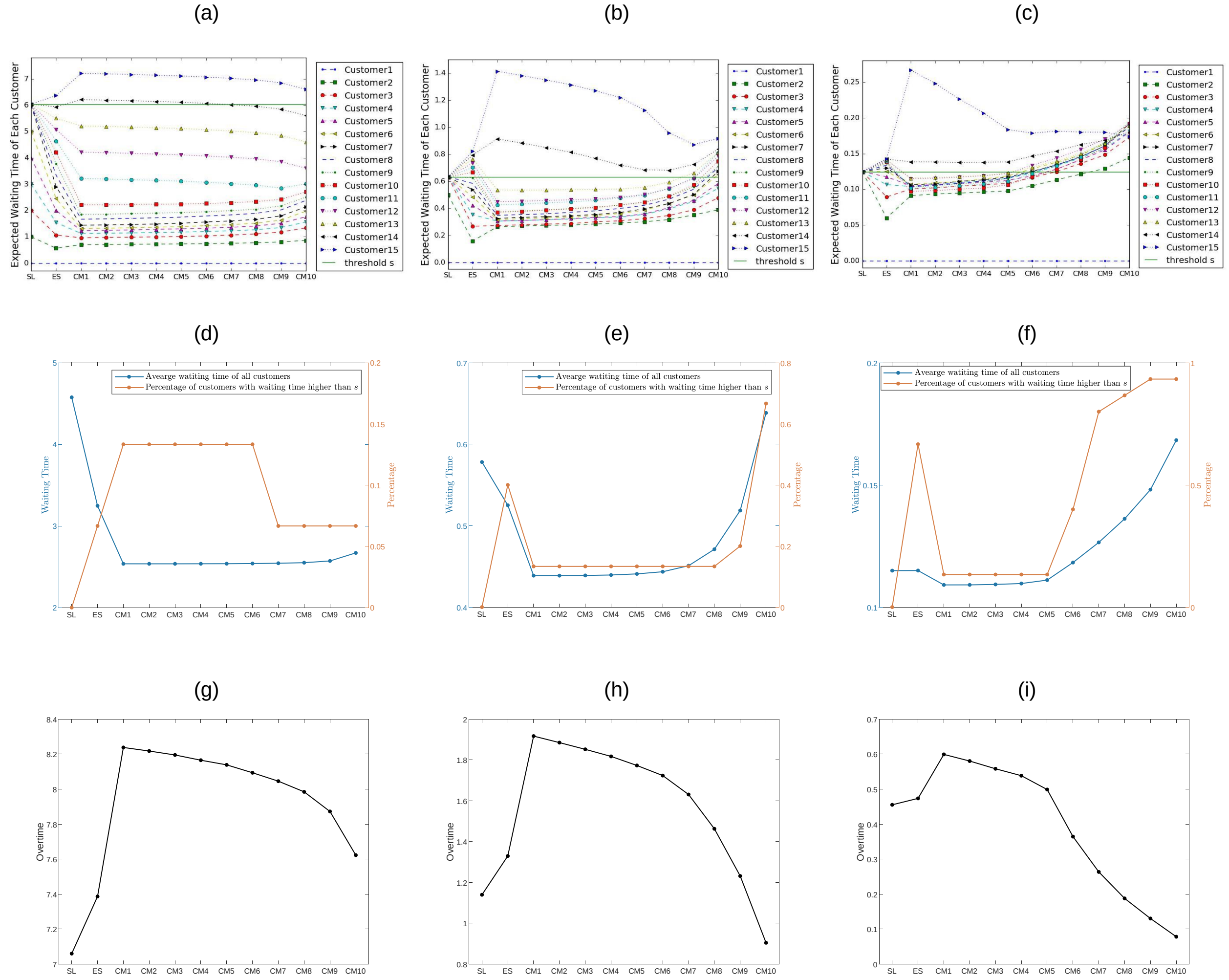
where

$$\begin{aligned} Z(x,S) = \min_{\mathbf{y}} \quad & \sum_{i=1}^{M} c_w y_i + c_o y_{M+1} \\ \text{subject to} \quad & y_{i+1} \geq y_i + S_i - x_{i+1}, \ \forall i = 1, \dots, M, \\ & y_1 = 0, \\ & y_i \geq 0, \qquad \forall i = 2, \dots, M+1. \end{aligned}$$

Here, x_i is the inter-appointment time between customers $i-1$ and i for $i = 2, \dots, M$; S_i is the expected service time of customer i for $i = 1, \dots, M$; y_i is the waiting time of customer i for $i = 1, \dots, M$; and y_{M+1} is the overtime of the system (defined as $D_M - T$). Notice that the solution of Problem (3) depends on the parameter values of waiting cost (c_w) and overtime cost (c_o). In our experiments, we set $c_w = 0.1$ and let $c_o = \frac{\gamma}{1-\gamma} c_w$. We vary the value of γ from zero to 0.9 with an increment of 0.1 and denote the corresponding systems as CM$_1$, …, CM$_{10}$, respectively. To solve Problem (3), we deploy a sample average approximation method.

Once the appointment schedule is computed, the corresponding expected waiting time of each customer and the expected system overtime can be calculated. In Figure 1, we compare the expected waiting time and expected overtime of the three systems under the settings with $M = 15$ and $T = 8$. We vary the service rate μ from one to three to cover three scenarios: (1) overloaded setting ($\frac{M}{\mu} > T$), (2) balanced setting ($\frac{M}{\mu} \approx T$), and (3) underutilized setting ($\frac{M}{\mu} < T$). The corresponding optimal waiting thresholds for

Figure 1. (Color online) Waiting Time and Overtime Comparisons



Notes. (a) Individual waiting ($\mu = 1$). (b) Individual waiting ($\mu = 2$). (c) Individual waiting ($\mu = 3$). (d) Average waiting ($\mu = 1$). (e) Average waiting ($\mu = 2$). (f) Average waiting ($\mu = 3$). (g) Overtime ($\mu = 1$). (h) Overtime ($\mu = 2$). (i) Overtime ($\mu = 3$).

the three scenarios are 6.04, 0.634, and 0.125, respectively. Figure 1, (a)–(c) plots the expected waiting time of each customer in the systems. There, we can compare the waiting times of customers at different positions in one system (by comparing points vertically) or the waiting times of customers at the same position in different systems (by comparing points horizontally). Figure 1, (d)–(f) plots the average expected waiting time among all customers and the percentage of customers with expected waiting time higher than the waiting threshold. Figure 1, (g)–(i) plots the expected system overtime.

From Figure 1, we can see that, first, the average waiting time of all customers under each system decreases in μ. When μ becomes higher, the service time becomes shorter, and, thus, the waiting time of each customer becomes lower. Besides, under the SL system, the expected waiting time of each customer is always below the waiting threshold. In contrast, under the ES and the CM systems, in general, the expected waiting time of each customer increases with the order of arrivals, and the expected waiting times of the later arrivals exceed the waiting threshold. The percentage of customers whose waiting times are higher than the waiting threshold can be as large as 90%. We also see that the average waiting time under the SL system can be either longer or shorter than that under the other systems. This is because under the SL system, the objective is to minimize the makespan such that the expected waiting time of each customer does not exceed the waiting threshold. According to Lemma 2, it is optimal to schedule a few customers to arrive together at time 0 and schedule the rest of the customers such that their expected waiting times are equal to the threshold. Under the EL and CM systems, customers who arrive earlier have expected waiting times less than s, and customers who arrive later have expected waiting times larger than s. Therefore, even though the average waiting time under the EL and CM systems can be shorter, these systems create unfair waiting experiences among customers. This phenomenon is more significant when the system is overloaded or balanced (see Figure 1, (d) and (e)). Notice that, in general, a higher average waiting time implies a lower server idle time. Under these cases, the SL schedule makes use of the service capacity more efficiently than the other schedules.

We also observe that the SL system performs quite well in terms of system overtime under the overloaded and balanced settings. For the underutilized setting (Figure 1(i)), as expected, overtime is quite low (with $T = 8$) in general for all systems.

In Figure 2, we plot the inter-appointment time and the arrival time of each customer for different systems under the overloaded and the balanced settings (with $M = 15$ and $T = 8$). As we can see, first, the inter-appointment times exhibit an increasing pattern under the SL system, a constant pattern under the ES system, and a dome-shaped pattern under the CM system. Under the SL system, the inter-appointment time increases dramatically only at the beginning; the growth diminishes after a few arrivals. This suggests that the simple heuristic schedule we proposed in Section 4.4 (to schedule the first $\lfloor \mu s \rfloor + 1$ customers to arrive together at time 0 and then set a constant inter-appointment time for the following customers) could perform well. From the arrival time plots, we can clearly see that multiple customers arrive together at time 0 under the SL system, and the customer at each position arrives earlier under the SL system than under the other systems. This also explains why the SL system performs quite well in terms of system overtime under the overloaded and balanced settings.

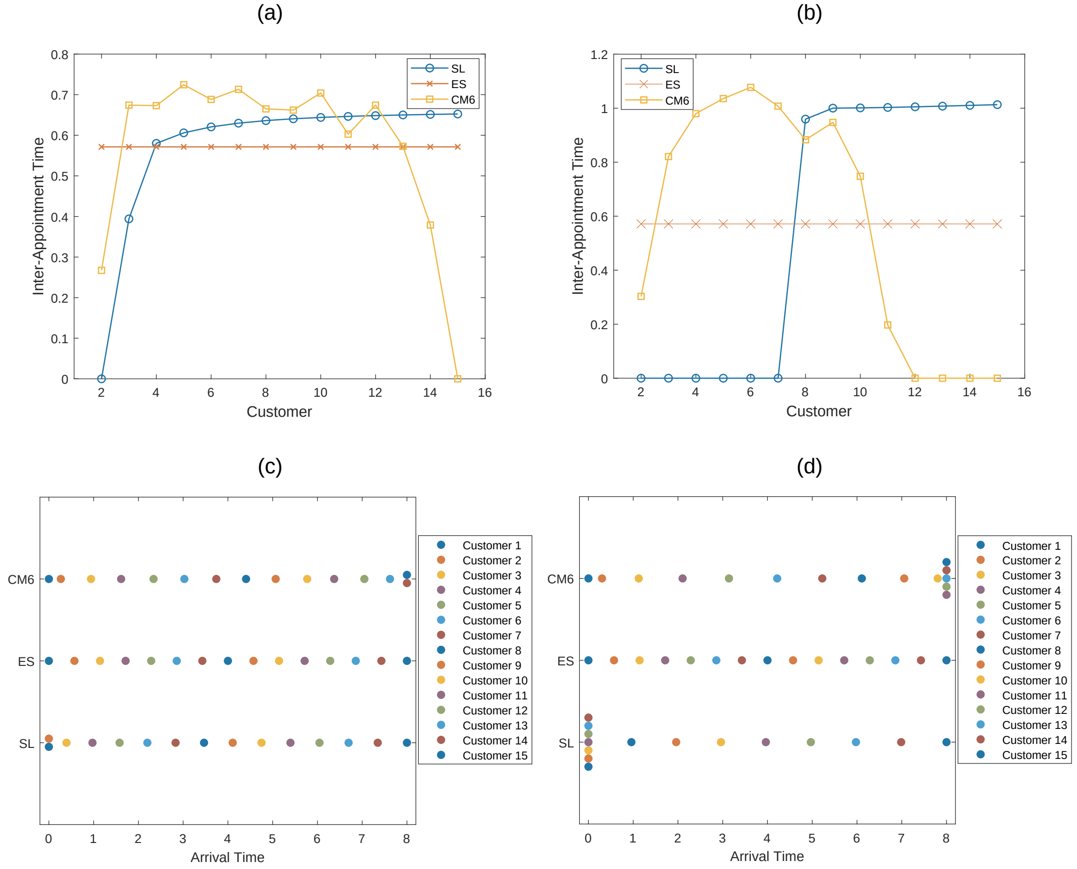
In summary, compared with the ES and the CM systems, the SL system leads to a lower upper bound on each customer's waiting time; ensures a fair waiting experience among customers; and performs quite well in terms of system overtime under the overloaded and balanced settings.

Next, we compare the different systems under given T (time window) and s (waiting threshold). Table 3 reports the total number of customers that can be scheduled into each system under various parameter settings. We see that under the same time window and waiting threshold, the SL system always admits the maximum number of customers, among all systems. This implies that under the service-level constraint, the SL system can be more profitable.

Finally, in Table 4, we compare the makespan of the SL system with that of the ES system under given M (number of customers to be scheduled) and s (waiting threshold).[1] We see that the SL system always leads to a shorter makespan. The relative difference $\left(\text{i.e., } \frac{D_M^{ES} - D_M^{SL}}{D_M^{SL}} \times 100\%\right)$ is larger than 10% under most of the parameter settings, and it reaches 40% for the case with $M = 10$, $\mu = 3$, and $s = 1$.

To conclude this section, we remark that, ultimately, whether the SL system (in general, a service-level-based modeling approach) or the CM system (in general, a cost-minimization modeling approach) is preferable depends on the underlying belief about how customers experience the waiting cost. If, indeed, customers are believed to incur cost linearly in the amount of waiting time they experience, then a cost-minimization approach is justifiable. On the other hand, if customers are tolerant of waiting when the waiting time is below a reference threshold or when it is perceived as being consistent across all customers, then a service-level-based approach would be appropriate. Furthermore, when the waiting cost experienced by customers is believed to increase at an increasing rate with the amount of waiting time (and this is accounted for in the cost-minimization-model formulation), one would expect the cost-minimization approach to favor schedules that equalize the waiting across all customers, which resembles the service-level-based approach.

Figure 2. (Color online) Inter-Appointment Time and Arrival Time Comparisons



Notes. (a) Inter-appointment time ($\mu = 1$). (b) Inter-appointment time ($\mu = 2$). (c) Arrival time ($\mu = 1$). (d) Arrival time ($\mu = 2$).

6. Extensions

In this section, we study several extensions of our system, including one with no-shows, one with mixed Erlang service times, one with multiple servers, and one with probability-based service-level constraints.

6.1. Systems with Customer No-Shows

We first study the possibility of customers' no-shows. Consider a system where each customer has a probability α of showing up, independent of all other events.[2] We interpret the service-level constraint in the way that, if a customer shows up, then his expected waiting time must be less than or equal to s.

Table 3. Number of Customers to Schedule Comparisons

	$s = 0.5$					$s = 1.0$				
	SL	ES	CM$_2$	CM$_6$	CM$_{10}$	SL	ES	CM$_2$	CM$_6$	CM$_{10}$
$T = 8$										
$\mu = 1$	6	4	5	5	4	8	6	6	7	6
$\mu = 2$	14	12	11	12	12	16	14	13	14	15
$\mu = 3$	22	19	17	18	20	25	21	20	21	23
$T = 12$										
$\mu = 1$	9	8	7	8	7	11	9	9	9	9
$\mu = 2$	19	18	16	16	17	23	20	18	19	21
$\mu = 3$	31	29	24	26	29	36	32	28	30	33
$T = 24$										
$\mu = 1$	16	15	14	14	14	19	18	15	17	17
$\mu = 2$	37	35	29	31	35	43	40	33	35	40
$\mu = 3$	59	57	45	49	56	67	63	52	55	62

Recall that when customer m shows up, R_m represents the number of customers found in the system by customer m, upon his arrival. When customer m does not show up, we also let $\bar{R}_m$ represent the number of customers who would have been found in the system by customer m, upon his arrival, if he had shown up. Per the assumption that a customer's probability of showing up is independent of all other events, we have $R_m^* = \bar{R}_m^*$.

If customer m finds i customers upon arrival, then customer $m-1$ must at least find $i-1$ customers upon arrival (or at least would have found i customers, if customer $m-1$ had shown up). Recall that $\Pr\{R_m(x) = i \mid R_{m-1}^* = j\} = c_{j-i+1}(x)$. It is straightforward to see that $\Pr\{R_m(x) = i \mid \bar{R}_{m-1}^* = j\} = c_{j-i}(x)$. Thus, we have

$$p_{m,i}(x) = \alpha \sum_{j=i-1}^{m-2} p_{m-1,j}^* c_{j-i+1}(x) + (1-\alpha) \sum_{j=i}^{m-2} p_{m-1,j}^* c_{j-i}(x).$$

Similar to Algorithm 1 in Section 4, Algorithm 2 can be used to sequentially compute the optimal appointment schedule $\mathbf{T}^* = (T_2^*, \ldots, T_M^*)$ for systems with no-shows.

Table 4. Makespan Comparisons

	$s = 0.5$			$s = 1.0$		
	SL	ES	Relative difference (%)	SL	ES	Relative difference (%)
$M = 10$						
$\mu = 1$	15.0	16.3	8.4	12.4	14.5	16.3
$\mu = 2$	6.2	7.2	16.3	5.4	7.0	29.3
$\mu = 3$	3.8	4.7	23.2	3.4	4.8	40.0
$M = 15$						
$\mu = 1$	23.3	24.6	5.5	19.3	21.4	10.9
$\mu = 2$	9.7	10.7	10.9	8.3	10.0	20.1
$\mu = 3$	5.9	6.8	15.7	5.2	6.7	27.8
$M = 20$						
$\mu = 1$	31.5	32.8	4.1	26.2	28.3	8.1
$\mu = 2$	13.1	14.2	8.1	11.3	13.1	15.3
$\mu = 3$	8.0	8.9	11.8	7.1	8.6	21.4

Algorithm 2 (Computing the Optimal Schedule (No-Shows))

```
Input: number of customers $M$; service rate $\mu$; waiting
       threshold $s$; showing-up rate $\alpha$.
Output: optimal appointment schedule $T^*_m$, for $m = 2, \ldots, M$.
Step 1: for $m = 2, \ldots, \lfloor \frac{\mu s}{\alpha} \rfloor + 1$, set $T^*_m = 0$;
        let $m = \lfloor \frac{\mu s}{\alpha} \rfloor + 2$;
        set $P^*_{m-1,j} = 1$ for $j = 0, \ldots, m-2$.
Step 2: solve for $x_m$ that satisfies
            $\sum_{i=1}^{m-1} [\alpha \sum_{i=n-1}^{m-2} P^*_{m-1,i} c_{i-n+1}(x_m) + (1-\alpha) \sum_{i=n}^{m-2} P^*_{m-1,i} c_{i-n}(x_m)] = \mu s$;
        set $T^*_m = x_m$.
Step 3: if $m = M$, then
            report output $T^*_m$ for $m = 2, \ldots, M$;
            stop;
        otherwise (i.e., if $m < M$),
            for $n = 0, \ldots, m-1$, compute
              $P^*_{m,n} = \alpha \sum_{i=n-1}^{m-2} P^*_{m-1,i} c_{i-n+1}(T^*_m) + (1-\alpha) \sum_{i=n}^{m-2} P^*_{m-1,i} c_{i-n}(T^*_m)$;
            let $m = m + 1$;
            go to Step 2.
```

We can also show that the main structural properties of the optimal schedule (as presented in statements in Section 4) continue to hold for systems with no-shows. The details are given in Theorem 4.

Theorem 4. *For systems with no-shows, the following results hold:*

a. $w_m(x)$ *is decreasing in* x *and* T^*_m *is decreasing in* s, *for* $m = 2, \ldots, M$;

b. $T^*_m = 0$ *for* $m = 2, \ldots, \lfloor \frac{\mu s}{\alpha} \rfloor + 1$; $w^*_m = s$ *for* $m = \lfloor \frac{\mu s}{\alpha} \rfloor + 2, \ldots, M$;

c. $T^*_{m+1} \geq T^*_m$, *for* $m = 2, \ldots, M$;

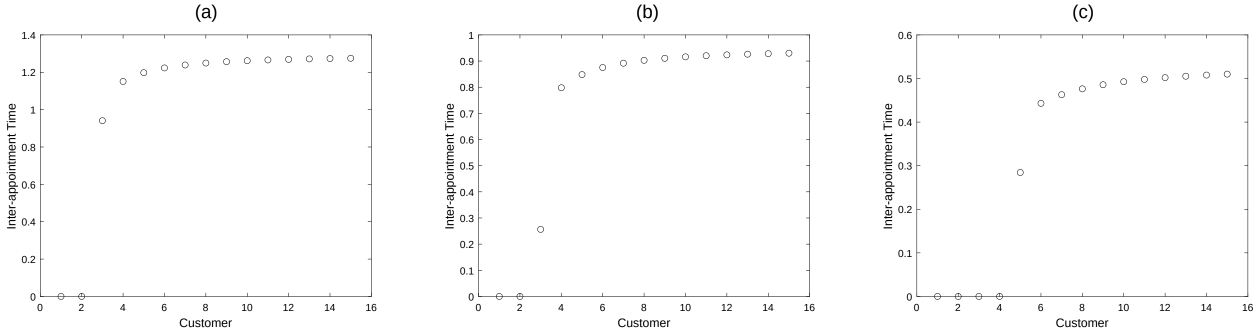
d. $p^*_{m+1,0} \geq p^*_{m,0}$, *for* $m = \lfloor \frac{\mu s}{\alpha} \rfloor + 2, \ldots, M$;

e. $\Pr\left\{W^*_{m+1} \leq \frac{1}{\mu}\right\} \geq \Pr\left\{W^*_m \leq \frac{1}{\mu}\right\}$, *for* $m = \lfloor \frac{\mu s}{\alpha} \rfloor + 2, \ldots, M$; *and*

f. $T^* = s\left(1 + \frac{1}{\mu s}\right) \ln\left(1 + \frac{\alpha}{\mu s}\right)$.

We numerically examine the impact of no-shows. In Figure 3, we plot the optimal inter-appointment times for different values of α ($M = 15$, $\mu = 1$, and $s = 1$). As we can see, first, the optimal inter-appointment time always increases with each arrival, converging quickly to a constant value. This suggests that the heuristic schedule we proposed in Section 4.4 $\left(\text{replacing the constant } s\left(1 + \frac{1}{\mu s}\right) \ln\left(1 + \frac{1}{\mu s}\right) \text{ by } s\left(1 + \frac{1}{\mu s}\right) \ln\left(1 + \frac{\alpha}{\mu s}\right)\right)$ should perform well. Second, as the no-show probability increases, it is optimal to set the inter-appointment time shorter and schedule

Figure 3. Optimal Appointment Schedule (No-Shows)



Notes. (a) $\alpha = 0.9$. (b) $\alpha = 0.6$. (c) $\alpha = 0.3$.

customers to arrive earlier. It is also optimal to schedule more customers to arrive at time 0.

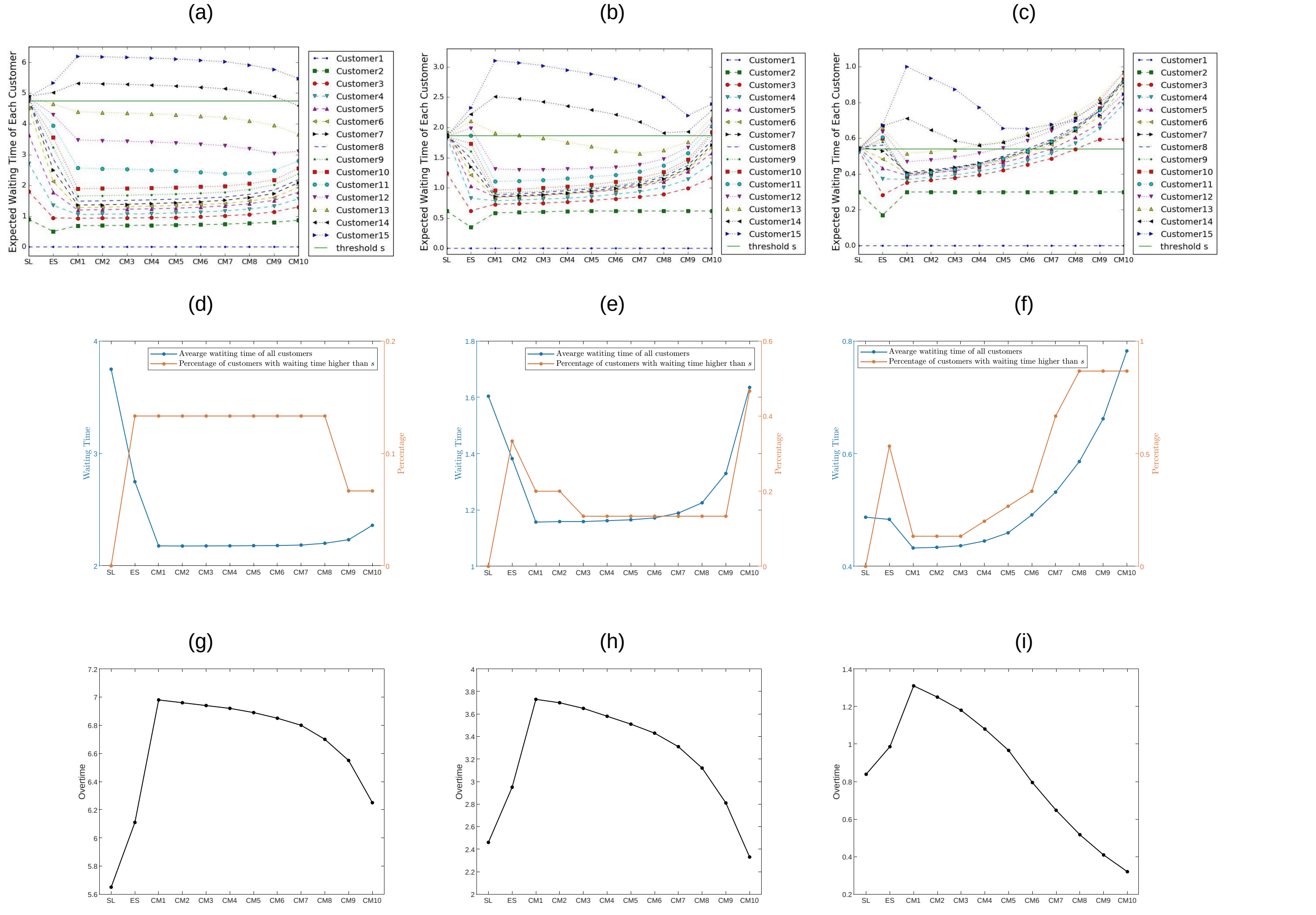
Next, we compare the SL, ES, and CM systems with no-shows. As in Section 5.2, we first look at the situation under given M and T. Figure 4 plots the expected waiting time and expected overtime of these systems with different values of α ($M = 15$, $T = 8$, and $\mu = 1$; the corresponding optimal waiting thresholds are 4.75, 1.86, and 0.54, respectively). We see that the SL system always leads to a lower upper bound on each customer's expected waiting time, which ensures fair waiting experiences among customers. Moreover, when the no-show rate is not high, the SL system performs quite well in terms of system overtime. Notice that low and median no-show rates correspond, respectively, to overloaded and balanced settings, as in Section 5.2. Thus, the observations for systems with no-shows are consistent with those for the original systems.

We then study the case under given T and s, with the results presented in Table 5 ($T = 8$); and the case under given M and s, with the results presented in Table 6 ($M = 15$). We see that, for the earlier case, the SL system always admits a higher number of customers and, thus, can be more profitable; for the later case, the SL system always leads to a shorter makespan, and the advantage is more significant under a higher no-show rate. These

Table 5. Number of Customers to Schedule Comparisons (No-Shows)

	$s = 0.5$					$s = 1.0$				
	SL	ES	CM$_2$	CM$_6$	CM$_{10}$	SL	ES	CM$_2$	CM$_6$	CM$_{10}$
$\mu = 1$										
$\alpha = 0.9$	7	6	6	6	5	8	7	7	7	6
$\alpha = 0.6$	8	7	7	7	6	11	9	9	10	8
$\alpha = 0.3$	14	12	11	11	8	20	16	15	18	15
$\mu = 2$										
$\alpha = 0.9$	15	13	12	13	13	18	15	14	15	16
$\alpha = 0.6$	20	18	15	17	16	25	21	19	20	22
$\alpha = 0.3$	35	31	25	30	28	47	39	34	38	41

Figure 4. (Color online) Waiting Time and Overtime Comparisons (No-Shows)



Notes. (a) Individual waiting ($\alpha = 0.9$). (b) Individual waiting ($\alpha = 0.6$). (c) Individual waiting ($\alpha = 0.3$). (d) Average waiting ($\alpha = 0.9$). (e) Average waiting ($\alpha = 0.6$). (f) Average waiting ($\alpha = 0.3$). (g) Overtime ($\alpha = 0.9$). (h) Overtime ($\alpha = 0.6$). (i) Overtime ($\alpha = 0.3$).

Table 6. Makespan Comparisons (No-Shows)

	s = 0.5			s = 1.0		
	SL	ES	Relative difference (%)	SL	ES	Relative difference (%)
μ = 1						
α = 0.9	21.7	23.0	5.9	17.7	19.7	11.2
α = 0.6	16.3	17.6	8.2	12.6	14.6	16.4
α = 0.3	9.2	10.6	15.7	6.3	8.4	32.1
μ = 2						
α = 0.9	8.9	9.9	11.1	7.6	8.9	17.8
α = 0.6	6.3	7.3	15.8	5.2	6.5	25.6
α = 0.3	3.2	4.2	30.5	2.5	3.6	46.3

observations for systems with no-shows again coincide with those for the original systems.

6.2. Systems with Mixed Erlang Service Times

Next, we extend our analysis to consider systems with mixed Erlang distributed service times. This extension is salient because any nonnegative random variable can be approximated arbitrarily closely by a mixed Erlang random variable (see Tijms 1995).

Suppose that the service time follows a mixed Erlang distribution with density

$$f(x;\boldsymbol{\beta},\mathbf{I},\mu)=\sum_{k=1}^{N}\beta_k\frac{\mu^{I_k}x^{I_k-1}e^{\mu x}}{(I_k-1)!}=\sum_{k=1}^{N}\beta_k f_{\mathrm{Er}}(x;I_k,\mu),$$

where the positive integers $\mathbf{I}=(I_1,\ldots,I_N)$ with $I_1<\cdots<I_N$ are the shape parameters of the Erlang distributions; and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_N)$ with $\beta_k>0$ and $\sum_{k=1}^{N}\beta_k=1$ are the weights of the Erlang distributions $f_{\mathrm{Er}}(x;I_k,\mu)$ in the mixture.

The service time of a customer following the above defined mixed Erlang distribution can be interpreted as that the customer needs to go through I_k phases of service, with probability β_k, for $k=1,\ldots,N$; the duration of each phase of service follows an exponential distribution with rate μ. Let $\hat{I}=\sum_{k=1}^{N}\beta_k I_k$ denote the expected number of phases each customer needs to go through. Now, the expected waiting time of a customer depends on the number of phases (instead of the number of customers) found in the system upon his arrival. Therefore, by using the same analysis as for systems with exponential service times but replacing "the number of customers" with "the number of phases," we can compute the optimal appointment schedule $\mathbf{T}^*=(T_1^*,\ldots,T_M^*)$ sequentially. We can also show that the main structural properties of the optimal schedule (as presented in statements in Section 4) still hold for systems with mixed Erlang service times. We summarize the analytical results in Theorem 5 and provide the computing procedures in Algorithm 3 (we define $P_{m,n}(x)=\sum_{i=n}^{(m-1)I_N}p_{m,i}(x)$ and $P_{m,n}^*=\sum_{i=n}^{(m-1)I_N}p_{m,i}^*$ for $n=0,\ldots,(m-1)I_N$). Detailed analytical developments can be found in the online supplement.

Theorem 5. *For systems with mixed Erlang service times, the following results hold*:

a. $w_m(x)$ *is decreasing in* x, *and* T_m^* *is decreasing in* s, *for* $m=2,\ldots,M$;

b. $T_m^*=0$ *for* $m=2,\ldots,\lfloor\frac{\mu s}{\hat{I}}\rfloor+1$; $w_m^*=s$ *for* $m=\lfloor\frac{\mu s}{\hat{I}}\rfloor+2,\ldots,M$;

c. $T_{m+1}^*\geq T_m^*$, *for* $m=2,\ldots,M$;

d. $T_m^*\geq\frac{\hat{I}}{\mu}$, *for* $m=\lfloor\frac{\mu s}{\hat{I}}\rfloor+3,\ldots,M$;

e. $p_{m+1,0}^*\geq p_{m,0}^*$, *for* $m=\lfloor\frac{\mu s}{\hat{I}}\rfloor+2,\ldots,M$;

f. $\Pr\left\{W_{m+1}^*\leq\frac{1}{\mu}\right\}\geq\Pr\left\{W_m^*\leq\frac{1}{\mu}\right\}$, *for* $m=\lfloor\frac{\mu s}{\hat{I}}\rfloor+2,\ldots,M$; *and*

g. $\lim_{M\to\infty}p_{M,i}^*$ *exists*, *for* $i=0,1,\ldots$.

Algorithm 3 (Computing the Optimal Schedule (Mixed Erlang Service Times))

```
Input: number of customers $M$; common service rate $\mu$; waiting threshold $s$; number of phases for
       each type of Erlang distribution $I_k$, for $k=1,\ldots,N$; weight of each Erlang distribution $\beta_k$,
       for $k=1,\ldots,N$.
Output: optimal appointment schedule $T_m^*$, for $m=2,\ldots,M$.
Step 1: for $m=2,\ldots,\lfloor\frac{\mu s}{\hat{I}}\rfloor+1$, set $T_m^*=0$;
        let $m=\lfloor\frac{\mu s}{\hat{I}}\rfloor+2$;
        set $P_{m-1,(n-I_k)}^*=1$, for $n=0,\ldots,(m-1)I_N$.
Step 2: solve for $x_m$ that satisfies $\sum_{n=1}^{(m-1)I_N}\sum_{k=1}^{N}\beta_k\sum_{i=0}^{(m-2)I_N-n+I_k}P_{m-1,(n-I_k+i)^+}^*c_i(x_m)=\mu s$;
        set $T_m^*=x_m$.
Step 3: if $m=M$, then
            report output $T_m^*$, for $m=2,\ldots,M$;
            stop;
        otherwise (i.e., if $m<M$),
            for $n=0,\ldots,(m-1)I_N$, compute
            $P_{m,n}^*=\sum_{k=1}^{N}\beta_k\sum_{i=0}^{(m-2)I_N-n+I_k}P_{m-1,(n-I_k+i)^+}^*c_i(T_m^*)$;
            let $m=m+1$;
            go to Step 2.
```

6.3. Systems with Multiple Servers

Now, we extend our analysis to study systems with multiple servers.

Consider a system with $\mathcal{N}$ (a finite positive integer) parallel and identical servers. We continue to use similar notation. Because there are $\mathcal{N}$ servers, it is optimal to schedule the first $\mathcal{N}$ customers to arrive together at time 0, and their waiting time equals zero. Notice that, when n servers are occupied (n customers are under service) simultaneously, the time it takes to complete one service is exponentially distributed with rate μn. Thus, if customers $1,\ldots,m$ are scheduled to arrive together at time 0, then the expected waiting time of customer m equals $\frac{(m-\mathcal{N})^+}{\mu\mathcal{N}}$. The following lemma corresponds to Lemmas 1 and 2 for systems with a single server.

Lemma 5. *For systems with multiple servers, the following results hold*:

a. *$w_m(x)$ is decreasing in x and T^*_m is decreasing in s, for $m=2,\ldots,M$; and*

b. *$T^*_m=0$ for $m=2,\ldots,\lfloor\mu\mathcal{N}s\rfloor+\mathcal{N}$; $w^*_m=s$ for $m=\lfloor\mu\mathcal{N}s\rfloor+\mathcal{N}+1,\ldots,M$.*

Similar to the case of systems with a single server, we can use the embedded Markov-chain approach to compute the optimal appointment schedule $\mathbf{T}^*=(T^*_1,\ldots,T^*_M)$ sequentially. We provide the computing procedures in Algorithm 4. Detailed developments can be found in the online supplement. However, because of the complexity of the $p^*_{m,i}$ expressions, it is difficult to apply the theory of majorization and analytically characterize the structural properties. Through extensive numerical experiments (see the online supplement), we find that important structural properties of the optimal schedule (e.g., T^*_m increases and converges) still hold for systems with multiple servers.

Algorithm 4 (Computing the Optimal Schedule (Multiple Servers))

```
Input: number of customers $M$; service rate $\mu$; waiting threshold $s$; number of servers $\mathcal{N}$.
Output: optimal appointment schedule $T^*_m$, for $m=2,\ldots,M$.
Step 1: for $m=2,\ldots,\lfloor\mu\mathcal{N}s\rfloor+\mathcal{N}$, set $T^*_m=0$;
        let $m=\lfloor\mu\mathcal{N}s\rfloor+\mathcal{N}+1$;
        set $p^*_{m-1,m-2}=1$ and $p^*_{m-1,j}=0$ for $j=0,\ldots,m-3$.
Step 2: solve for $x_m$ that satisfies $\sum_{i=\mathcal{N}}^{m-1}\frac{i-\mathcal{N}+1}{\mu\mathcal{N}}\sum_{j=i-1}^{m-2} p^*_{m-1,j}\frac{(\mu\mathcal{N}x)^{j-i+1}}{(j-i+1)!}e^{-\mu\mathcal{N}x}=s$;
        set $T^*_m=x_m$.
Step 3: if $m=M$, then
            report output $T^*_m$, for $m=2,\ldots,M$;
            stop;
        otherwise (i.e., if $m<M$),
            for $i=1,\ldots,\mathcal{N}-1$, compute
                $p^*_{m,i}=\sum_{j=i-1}^{\mathcal{N}-1}p^*_{m-1,j}\frac{(j+1)!}{i!(j-i+1)!}(1-e^{-\mu T^*_m})^{j-i+1}(e^{-\mu T^*_m})^i+\sum_{j=\mathcal{N}}^{m-2}p^*_{m-1,j}\cdot\int_0^{T^*_m}\left[\frac{\mathcal{N}!}{i!(\mathcal{N}-i)!}(1-e^{-\mu(T^*_m-t)})^{\mathcal{N}-i}e^{-\mu(T^*_m-t)i}\right]\frac{(\mu\mathcal{N})^{j-\mathcal{N}+1}t^{j-\mathcal{N}}e^{-\mu\mathcal{N}t}}{(j-\mathcal{N})!}dt$;
            for $i=\mathcal{N},\ldots,m-1$, compute $p^*_{m,i}=\sum_{j=i-1}^{m-2}p^*_{m-1,j}\frac{(\mu\mathcal{N}T^*_m)^{j-i+1}}{(j-i+1)!}e^{-\mu\mathcal{N}T^*_m}$;
            compute $p^*_{m,0}=1-\sum_{i=1}^{m-1}p^*_{m,i}$;
            let $m=m+1$;
            go to Step 2.
```

6.4. Systems with Probability-Based Service-Level Constraints

Last, we extend our analysis to consider systems with probability-based service-level constraints.

Up to now, our modeling and analysis have been focusing on the case where the service-level constraint is in the form of a limit on the expected waiting time of each individual customer. Notice that it is also common and natural that the service-level constraint is in another form—namely, a limit on the probability of long waiting (see Millhiser et al. 2012 and Millhiser and Veral 2015 for detailed discussion). In particular, we consider a service-level constraint of the form:

$$\Pr\left\{W_m>\frac{1}{\mu}\right\}\le s,\ \text{for } m=1,\ldots,M.$$

This constraint can be interpreted as, for every customer, the probability that he waits longer than his expected service time must be less than or equal to a predetermined value s.

First, the same as Proposition 1 for the case under expectation-based service-level constraints, we prove that under probability-based service-level constraints, the prospective scheduling system (where the appointment times of all customers are decided together at once) and the sequential scheduling system (where the appointment time of each customer is set one after another at the time when service is requested) are equivalent, and, thus, a prospective schedule can be obtained from a sequential scheduling approach.

Proposition 5. *For systems with probability-based service-level constraints, the following two problems share the same solution*:

1. *Solving*

$$\min \quad \sum_{i=1}^{M}T_m+w_M+\frac{1}{\mu}$$
$$\text{subject to}\quad \Pr\left\{W_m>\frac{1}{\mu}\right\}\le s,\ \text{for } m=1,\ldots,M.$$

2. *For $m=2,\ldots,M$, sequentially solving*

$$\min \quad T_m$$
$$\text{subject to}\quad \Pr\left\{W_m>\frac{1}{\mu}\,\middle|\,T_n=T^*_n,\ \text{for } n=1,\ldots,m-1\right\}\le s.$$

Let $q_m(x)$ denote the probability that the waiting time of customer m is larger than $\frac{1}{\mu}$, with $T_m=x$ and $T_n=T^*_n$ for $n=2,\ldots,m-1$. Then, we have $T^*_m=\arg\min_{q_m(x)\le s}x$.

We also define $q^*_m = q_m(T^*_m)$. Let $N^* = \max\{n \geq 2 \mid \sum_{i=0}^{n-1} \frac{1}{i!e} \leq s\}$. Now, we can obtain results, which correspond to Lemmas 1 and 2 for systems with expectation-based service-level constraints.

Lemma 6. *For systems with probability-based service-level constraints, the following results hold:*

a. *$q_m(x)$ is decreasing in x and T^*_m is decreasing in s, for $m = 2, \ldots, M$; and*

b. *$T^*_m = 0$ for $m = 2, \ldots, N^*$; $q^*_m = s$ for $m = N^* + 1, \ldots, M$.*

Next, we apply the transient analysis. If customer m finds i customers ($i \geq 1$) in the system, upon his arrival, then his waiting time is Erlang-distributed with shape i and rate μ. Therefore, $\Pr\{W_m > \frac{1}{\mu} \mid R_m = i\} = \sum_{j=0}^{i-1} \frac{1}{j!e}$. Following an analysis similar to that in Section 4.1, we obtain $q_m(x) = \sum_{i=0}^{m-2} \frac{1}{i!e} P_{m,i+1}(x)$, $P_{m,0}(x) = 1$, and $P_{m,n}(x) = \sum_{i=n-1}^{m-2} P^*_{m-1,i} c_{i-n+1}(x)$ for $n = 1, \ldots, m-1$; as well as $q^*_m = \sum_{i=0}^{m-2} \frac{1}{i!e} P^*_{m,i+1}$, $P^*_{m,0} = 1$, and $P^*_{m,n} = \sum_{i=n-1}^{m-2} P^*_{m-1,i} c^*_{i-n+1}$ for $n = 1, \ldots, m-1$.

Using the above formulas, we can again compute the optimal appointment schedule $\mathbf{T}^* = (T^*_2, \ldots, T^*_M)$ sequentially (see Algorithm 5). Because of the complexity of the $p^*_{m,i}$ expressions, it is difficult to analytically characterize the structural properties of T^*_m. Through extensive numerical experiments (see the online supplement), we observe that T^*_m converges quickly in all the cases tested. However, T^*_m increases with m only when μ is relatively large. When the service rate is low, T^*_m may not be monotonic. We remark that these observations are consistent with results (figure 3) in Millhiser and Veral (2015).

Algorithm 5 (Computing the Optimal Schedule (Probability-Based Service-Level Constraints))

```
Input: number of customers $M$; service rate $\mu$; waiting
        threshold $s$.
Output: optimal appointment schedule $T^*_m$, for $m =$
        $2, \ldots, M$.
Step 1: set $N^* = \max\left\{n \geq 2 \middle| \sum_{i=0}^{n-1} \frac{1}{i!e} \leq s\right\}$;
        for $m = 2, \ldots, N^* + 1$, set $T^*_m = 0$;
        let $m = N^* + 2$;
        set $P^*_{m-1,j} = 1$ for $j = 0, \ldots, m-2$.
Step 2: solve for $x_m$ that satisfies $\sum_{i=1}^{m-1} \frac{1}{(i-1)!e} \sum_{j=i-1}^{m-2}$
          $P^*_{m-1,j} c_{j-i+1}(x_m) = \mu s$;
        set $T^*_m = x_m$.
Step 3: if $m = M$, then
        report output $T^*_m$ for $m = 2, \ldots, M$;
        stop;
otherwise (i.e., if $m < M$),
        set $P^*_{m,0} = 0$;
        for $n = 1, \ldots, m-1$, compute $P^*_{m,n} = \sum_{i=n-1}^{m-2} P^*_{m-1,i}$
        $c_{i-n+1}(T^*_m)$;
        let $m = m + 1$;
        go to Step 2.
```

7. Conclusion

In this paper, we consider an appointment-scheduling problem with a service-level constraint on the expected waiting time of each individual customer. We show that under such a setting, a prospective schedule can be obtained from a sequential scheduling approach. By applying the theory of majorization, we analytically characterize the structure of the optimal schedule. We study the limiting behavior of the system and prove the convergence to the D/M/1 queueing system. Based on this limiting behavior, we propose a simple, yet practical, heuristic schedule that is asymptotically optimal. We also develop algorithms to compute the optimal appointment schedules in a fixed time window and use these algorithms to compare the service-level-constrained appointment system with the widely studied equal-space and cost-minimization systems. Finally, we investigate various extended settings of our analysis, including customer no-shows, mixed Erlang service times, multiple servers, and probability-based service-level constraints.

For future research, it will be useful to further extend our analysis to study systems consisting of multiple classes of customers with heterogeneous service times. It will also be interesting to investigate customers' choice preferences under the service-level-constrained appointment system. For example, will customers prefer later appointments to earlier ones, due to the potential better waiting experience, as we discussed at the end of Section 4.3?

Acknowledgments

The authors thank Francis de Véricourt (the department editor), the associate editor, and the reviewers for valuable comments and suggestions. The authors are alphabetically ordered and with equal contributions.

Endnotes

[1] The CM system is not valid in this setting because it needs a predetermined time window T as input.

[2] There exists literature that considers customer- or time-dependent no-show rates. For mathematical tractability, we consider constant no-show rates in this paper. However, our analysis can also be extended to consider systems with customer- or time-dependent no-show rates.

References

Ahmadi-Javid A, Jalali Z, Klassen KJ (2017) Outpatient appointment systems in healthcare: A review of optimization studies. *Eur. J. Oper. Res.* 258(1):3–34.

Baron O, Berman O, Krass D, Wang J (2017) Strategic idleness and dynamic scheduling in an open-shop service network: Case study and analysis. *Manufacturing Service Oper. Management* 19(1):52–71.

Begen MA, Queyranne M (2011) Appointment scheduling with discrete random durations. *Math. Oper. Res.* 36(2):240–257.

Cayirli T, Veral E (2003) Outpatient scheduling in healthcare: A review of literature. *Production Oper. Management* 12(4): 519–549.

Chen RR, Robinson LW (2014) Sequencing and scheduling appointments with potential call-in patients. *Production Oper. Management* 23(9):1522–1538.

Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35(11):1003–1016.

Fries BE, Marathe VP (1981) Determination of optimal variable-sized multiple-block appointment systems. *Oper. Res.* 29(2):324–345.

Gupta D, Denton B (2008) Appointment scheduling in healthcare: Challenges and opportunities. *IIE Trans.* 40(9):800–819.

Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management Sci.* 54(3):565–572.

Jansson B (1966) Choosing a good appointment system—A study of queues of the type (*D*, *M*, 1). *Oper. Res.* 14(2):292–312.

Jiang R, Ryu M, Xu G (2020) Data-driven distributionally robust appointment scheduling over Wasserstein balls. Working paper, University of Michigan, Ann Arbor, MI.

Jiang R, Shen S, Zhang Y (2017) Integer programming approaches for appointment scheduling with random no-shows and service durations. *Oper. Res.* 65(6):1638–1656.

Jouini O, Benjaafar S, Lu B, Legros B (2022) Appointment-driven queueing systems with non-punctual customers. *Queueing Systems* 101(1–2):1–56.

Keskinocak P, Savva N (2020) A review of the healthcare-management (modeling) literature published in *Manufacturing & Service Operations Management*. *Manufacturing Service Oper. Management* 22(1):59–72.

Kong Q, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Oper. Res.* 61(3):711–726.

Kong Q, Li S, Liu N, Teo CP, Yan Z (2021) Appointment scheduling under time-dependent patient no-show behavior. *Management Sci.* 66(8):3480–3500.

Kuiper A, Kemper B, Mandjes M (2015) A computational approach to optimized appointment scheduling. *Queueing Systems* 79(1):5–36.

Luo J, Kulkarni VG, Ziya S (2012) Appointment scheduling under patient no-shows and service interruptions. *Manufacturing Service Oper. Management* 14(4):670–684.

Mak HY, Rong Y, Zhang J (2015) Appointment scheduling with limited distributional information. *Management Sci.* 61(2):316–334.

Marshall AW, Olkin I, Arnold BC (2011) *Inequalities: Theory of Majorization and Its Applications*, *Springer Series in Statistics* (Springer, New York).

Millhiser WP, Veral EA (2015) Designing appointment system templates with operational performance targets. *IIE Trans. Healthcare Systems Engrg.* 5(3):125–146.

Millhiser WP, Veral EA, Valenti BC (2012) Assessing appointment systems' operational performance with policy targets. *IIE Trans. Healthcare Systems Engrg.* 2(4):274–289.

Mondschein SV, Weintraub GY (2003) Appointment policies in service operations: A critical analysis of the economic framework. *Production Oper. Management* 12(2):266–286.

Müller A, Stoyan D (2002) *Comparison Methods for Stochastic Models and Risks*, *Wiley Series in Probability and Statistics* (Wiley, Chichester, UK).

Olivares M, Terwiesch C, Cassorla L (2008) Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Sci.* 54(1):41–55.

Qi J (2017) Mitigating delays and unfairness in appointment systems. *Management Sci.* 63(2):566–583.

Robinson LW, Chen RR (2003) Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Trans.* 35(3):295–307.

Robinson LW, Chen RR (2011) Estimating the implied value of the customer's waiting time. *Manufacturing Service Oper. Management* 13(1):53–57.

Tijms HC (1995) *Stochastic Models: An Algorithmic Approach*, *Wiley Series in Probability and Statistics* (Wiley, Chichester, UK).

Wang WY, Gupta D (2014) Nurse absenteeism and staffing strategies for hospital inpatient units. *Manufacturing Service Oper. Management* 16(3):439–454.

Wang S, Liu N, Wan G (2020) Managing appointment-based services in the presence of walk-in customers. *Management Sci.* 66(2):667–686.

Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production Oper. Management* 23(5):788–801.

Zacharias C, Pinedo M (2017) Managing customer arrivals in service systems with multiple identical servers. *Manufacturing Service Oper. Management* 19(4):639–656.

Zacharias C, Yunes T (2020) Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management Sci.* 66(2):744–763.

Zhou Y, Parlar M, Verter V, Fraser S (2021) Surgical scheduling with constrained patient waiting times. *Production Oper. Management* 30(9):3253–3271.