OPERATIONS RESEARCH
Articles in Advance, pp. 1–16
ISSN 0030-364X | EISSN 1526-5463

DOI 10.1287/opre.1080.0585

Staffing to Maximize Profit for Call Centers with Alternate Service-Level Agreements

Opher Baron, Joseph Milner
Rotman School of Management, University of Toronto, Toronto, Ontario, Canada M5S 3E6
{opher.baron@rotman.utoronto.ca, milner@rotman.utoronto.ca}

To ensure quality from outsourced call centers, firms sign *service-level agreements* (SLAs). These define service measures such as what constitutes an acceptable delay or an acceptable abandonment rate. They may also dictate penalties for failing to meet agreed-upon targets. We introduce a *period-based SLA* that measures performance over a short duration such as a rush hour. We compare it to alternate SLAs that measure service by individual and over a long horizon. To measure the service levels for these SLAs, we develop several approximations. We approximate the probability an acceptable delay is met by generalizing the heavy-traffic quality and efficiency driven regime. We also provide a new approximation for the abandonment rate. Further, we prove a central limit theorem for the probability of meeting a service level measured by the percentage of customers acceptably served during a period. We demonstrate how an outsourced call center operating in an environment with uncertain demand and abandonment can determine its staffing policy to maximize the expected profit for these SLAs. Numerical experiments demonstrate a high degree of accuracy for the approximations and the resulting staffing levels. We indicate several salient features of the behavior of the period-based SLA.

Subject classifications: queues: applications; approximations; balking and reneging.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received February 2006; revisions received September 2006, August 2007, December 2007, February 2008; accepted March 2008. Published online in *Articles in Advance*.

1. Introduction

Firms sign service-level agreements (SLAs) with outsourced call centers to ensure quality in the handling of their customers' calls. Within these agreements are terms that describe services to be provided, such as the hours of operation and the types of facilities; and terms that describe the service level, such as the average actual handling time (the service time), the acceptable abandonment rate, and acceptable customer delay times. Although such SLAs are common in industry, their terms and the extent of their specifications vary. In some cases, the service level is defined simply as requiring a given percentage of customers to be served within a given delay, e.g., the so-called 80/20 contract, where 80% of calls are to be answered in 20 seconds (see, e.g., Jackson 2002). In other cases, greater detail is placed in the contract, such as specifying maximum abandonment rates, nonlinear penalties, escalation procedures, etc.

Although an SLA states a number of terms that guide a call center's operations, the main determinants of how a call center will work are often left to its own management. In particular, call centers must determine how many servers to staff at any given time, how to train their staff, and how calls should be routed within the call center. In this paper, we investigate how staffing levels affect compliance with alternate SLAs. Because the staffing decision has been the focus of quite a few recent papers (see Gans et al. 2003 for an excellent tutorial and literature review on the topic, and Mandelbaum 2004 for an extensive bibliography), we distinguish our work in several ways.

First, we consider how the period length over which a service level is measured affects the customers' experience. We show this period to be an important consideration that has not been well studied in the literature. Second, we consider uncertain arrival rates and allow abandonment from the queue. Third, we apply our results to the problem of determining the staffing level that maximizes the expected profit of an outsourced call center subject to the alternate SLAs.

The SLAs we study measure the number or percentage of customers acceptably served. Acceptable service may be defined as being served within an acceptable delay or simply as being served (i.e., not abandoning). In practice, failure to meet a specified service level can result in a penalty to the call center, often in the form of a discount in the revenue received per call (e.g., OutsourcingBestPractices.com 2001). Similar measurements of service are common in other contexts—e.g., de Vericourt and Jennings (2006) study regulations on staffing levels for nurses where service level is measured by an acceptable delay.

The SLAs we consider differ in the time scale for which the service level is measured. Suppose that the mean time to serve a customer is on the order of minutes, the time over which the mean arrival rate may be presumed constant, which we refer to as a *natural period*, is on the order of an

hour, and a contract is established for several months to a year. The individual-based (IB) SLA considers the individual experience; the period-based (PB) SLA, the experience of customers during a natural period; and the horizon-based (HB) SLA, the experience of all customers over the contract duration. Under the IB-SLA, the call center is penalized for each customer not acceptably served. Under the PB-SLA, a penalty proportional to a period's demand is incurred if the call center does not achieve a service level (SL) measured by the percentage of customers acceptably served. Under the HB-SLA, a penalty proportional to the number of customers arriving over a horizon is incurred if the service level is not achieved. We determine the probability that a penalty is assessed for each of these SLAs.

In this paper, we introduce and analyze the PB-SLA, and contrast the staffing decisions for it with those of the IB- and HB-SLAs. In practice, many contracts have been defined only in terms of the HB-SLA (see www.techagreements.com for several public domain agreements), whereas much of the previous research has focussed on using costs as in the IB-SLA (e.g., Borst et al. 2004) to determine staffing levels. We believe that the popularity of the HB-SLA stems from its allowance for variability in the system because it only penalizes a call center that does not meet a service level on average. However, the HB-SLA allows the call center to provide disparate service to customers that arrive in periods of greater and lesser demand. For example, a call center may provide good service during natural periods of expected high demand while reducing service levels during low-demand hours. As we discuss below, staffing to meet an acceptable service level results in economies of scale so that there is incentive to provide such disparate service. Such behavior has been observed in practice (a case is discussed in Milner and Olsen 2008). Further, it is not clear that the firm outsourcing its call center needs would observe this behavior unless it audited the call center's performance by the period, which is the point of the PB-SLA we introduce. In comparison, the IB-SLA treats all customers similarly, but does not allow for the variability inherent in call centers. Because of variability, penalties would be incurred under the IB-SLA in all circumstances. This is in contrast to common practice in outsourcing relationships, where penalties are used only for consistent failure to meet specified targets.

The PB-SLA captures the benefits of each, while mitigating their weaknesses. Through the PB-SLA, the service level is imposed on all periods, but variability within a period is recognized and not penalized. By measuring the service each period, the call center must provide more even treatment of customers across periods. In addition, we demonstrate that the PB-SLA responds to uncertainty in the demand by staffing to a higher percentile of the demand than either the IB- or HB-SLA. Further, many standard call center data reports provide aggregate data that support measurement of a PB-SLA (e.g., Gans et al. 2003) so that its implementation is relatively facile.

Previous related research includes work on approximations for large-scale queuing systems, optimization of staffing levels for call centers, and the inclusion of variable demand and abandonments in call center models. In their seminal paper on the so-called quality and efficiency driven (QED) regime, Halfin and Whitt (1981) analyzes the $M/M/N$ queue and establish that if and only if N increases with λ so that $\sqrt{N}(1-\rho) \to \beta$ as $N \to \infty$, where $\rho = \lambda/(N\mu)$ and β is a positive constant, then the probability of waiting approaches a limit strictly between zero and one. This provides theoretical justification for the well-known square root safety staffing rule, $N \approx R + \beta\sqrt{R}$, where R is the offered service load, λ/μ (Whitt 1992). Garnett et al. (2002) extends these results to include abandonment (in the Erlang-A Markovian $M/M/N+M$ model) and shows that in this case $\beta \leqslant 0$ is possible. Additional models that include customer abandonment and address more general arrival or patience distributions may be found in Brandt and Brandt (1999, 2002), Zeltyn and Mandelbaum (2005), Mandelbaum and Zeltyn (2006), Whitt (2005b, 2006a), and references therein. Although we focus on the Markovian $M/M/N+N$ model, we develop approximations where, rather than fixing the probability of waiting, we fix the probability that the wait exceeds a given acceptable service delay, d. Because many contracts in practice are based on this acceptable delay, a scaling that fixes the latter probability is natural. This scaling generalizes the QED regime. In concurrent research, Mandelbaum and Zeltyn (2006) investigates a similar scaling for the $M/M/N+G$ model. There exist exact algorithms to determine many of the desired performance measures for the $M/M/N+M$ model (see Whitt 2005a). This implies that the value of any approximations would be in their simplicity of evaluation, their accuracy, and their ability to provide insight (cf. Whitt 2004, p. 1451). We find that our approximations satisfy these criteria in the settings we propose.

Several papers consider how contractual costs and terms affect the relationship between firms and outsourced call centers. Borst et al. (2004) considers how to determine an asymptotically optimal staffing level to minimize staffing and delay costs, or to satisfy a delay constraint. Its discussion of cost minimization is similar to our model of the IB-SLA, while its treatment of constraint satisfaction is similar to our HB-SLA (see Proposition 2(a) below). Borst et al. (2004) notably points to a need to include abandonment and uncertain demand in future work. Harrison and Zeevi (2004) considers how to minimize call center staffing and abandonment costs for a multiple customer class, multiple agent pool model with stochastic, time-varying arrival rates over a finite horizon. Their framework may be interpreted as considering an IB-SLA with penalty terms for abandonment. Bassamboo et al. (2006) extends the framework of Harrison and Zeevi (2004) to include dynamic control in the context of skill-based routing. Akşin et al. (2008) considers the strategic choice of defining a contract type within a call center outsourcing supply chain

and suppresses the operational details we consider. Ren and Zhou (2008) introduces a notion of service quality into the contract. These papers differ from ours in that we explicitly consider how the duration of a period over which a service level is measured affects the staffing. Further, we consider a generic penalty term. It can measure, for example, service delay or abandonment.

As noted, we consider models with abandonments and uncertain arrival rates. Brown et al. (2005) shows that arrival rates for customers are both temporally and stochastically variable and develop autoregressive models to predict demand levels. They also discuss the importance of modeling abandonments in call centers. Chen and Henderson (2001) considers how uncertainty in arrival rates may affect performance measures. Ross (2001) studies how the square root safety staffing rule may be amended to capture uncertain arrival rates. Jongbloed and Koole (2001) proposes an approach based on choosing a likelihood that a service level is met and staffing to that level. Whitt (2006b) considered a fluid approximation for the case of uncertain arrival rates and absenteeism. Steckley et al. (2005) studies the empirical justification for random arrival rates. Similar to our paper, they consider how to measure the fraction of customers satisfactorily served both over a horizon and within a period, and propose an approximation based on the strong law of large numbers. We demonstrate that such an approximation is too crude for the PB-SLA, and therefore develop a more accurate approximation using a central limit theorem (CLT).

The main contributions of this paper are twofold: we develop the tools necessary to analyze the three SLAs, and we draw conclusions regarding how an outsourced call center's staffing may differ under these SLAs. To measure the performance of each of the SLAs for large-scale queueing systems, we develop three approximations. We approximate the probability that a given service delay is achieved, the probability a customer abandons the queue, and the probability of meeting a service level measured by the percentage of customers acceptably served during a period. To this end, we generalize the QED regime for environments with abandonment so that the asymptotic probability of exceeding some acceptable delay, d, is predetermined, and refer to this as a QED(d) regime. This regime justifies a square root safety staffing rule that approximately maintains this probability. We establish sufficient conditions under which a CLT holds for the percentage of customers with acceptable service in a $GI/G/N+G$ queue. We show that these conditions are satisfied in the case of an $M/M/N+M$ queue and approximate the normalizing mean and variance for the CLT.

We consider an outsourced call center that receives revenue for each served call, pays penalties for failing to meet the terms of an SLA, and incurs operating costs. We demonstrate how our approximations may be used to determine the staffing level that maximizes the profit rate of such a call center. We compare how the alternate SLAs perform in cases of known and uncertain arrival rates. We find that the different SLAs lead to markedly different behaviors in the presence of demand rate uncertainty.

The remainder of this paper is organized as follows. In §2, we formally introduce the model and the three SLAs. We then turn to approximating the probability a customer is excessively delayed in §3.1 and the probability a customer abandons the queue in §3.2. We approximate the probability a penalty is incurred in the Period-Based SLA in §4. We present numerical results on the accuracy of the approximations in §5. We then discuss the application of the approximations to determining staffing levels in §6. We draw conclusions in §7.

2. Model

We model a call center as an $M/M/N+M$ queueing system, i.e., we assume Poisson arrivals with demand rate λ, exponentially distributed service with mean $1/\mu$, N servers, and exponentially distributed customer patience with mean $1/\theta$. Let $\rho=\lambda/(N\mu)$. Similar models have been used to model the operations of moderate to large-sized call centers, where there may be 50 to 500 servers and hundreds of calls per minute. We note that the exponentially distributed service is limiting; however, we require it to proceed. (Some notation is given in Table 1; additional notation will be defined as needed.)

Call centers observe demand that has both predictable variability, characterized by time variations in the mean demand level, and stochastic variability expressed by random fluctuations around this mean. We consider how a call center should determine its staffing for the different SLAs for each natural period where there is only stochastic variability. See, e.g., Jennings et al. (1996), Harrison and Zeevi (2004), and Feldman et al. (2008) for work considering predictable variability. We assume alternately that the mean arrival rate, λ, is a fixed constant (referred to as the *Fixed* λ case), or that in each natural period j, the arrival rate, λ_j, is determined as a realization of a time-homogeneous random variable, Λ, with distribution $H_\Lambda(\lambda)$

Table 1. Notation.

λ	arrival rate, Λ is the associated r.v. with distribution $H_\Lambda(\lambda)$
μ	service rate
N	number of servers
ρ	$=\lambda/(N\mu)$
$1/\theta$	average patience
SL	designated service level (for PB- and HB-SLAs)
d	acceptable delay
T	natural period length
$M(T)$	number of arrivals in a period of length T
V_i	virtual waiting time for customer i
X_i	patience of customer i
W_i	waiting time for customer i, $W_i=V_i\wedge X_i$
Y_i	interarrival time of customer i
Z_i	service time of customer i (if served)
$I\{K_i\}$	indicator function, 1 if event K_i occurs, 0 otherwise

(referred to as the *Uncertain* λ case). That is, for the Uncertain λ case, we view arrivals as occurring over an infinite sequence of natural periods, each with an i.i.d. λ drawn from $H_\Lambda(\lambda)$. Such a model of uncertain demand rates is consistent with Brown et al. (2005). We assume that each natural period is independent of the others. This is a reasonable assumption if, for example, each natural period is a particular hour of the day on a number of consecutive (and similar) days. We further assume that the system achieves steady state instantaneously from period to period. Such an assumption is supported by Steckley et al. (2005).

To model abandonments, let V_i, the virtual waiting time for customer i, be the time the customer would have waited if served, and let X_i be his/her patience. Then, the customer abandons if $X_i < V_i$ and is served otherwise. The observed waiting time for customer i is then $W_i = V_i \wedge X_i$.

To motivate the staffing problem, consider an outsourced call center that wishes to maximize its expected profit rate. Suppose that it receives a revenue for each served call, has an operating cost per unit time reflecting its capacity, and pays a penalty for failing to meet a specified service level. Let $Rev(\lambda, N)$ be the expected revenue rate given demand rate λ and N servers, and let $F(N)$ be the cost per unit time of operating with N servers. When λ is a known, fixed value, the expected profit rate is

$$z_\lambda(N) = Rev(\lambda, N) - E[\text{Penalty Rate} \mid N, \lambda] - F(N), \quad (1)$$

where $E[\text{Penalty Rate} \mid N, \lambda]$ is the rate at which penalty costs are incurred. When λ is uncertain, let the expected profit be $z_\Lambda(N) = E[z_\lambda(N)]$.

The focus of this paper is comparing the three SLAs and the staffing levels they induce. To this end, we determine the value of the $E[\text{Penalty Rate}]$ for each SLA. Let K_i be the event that customer i is not acceptably served. We assume that K_i is defined so that a customer is always acceptably served if upon arrival the customer finds fewer than N busy servers and that if all servers are busy there is some probability that the customer is not acceptably served. For the case of excessive delay (the customer waits more than d), $K_i \equiv \{W_i > d\}$; for the case of customer abandonment, $K_i \equiv \{V_i > X_i\}$. Other definitions of K_i are possible such as the event that a customer is excessively delayed given that s/he is served. Clearly, a contract can be written with multiple penalty terms for different measures of service; we consider a single generic penalty in our analysis and specify K_i for our numerical results.

Based on our previous discussion, for Fixed λ, the penalty rate for the IB-SLA is proportional to the likelihood that K_i occurs, i.e.,

$$\begin{aligned} E[\text{Penalty Rate}^{\text{IB, Fixed}}] &= c_p \lambda P(K_i \mid \lambda, N) \\ &= c_p \lambda E[I\{K_i\} \mid \lambda, N], \end{aligned}$$

where c_p is the penalty cost paid by the firm for each customer not acceptably served and $I\{\cdot\}$ is the indicator function. Taking expectations over Λ gives the penalty rate of the Uncertain λ case.

For the PB-SLA, the penalty rate is proportional to the number of arrivals in the period, $M(T)$. A penalty of $c_p M(T)$ is paid if more than $(1 - SL)M(T)$ customers receive unacceptable service, i.e., for Fixed λ,

$$\begin{aligned} &E[\text{Penalty Rate}^{\text{PB, Fixed}}] \\ &= \frac{c_p}{T} E\left[M(T) I\left\{ \sum_{i=1}^{M(T)} I\{K_i\} > (1 - SL)M(T) \right\} \middle| \lambda, N \right] \quad (2) \\ &\approx c_p \lambda P(\text{Penalty}^{\text{PB, Fixed}} \mid \lambda, N, T), \quad (3) \end{aligned}$$

where

$$\begin{aligned} &P(\text{Penalty}^{\text{PB, Fixed}} \mid \lambda, N, T) \\ &\quad \equiv P\left(\sum_{i=1}^{M(T)} I\{K_i\} > (1 - SL)M(T) \mid \lambda, N, T \right). \end{aligned}$$

(We use common notation for the penalty cost, c_p, though in practice its value would be contract dependent.) Again taking expectations over Λ gives the expected penalty rate in the Uncertain λ case. Note that the approximation in (3) expresses the expectation that λT customers will arrive in a period of length T. (Throughout, we use the notation $\approx$ to designate an approximation. We test this approximation and all others in our numerical results—see §§5 and 6.)

For the HB-SLA, for Fixed λ, the law of large numbers implies that a penalty is incurred with probability one if the probability that a customer is not acceptably served exceeds $(1 - SL)$, and none otherwise. Thus, $E[\text{Penalty Rate}^{\text{HB, Fixed}}] = c_p \lambda P(\text{Penalty}^{\text{HB, Fixed}} \mid \lambda, N)$, where

$$P(\text{Penalty}^{\text{HB, Fixed}} \mid \lambda, N) = I\{E[I\{K_i\} \mid \lambda, N] > (1 - SL)\}.$$

Similarly, for the case of Uncertain λ, a penalty is incurred either with probability one or zero. In this case, we show:

Observation 1.

$$\begin{aligned} &P(\text{Penalty}^{\text{HB, Uncertain}} \mid N) \\ &\quad = I\left\{ \frac{E[\lambda T E[I\{K_i\} \mid \lambda, N]]}{E[\Lambda T]} > (1 - SL) \right\} \\ &\quad = I\left\{ \frac{\int_0^\infty \lambda E[I\{K_i\} \mid \lambda, N] \, \mathrm{d}H_\Lambda(\lambda)}{E[\Lambda]} > (1 - SL) \right\}. \end{aligned}$$

All proofs appear in the online companion to this paper that can be found at http://or.journal.informs.org/.

Comment on the Use of Approximations. As we assume Markovian arrivals, service times, and abandonment times, one could compute the exact values of $P(W > d)$ and $P(Ab)$ using standard queueing theory. Therefore, we could solve for the optimal value of N for the IB-SLA and HB-SLA cases without resorting to approximations. However, the probability that a penalty is incurred for the PB-SLA cannot be computed exactly. It can be found

either through approximation (as we do) or through extensive simulation. Our approximation using a CLT requires an estimate of both the expected performance (given by our approximations of $E[I\{K_i\}]$) and its variance (provided by an approximation for the distribution of the waiting time). Further, based on our approximation for $P(W > d)$, we show that a square root safety staffing rule provides an asymptotically optimal solution to the HB-SLA. We next develop these approximations.

3. Approximations for Delay and Abandonment

In this section, we develop heavy-traffic approximations for the probability a customer is delayed in excess of d, $P(W > d \mid \lambda, N)$, and the probability a customer abandons the queue, $P(Ab \mid \lambda, N) = 1 - P(\text{Customer is Served} \mid \lambda, N)$. Throughout, we denote the c.d.f. and the p.d.f. of a standard normal random variable by $\Phi(x)$ and $\phi(x)$, respectively.

3.1. Approximating the Probability of Delay

We develop an asymptotic approximation for $P(W > d \mid \lambda, N)$. We do so by first presenting in Lemma 1 an analytic expression for π_k, where π_k denotes the steady-state probability that there are k customers in an $M/M/N + M$ queue (including in service). We then establish an analytic expression for $P(W > d \mid \lambda, N)$ in Lemma 2. Both of these use Gamma function notation. Recall that the Gamma function, the upper incomplete Gamma function, and the lower incomplete Gamma function are given by $\Gamma(a) \equiv \int_0^\infty e^{-t} t^{a-1}\,dt$, $\Gamma(a, x) \equiv \int_x^\infty e^{-t} t^{a-1}\,dt$, and $\gamma(a, x) \equiv \int_0^x e^{-t} t^{a-1}\,dt$, respectively. Using these lemmas, we state and prove the main theorem giving the nondegenerate asymptotic probability of an unacceptable delay. To do so, we consider a series of $M/M/N + M$ queues with arrival rates λ_N, service rate $\mu_N = \mu$, and patience rate $\theta_N = \theta > 0$ as N increases. Let $\rho_N = \lambda_N/(N\mu_N)$. Throughout, let $c \equiv N\mu/\theta$; this is a simplifying constant in our development. Let W_N be distributed as the steady-state waiting time in such a queue.

Lemma 1. *Denoting by π_k the steady-state probability that an $M/M/N+M$ queueing system has k calls in it, we have*

$$\pi_k = \begin{cases} \dfrac{(\lambda/\mu)^k}{k!}\pi_0, & 0 \leqslant k \leqslant N, \\ \dfrac{(\lambda/\mu)^N}{N!}\pi_0 \displaystyle\prod_{j=N+1}^{k} \frac{\lambda}{N\mu + (j-N)\theta}, & k > N, \end{cases}$$

and

$$\pi_0 = \left(e^{\lambda/\mu}\frac{\Gamma(N+1, \lambda/\mu)}{\Gamma(N+1)} + \frac{(\lambda/\mu)^N}{N!}\left(\frac{\theta}{\lambda}\right)^c \cdot e^{\lambda/\theta}\left(\gamma\left(1+c, \frac{\lambda}{\theta}\right)\right)\right)^{-1}.$$

Next, we express the probability of exceeding a stated delay in analytic form.

Lemma 2. *In an $M/M/N + M$ queue, the steady-state probability of waiting time to exceed $d > 0$ is given by*

$$P(W > d \mid \lambda, N) = \pi_N \frac{N\mu}{\theta} e^{-\theta d}\left(\frac{\theta}{\lambda}\right)^c \gamma\left(c, \lambda\frac{e^{-\theta d}}{\theta}\right) e^{\lambda/\theta}. \quad (4)$$

For the case of $d = 0$, Lemmas 1 and 2 imply:

Corollary 1. *The probability of waiting is given by*

$$\begin{aligned} &P(W > 0 \mid \lambda, N) \\ &= \pi_N \frac{N\mu}{\theta}\left(\frac{\theta}{\lambda}\right)^c \gamma\left(c, \frac{\lambda}{\theta}\right) e^{\lambda/\theta} \\ &= \left(\frac{\lambda}{\mu}\right)^N \\ &\quad \cdot \frac{(N\mu/\theta)(\theta/\lambda)^c \gamma(c, \lambda/\theta) e^{\lambda/\theta}}{e^{\lambda/\mu}\Gamma(N+1, \lambda/\mu) + (\lambda/\mu)^N(\theta/\lambda)^c e^{\lambda/\theta}(\gamma(1+c, \lambda/\theta))}. \end{aligned}$$

We now state our main result for the probability that the waiting time exceeds a stated delay, the proof of which uses Lemmas 1 and 2. For the case of $d = 0$, the result is as in Garnett et al. (2002), and we adopt their notation in this case. Let $h(x) = \phi(x)/(1 - \Phi(x))$ and $w(x, y) = (1 + h(-xy)/yh(x))^{-1}$.

Theorem 1. *In an $M/M/N+M$ queue, the probability of waiting more than $d \geqslant 0$ has a nondegenerate limit, i.e.,*

$$\lim_{N\to\infty} P(W_N > d) = \alpha, \quad 0 < \alpha < e^{-\theta d},$$

if and only if

$$\lim_{N\to\infty} \sqrt{N}(1 - \rho_N e^{-\theta d}) = \beta, \quad -\infty < \beta < \infty. \quad (5)$$

Then,

$$\alpha = \begin{cases} e^{-\theta d}\Phi\left(-\sqrt{\dfrac{\mu}{\theta}}\beta\right) & \text{if } d > 0, \\ w\left(-\beta, \sqrt{\dfrac{\mu}{\theta}}\right) & \text{if } d = 0. \end{cases} \quad (6)$$

We make several observations with regard to the theorem. First, sequences of systems operating under staffing policies that satisfy (5) exhibit properties analogous to those under QED staffing. In particular, the asymptotic probability of unacceptable delay converges to a nondegenerate limit. In contrast to the QED regime as refined by Garnett et al. (2002) where $d = 0$, we allow $d > 0$. To satisfy (5), we require that $\lim_{N\to\infty} \rho_N = e^{\theta d}$. This implies that for nonzero d, ρ_N approaches a limit strictly larger than one so that the probability of waiting approaches one. Moreover, asymptotically the delay for customers in excess of d decreases to zero and the fraction of customers abandoning after waiting d also decreases to zero. Because of

this focus on d, we refer to the scaling regime given by (5) as the QED(d) regime because it generalizes the standard QED regime to the case of $d > 0$ with impatient customers.

Second, because $e^{-\theta d}$ is the probability that a customer has patience in excess of d, $\lambda_N e^{-\theta d}$ is the demand rate of customers with patience larger than d. Therefore, $\lim_{N\to\infty} \lambda_N e^{-\theta d}/(N\mu) = 1$ implies that the limiting system is similar to a call center that only serves customers with patience larger than d. Thus, Theorem 1 implies that as $N \to \infty$, there is negligible effect on the system from customers with little patience, beyond insuring stability. This agrees with the fluid model considered by Whitt (2006a).

Third, recall that the standard QED regime keeps $P(W > 0)$ approximately fixed for large N and implies a square root safety staffing rule. Analogously, the QED(d) regime keeps $P(W > d)$ approximately fixed for large N and implies the following square root safety staffing rule:

$$N \approx Re^{-\theta d} + \beta\sqrt{Re^{-\theta d}}. \tag{7}$$

In §6, we show that the above rule provides the asymptotically optimal solution to the profit maximization problem for the HB-SLA with $K_i \equiv \{W_i > d\}$ for the Fixed λ case (under stated conditions—see Corollary 3).

3.2. Approximating the Probability of Abandonment

We develop an approximation for the probability of abandonment $P(Ab)$, based on an exact formula for it. Because $P(Ab) = \theta E[W] = (\theta/\lambda)E[\text{Queue Length}]$, doing so also provides approximations for the mean waiting time and queue length.

THEOREM 2. *In an $M/M/N + M$ queue, the steady-state probability of abandonment is given by*

$$P(Ab) = \frac{\pi_N}{\rho} + \left(\frac{\rho - 1}{\rho}\right)P(W > 0). \tag{8}$$

Theorem 2 can be used to argue that an asymptotic approximation for the probability of abandonment is given by $\max[0, (\rho - 1)/\rho]$, which is the approximation given in Whitt (2004) for the efficiency-driven (ED) regime. That is, by letting N get large so that the probability π_N is small, and assuming $d > 0$ so that in the limit all customers wait either in our regime or the ED regime, $P(W > 0) \to 1$. (A formal proof follows along these lines by adapting Theorems 2.1 and 2.3 of Whitt 2004 to our regime.) Although this approximation is very clean and simplifies solving for the optimal staffing levels, we find that in numerical testing it is too crude. We propose an approximation based on Theorem 2 that performs well in numerical testing.

Observing that $P(W > 0)$ is approximated in Theorem 1, we require only an approximation for π_N. We do so using a fluid approximation. The wait of a customer with infinite patience that sees exactly N calls in the system upon its arrival is exponentially distributed with an average waiting time of $1/(N\mu)$. Thus, because this customer's waiting time is the virtual waiting time of the system, we approximate

$$\pi_N \approx P(V > 0) - P(V > \tau),$$

with $\tau = 1/(N\mu)$. That is, we approximate π_N by the probability that there is a wait, less the probability that the wait exceeds the expected waiting time of a single waiting customer. Note that $P(V > 0) = P(W > 0)$ and that Garnett et al. (2002) approximate

$$P(V > \tau) \approx P(W > 0)\frac{h(\beta_0\sqrt{\mu/\theta})}{\Psi(\beta_0\sqrt{\mu/\theta}, \tau\sqrt{N\mu\theta})},$$

where $h(x) = \phi(x)/(1 - \Phi(x))$ as above and $\Psi(x, y) = \phi(x)/(1 - \Phi(x + y))$. Substituting in $P(W > 0)$ from (6) and simplifying, we have

$$P(Ab) \approx w(-\beta_0, \sqrt{\mu/\theta}) \cdot \left(1 - \frac{h(\beta_0\sqrt{\mu/\theta})}{\rho\Psi(\beta_0\sqrt{\mu/\theta}, \sqrt{\theta/N\mu})}\right). \tag{9}$$

The approximation is of similar form to that of Garnett et al. (2002). In §5, we compare these approximations.

4. Approximating the Penalty for the PB-SLA

Recall that for the period-based SLA with a Fixed λ, the probability that a penalty is paid in a period of length T when there are $M(T)$ customers is given by

$$P(\text{Penalty}^{\text{PB, Fixed}} \mid \lambda, N, T) = P\left(\sum_{i=1}^{M(T)} \frac{I\{K_i\}}{M(T)} > (1 - SL) \,\middle|\, \lambda, N, T\right). \tag{10}$$

We approximate the probability of penalty by establishing conditions for which a CLT holds for (10). We use the approach of Meyn and Tweedie (1993) on an appropriately defined Markov chain representation of a $GI/G/N + G$ queue. We then establish, using Anscombe's Theorem (Anscombe 1952), that the CLT holds when the number of arrivals within time T is a random variable. Then, we show that the conditions for the CLT hold for the $M/M/N + M$ case. Finally, we approximate the mean and variance of the normal random variable defined by the CLT for this case.

4.1. The Embedded Workload Markov Chain for a $GI/G/N + G$ Queue

Consider the following infinite (and nondenumerable) state space Markov chain representation for the embedded workload process of a $GI/G/N + G$ queue denoted by $\widetilde{\mathbf{U}}$.[1] (Throughout, we denote processes using capital, bold letters and vectors using lower-case, bold letters.) We assume,

without loss of generality, that the service and patience of each customer are known at the time of arrival. Upon arrival, the ith customer observes the workload of each server (i.e., the residual cumulative service time for each server) and joins the queue with the smallest workload if it is less than the customer's patience; otherwise, the customer abandons the system. In case of a tie, the newly arriving job is routed to the station with the smallest index. (This is a formal way of viewing the operation of the system; in practice, customers will in fact wait prior to abandoning the queue.) Let $\widetilde{U}_i^j$ for $i = 0, 1, \ldots$ and $j \in J = \{1, \ldots, N\}$ denote the workload of the jth server immediately before the ith arrival. The smallest index of the least-busy server just before the ith arrival is $j_i^* = \min(\arg\min_{j=1,\ldots,N}(\widetilde{U}_i^j))$. Observe that with this description $V_i = \widetilde{U}_i^{j_i^*}$. (For notational convenience, we omit the subscript i from j_i^* in the sequel.) Denote the interarrival time of the ith customer by Y_i, the patience of the ith customer by X_i, and the service time of the ith customer by Z_i, and let F_Y, F_X, and F_Z be their associated c.d.f.s, respectively. We let $\mathbf{e}_j$ be an $N \times 1$ vector with one in the jth element and zero for all others, $\mathbf{e}$ be the $N \times 1$ unity vector, and $\mathbf{0}$ be the $N \times 1$ zero vector. Then, the evolution of the embedded workload process, assuming that the 0th customer arrives at an empty system, is

$$\widetilde{\mathbf{U}}_0 = (\mathbf{0}),$$

$$\widetilde{\mathbf{U}}_{i+1} = \begin{cases} (\widetilde{\mathbf{U}}_i - Y_{i+1}\mathbf{e})^+ & \text{if } X_i < \widetilde{\mathbf{U}}_i^{j^*}, \\ (\widetilde{\mathbf{U}}_i + Z_i\mathbf{e}_{j^*} - Y_{i+1}\mathbf{e})^+ & \text{if } X_i \geqslant \widetilde{\mathbf{U}}_i^{j^*}, \end{cases}$$

where $(\widetilde{\mathbf{U}}_i)^+$ is an $N \times 1$ vector with nonnegative elements $\max[\widetilde{U}_i^j, 0]$.

Let $\mathbf{U}$ be the (continuous-time) Markov chain with $\widetilde{\mathbf{U}}$ augmented with the patience of the arriving customer, i.e., $\mathbf{U}_i = (\widetilde{\mathbf{U}}_i; X_i)$, and let U be its state space. Note that the events $K_i \equiv I\{W_i > d\}$ and $K_i \equiv I\{V_i > X_i\}$ are defined on $\mathbf{U}$.

Let $\boldsymbol{\eta} = \bigcup_{x\in[0,\infty)}\{\mathbf{0}; x\}$ be the set of states of U when an arrival (with any patience x) finds an empty system. Assume that the $GI/G/N + G$ can reach a steady state (sufficient conditions for this are that both $E[X] < \infty$ and $E[Z] < \infty$). Let p_0 be the limiting probability that an arrival sees an empty system in this queue. Sufficient conditions for $p_0 > 0$ are that both $P(X < Y) > 0$ and $P(Z < Y) > 0$, as is true if Y has support on $[0, \infty)$ (see the discussion following Example 12.2.1 in Asmussen 2003). Observe that once the Markov chain reaches the set $\boldsymbol{\eta}$, its future is independent of its past. Then, assuming $p_0 > 0$, we show that $\mathbf{U}$ is a regenerative process and the corresponding Markov chain is well-behaved. For denumerable state space Markov chains, this is equivalent to saying that the chain is irreducible and recurrent. However, because $\mathbf{U}$ is nondenumerable, we require a notion of irreducibility of a Markov chain on a general set.

We therefore follow the development of Meyn and Tweedie (1993, pp. 89–91) and use the concept of ψ-irreducibility, which implies that there exists a measure ψ such that any set of states A in the chain with positive measure $\psi(A)$ can be reached from every state x in the chain. That such a measure exists for our representation of the embedded workload process is not surprising given that $\mathbf{U}$ is a regenerative process. Formally, let Q be a topological set and let $\mathcal{B}(Q)$ be its Borel σ-field.

Definition 1. A Markov chain $\boldsymbol{\Psi}$ is *ψ-irreducible* if there exists a measure φ on $\mathcal{B}(Q)$ such that for $A \in \mathcal{B}(Q)$, whenever $\varphi(A) > 0$, $P(\boldsymbol{\Psi}$ starting at q ever enters $A) > 0$ for all $q \in Q$. ψ is the maximal such measure, i.e., $\psi(A) = 0$ implies $\varphi(A) = 0$.

We show:

Proposition 1. *If $p_0 > 0$, the Markov chain $\mathbf{U}$ satisfies the following*:

(1) $\mathbf{U}$ *is a regenerative process*; *thus, it has a steady-state distribution, φ and $\varphi(\boldsymbol{\eta}) = p_0$*;

(2) $\mathbf{U}$ *is ψ-irreducible.*

4.2. The Central Limit Theorem for $I\{K\}$

For a Markov chain, $\boldsymbol{\Psi}$, let $S_m(g) \equiv \sum_{i=1}^m g(\Psi_i)$ for any function g on the state space. The CLT for a Markov chain with an invariant probability gives conditions under which $S_m(g)$, when properly centered and normalized, is asymptotically distributed as a standard normal random variable. To apply the theorem to $\mathbf{U}$, let

$$f(\mathbf{U}_i) = I\{K_i\} - E[I\{K_i\} \mid \mathbf{u}_{i-1}]. \tag{11}$$

Then, $|f(\mathbf{U}_i)| \leqslant 1$ and because $E[E[I\{K_i\} \mid \mathbf{u}_{i-1}]] = E[I\{K_i\}]$, we have $E[f(\mathbf{U}_i)] = 0$. Let

$$S_0 = 0,$$
$$S_m = \sum_{i=1}^m f(\mathbf{U}_i), \tag{12}$$

and observe that S_m is the sum of dependent zero mean random variables, and therefore is a zero mean martingale. Let $\tau_\eta = \min\{n \geqslant 1\colon \mathbf{U}_n \in \boldsymbol{\eta} \mid \mathbf{U}_0 \in \boldsymbol{\eta}\}$ be the number of steps for the process to reenter η given that it starts there.

Theorem 3. *If $E[\tau_\eta^2] < \infty$ for a $GI/G/N + G$ queue, then for $f(\mathbf{U}_i)$ defined in* (11) *and S_m defined in* (12), *the CLT for Markov chains holds, i.e.*,

$$\lim_{m\to\infty} P\left(\frac{S_m}{\sqrt{m\nu_f^2}} \leqslant u\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-v^2/2}\,\mathrm{d}v = \Phi(u)$$

for some constant ν_f^2, $0 < \nu_f^2 < \infty$.

Note that ν_f^2 is independent of the initial state of the Markov chain as long as the system empties with positive probability from that state.

For the case where the number of arrivals is a random variable, $M(T)$, we use Anscombe's Theorem (1952) to establish:

Theorem 4. *Let $M(T)$ represent the number of arrivals during period T for a $GI/G/N + G$. Then, if $E[\tau_\eta^2] < \infty$*

for $f(\mathbf{U}_i)$ in (11) *and $S_{M(T)}$ defined analogous to* (12)*, the CLT holds, i.e.,*

$$\lim_{\lambda T\to\infty} P\left(\frac{S_{M(T)}}{\sqrt{E[M(T)]\nu_f^2}} \leqslant u\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-v^2/2}\,\mathrm{d}v = \Phi(u) \tag{13}$$

for some constant ν_f^2, $0 < \nu_f^2 < \infty$.

The key premise to establish for Theorem 4 is $E[\tau_\eta^2] < \infty$. Because this is the case for the $M/M/N+M$ queue, we have:

Corollary 2. *For an $M/M/N+M$ queue, the CLT in Theorem* 4 *holds.*

When $E[\tau_\eta^2] < \infty$ for more general queues, the CLT will follow. Doing so is beyond the scope of this paper.

In addition to the CLT, the strong law of large numbers (c.f. Theorem 17.0.1 of Meyn and Tweedie) also holds, i.e.,

$$\lim_{m\to\infty}\sum_{i=1}^{m}\frac{I\{K_i\}}{m} = \lim_{m\to\infty}\sum_{i=1}^{m}\frac{E[I\{K_i\}\mid \mathbf{u}_{i-1}]}{m} = E[I\{K_i\}]. \tag{14}$$

Using Corollary 2 and (10), we can present our approximation for the probability of a penalty:

$$\begin{aligned}
&P(\text{Penalty}^{\text{PB, Fixed}} \mid \lambda, N, T)\\
&= P\left(\sum_{i=1}^{M(T)}\frac{I\{K_i\}}{M(T)} > (1-SL)\,\middle|\,\lambda, N, T\right)\\
&= P\left(\frac{S_{M(T)}}{M(T)} > (1-SL) - \frac{\sum_{i=1}^{M(T)}E[I\{K_i\}\mid u_{i-1}]}{M(T)}\,\middle|\,\lambda, N, T\right)\\
&= P\left(\frac{S_{M(T)}}{\sqrt{M(T)\nu_f^2}} > \frac{\sqrt{M(T)}}{\sqrt{\nu_f^2}}\right.\\
&\qquad\left.\cdot\left((1-SL) - \frac{\sum_{i=1}^{M(T)}E[I\{K_i\}\mid u_{i-1}]}{M(T)}\,\middle|\,\lambda, N, T\right)\right).
\end{aligned} \tag{15}$$

Then, using (13), (14), and Theorems 2.7 and 3.9 of Billingsley (1999), which imply that we can take the limit as $T\to\infty$ on the right-hand side of the inequality in (15), we approximate

$$P(\text{Penalty}^{\text{PB, Fixed}} \mid \lambda, N, T) \overset{\infty}{\sim} 1 - \Phi\left(\frac{\sqrt{\lambda T}((1-SL) - E[I\{K_i\}])}{\sqrt{\nu_f^2}}\right), \tag{16}$$

where $a(T) \overset{\infty}{\sim} b(T)$ denotes $\lim_{T\to\infty}(a(T)/b(T)) = 1$.

4.3. Approximating the Variance of the CLT

Based on Theorem 17.5.3 of Meyn and Tweedie, for chains with regenerative sets

$$\nu_f^2 = \mathrm{Var}(f(\mathbf{U}_1)) + 2\sum_{i=1}^{\infty}\mathrm{Cov}(f(\mathbf{U}_1), f(\mathbf{U}_{i+1})),$$

where the variance is given under the invariant probability of $\mathbf{U}$, i.e., $\mathbf{U}_1$ is redefined now as a steady-state realization of $\mathbf{U}$. This implies, in our case,

$$\begin{aligned}
\nu_f^2 &= \mathrm{Var}[I\{K_1\} - E[I\{K_1\}\mid \mathbf{u}_0]]\\
&\quad + 2\sum_{i=1}^{\infty}\mathrm{Cov}\big(I\{K_1\} - E[I\{K_1\}\mid \mathbf{u}_0],\\
&\qquad\qquad I\{K_{i+1}\} - E[I\{K_{i+1}\}\mid \mathbf{u}_i]\big)\\
&= \mathrm{Var}(I\{K_1\}) + 2\sum_{i=1}^{\infty}\mathrm{Cov}(I\{K_1\}, I\{K_{i+1}\}),
\end{aligned}$$

where the equality follows because $E[I\{K_i\}\mid \mathbf{u}_{i-1}]$ is not random.

Because of the Markovian structure of $\mathbf{U}$, $I\{K_1\}$ and $I\{K_{i+1}\}$ are related only through the embedded workload process. Let $\varrho_i \equiv \text{Correlation}(V_1, V_{i+1})$. We assume that the covariance, $\mathrm{Cov}(I\{K_1\}, I\{K_{i+1}\})$, decays geometrically through ϱ_i.[2] That is, we assume that for $i \geqslant 1$, $\varrho_i = (\varrho_1)^i$ and $\mathrm{Cov}(I\{K_1\}, I\{K_{i+1}\}) = \mathrm{Cov}(I\{K_1\}, I\{K_2\})(\varrho_1)^{i-1}$. Thus, we suggest the approximation

$$\begin{aligned}
\nu_f^2 \approx \hat{\nu}_f^2 &\equiv \mathrm{Var}(I\{K_1\}) + 2\,\mathrm{Cov}(I\{K_1\}, I\{K_2\})\sum_{i=0}^{\infty}(\varrho_1)^i\\
&= E[I\{K_1\}](1 - E[I\{K_1\}])\\
&\quad + \frac{2\,\mathrm{Cov}(I\{K_1\}, I\{K_2\})}{1-\varrho_1}.
\end{aligned} \tag{17}$$

For example, when $K_i = \{W_i > d\}$, finding $\hat{\nu}_f^2$ requires $P(W > d)$ (approximated using Theorem 1), $\mathrm{Cov}(I\{W_1 > d\}, I\{W_2 > d\})$, and ϱ_1, the correlation of V_1 and V_2.

To calculate $\mathrm{Cov}(I\{K_1\}, I\{K_2\})$ and ϱ_1, we require the distribution of V. We obtain an approximation for it by first observing that both V_1 and V_2 follow the same distribution. Thus, they are identically distributed but not independent. Therefore,

$$\varrho_1 = \frac{\mathrm{Cov}(V_1, V_2)}{\sqrt{\mathrm{Var}(V_1)\,\mathrm{Var}(V_2)}} = \frac{\mathrm{Cov}(V_1, V_2)}{\mathrm{Var}(V_1)}. \tag{18}$$

We approximate the c.d.f. of V using Theorem 1. Because V_i and X_i are independent random variables, $P(W > w) = P(X > w)P(V > w)$, and because X_i is exponentially distributed, the c.d.f. of V is given as

$$\begin{aligned}
P(V < v) &= 1 - P(W > v)/P(X > v)\\
&\approx 1 - e^{-\theta v}\Phi\left(-\sqrt{\frac{N\mu}{\theta}}(1-\rho_N e^{-\theta v})\right)\Big/ P(X > v)\\
&= \Phi\left(\sqrt{\frac{N\mu}{\theta}}(1-\rho_N e^{-\theta v})\right),
\end{aligned} \tag{19}$$

Table 2. Exact values and error of the approximations for $P(W > d)$ with $\lambda = 100$, and the mean absolute deviation (MAD) of each for $N = 80$ to 120 and $N = 80$ to 100.

		ρ	1.250	1.176	1.111	1.053	1.000	0.952	0.909	0.870	0.833	MAD	
d	θ	N	80	85	90	95	100	105	110	115	120	80–120	80–100
0.1	0.5	Exact	0.9406	0.8914	0.7501	0.5156	0.2775	0.1173	0.0398	0.0110	0.0025		
		Approx	0.0026	0.0026	−0.0105	−0.0332	−0.0442	−0.0351	−0.0184	−0.0069	−0.0019	0.0187	0.0176
		GMR	0.0052	0.0136	0.0118	−0.0028	−0.0103	−0.0068	−0.0020	0.0000	0.0003	0.0063	0.0091
	1.0	Exact	0.7938	0.6622	0.4834	0.2997	0.1548	0.0659	0.0231	0.0066	0.0016		
		Approx	0.0020	−0.0071	−0.0126	−0.0087	−0.0004	0.0049	0.0053	0.0034	0.0016	0.0055	0.0071
		GMR	0.0298	0.0251	0.0070	−0.0082	−0.0112	−0.0069	−0.0026	−0.0005	0.0000	0.0096	0.0153
	2.0	Exact	0.4705	0.3404	0.2211	0.1269	0.0633	0.0270	0.0096	0.0029	0.0007		
		Approx	−0.0130	−0.0087	0.0019	0.0126	0.0185	0.0183	0.0141	0.0090	0.0050	0.0115	0.0101
		GMR	0.0376	0.0129	−0.0051	−0.0117	−0.0100	−0.0056	−0.0023	−0.0007	−0.0001	0.0083	0.0132
1/3	0.5	Exact	0.6541	0.4142	0.1838	0.0553	0.0115	0.0018	0.0002	0.0000	0.0000		
		Approx	−0.0033	−0.0092	−0.0039	0.0010	0.0011	0.0004	0.0001	0.0000	0.0000	0.0022	0.0041
		GMR	0.0689	0.0525	0.0109	−0.0039	−0.0024	−0.0006	−0.0001	0.0000	0.0000	0.0137	0.0263
	1.0	Exact	0.1262	0.0484	0.0145	0.0034	0.0006	0.0001	0.0000	0.0000	0.0000		
		Approx	−0.0006	0.0045	0.0045	0.0025	0.0010	0.0003	0.0001	0.0000	0.0000	0.0016	0.0030
		GMR	0.0372	0.0047	−0.0020	−0.0012	−0.0003	−0.0001	0.0000	0.0000	0.0000	0.0032	0.0062

where, given any $v > 0$, we have used our approximation for $P(W > v)$ given in Theorem 1 with $\rho_N = \lambda/(N\mu)$. Observe that as $v \to \infty$, (19) does not go to one. Thus, it is not a c.d.f., and so we normalize it by dividing by $\Phi(\sqrt{N\mu/\theta})$. (Observe that the normalization factor is effectively one for $N\mu/\theta$ larger than 10.) This gives the following approximation for the c.d.f. for V:

$$F_V(v) \approx \Phi\left(\sqrt{\frac{N\mu}{\theta}}(1-\rho_N e^{-\theta v})\right) \Big/ \Phi(\sqrt{N\mu/\theta}),$$

and differentiating with respect to v implies

$$f_V(v) \approx \frac{\lambda e^{-\theta v}}{\Phi(\sqrt{N\mu/\theta})}\sqrt{\frac{\theta}{N\mu}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(N\mu-\lambda e^{-\theta v})^2}{2N\mu\theta}\right) \quad \text{for } v > 0, \quad (20)$$

and

$$P(V=0) = P(W=0) \approx \Phi\left(\sqrt{\frac{N\mu}{\theta}}(1-\rho_N)\right) \Big/ \Phi(\sqrt{N\mu/\theta}). \quad (21)$$

(Alternate approximations for the c.d.f. can be found using the methods of Puhalskii 1994—see Garnett et al. 2002, p. 217 and Whitt 2004, p. 1457.)

Using this distribution, the mean and variance of V can be calculated using numerical integration. The calculations of our approximations for $\mathrm{Cov}(I\{K_1\}, I\{K_2\})$ and $\varrho_1 \equiv \mathrm{Correlation}(V_1, V_2)$ are given in Online Appendix B.

5. Accuracy of Approximations

Through numerical experiments, we study the three approximations, $P(W > d)$, $P(Ab)$, and $E[\text{Penalty Rate}^{\text{PB, Fixed}}]$. These approximations drive the staffing results. Thus, good performance of each should lead to good performance in determining the staffing. We solve for $P(W > d)$ for a fixed d and N using Theorem 1. That is, given λ, μ, θ , d, and N, we approximate $\beta \approx \sqrt{N}(1-\rho e^{-\theta d})$ based on (5). Substituting β in (6) provides $P(W > d) = \alpha$. To approximate $P(Ab)$, we use the formula given in (9). We approximate the value of $E[\text{Penalty Rate}^{\text{PB}}]$ using (3) where we determine $P(\text{Penalty}^{\text{PB, Fixed}} \mid N, \lambda, T)$ using (16). For this, $E[I\{K_i\}]$ is given either by $P(W > d)$ or $P(Ab)$ in our examples, and ν_f^2 is approximated by $\hat{\nu}_f^2$ in (17).

In our test cases, we let $\mu = 1$; $\lambda = 100$ or 200; $\theta = 0.5$, 1.0, or 2.0; $d = 0.1$ or 1/3, i.e, acceptable delays of 6 seconds and 20 seconds. We vary ρ from 0.833 to 1.25 by varying N from 80 to 120.

Probability of Acceptable Delay, $P(W \geqslant d)$. In Table 2, we present the exact value of the probability of acceptable delay and the error of our approximation defined as (*exact* − *approximation*), denoted by Approx. We also present the mean absolute deviation (MAD) calculated for $N = 80$ to 120 and for $N = 80$ to 100 (the latter are cases where delays are expected). We compare our results to the error of the approximations provided by Garnett et al. (2002) (GMR). We find that our approximation is very accurate (with absolute errors less than 1%) for typical service levels when some abandonment is expected. In particular, we find nearly exact performance for $d = 1/3$ and $\rho > 1$. Because our asymptotic approximation is given for a regime where almost all customers wait, the strong performance here is as expected. That is, from Theorem 1, our approximation is expected to do well when $\rho > 1$ for $d > 0$. In Table 2, this corresponds to smaller N. In contrast, the GMR approximation is based on the scaling where the waiting time is asymptotically zero. This would be the case for smaller ρ or larger N. (We omit the case of $d = 1/3$

Table 3. Exact values and error of the approximations for $P(Ab)$ with $\lambda = 100$, and the mean absolute deviation (MAD) of each for $N = 80$ to 120 and $N = 80$ to 100.

	ρ	1.250	1.176	1.111	1.053	1.000	0.952	0.909	0.870	0.833	MAD	
θ	N	80	85	90	95	100	105	110	115	120	80–120	80–100
0.5	Exact	0.2001	0.1506	0.1034	0.0627	0.0330	0.0151	0.0060	0.0021	0.0006		
	Approx	−9.14E−06	−1.21E−05	5.45E−05	1.33E−04	−8.25E−06	−3.36E−04	−5.54E−04	−5.35E−04	−3.79E−04	2.34E−04	5.32E−05
	GMR	1.87E−02	9.18E−03	2.54E−03	−8.24E−04	−1.39E−03	−7.57E−04	−1.56E−04	9.04E−05	1.12E−04	3.07E−03	5.61E−03
1.0	Exact	0.2007	0.1526	0.1079	0.0695	0.0399	0.0200	0.0087	0.0032	0.0010		
	Approx	1.18E−04	3.12E−04	4.31E−04	2.89E−04	−3.32E−05	−2.95E−04	−3.77E−04	−3.24E−04	−2.23E−04	2.82E−04	2.71E−04
	GMR	1.62E−02	6.70E−03	5.77E−04	−2.17E−03	−2.34E−03	−1.35E−03	−4.01E−04	6.37E−05	1.55E−04	2.77E−03	4.75E−03
2.0	Exact	0.2027	0.1565	0.1138	0.0766	0.0467	0.0252	0.0118	0.0047	0.0016		
	Approx	1.08E−03	1.27E−03	1.00E−03	4.16E−04	−9.96E−05	−2.85E−04	−2.11E−04	−8.98E−05	−2.97E−05	4.95E−04	8.06E−04
	GMR	1.15E−02	2.81E−03	−2.35E−03	−4.25E−03	−3.79E−03	−2.27E−03	−8.39E−04	−4.81E−05	1.80E−04	2.70E−03	4.16E−03

and $\theta = 2.0$ as $P(W > d) \approx 0$ in all cases.) Similar results hold for $\lambda = 200$.

Probability of Abandonment, $P(Ab)$. For $P(Ab)$, we find that our approximation is excellent, with absolute errors less than 0.1% (see Table 3). The approximation is generally an order of magnitude better than that of Garnett et al. (2002). We believe the improved performance stems from approximating the exact abandonment probability given in Theorem 2.

Penalty Approximation for Excessive Delay. To determine the accuracy of the penalty approximation for the PB-SLA, we compare the expected penalty rate $E[\text{Penalty Rate}^{\text{PB, Fixed}}]$ given by simulating (2) to the approximation given by (3), where $P(\text{Penalty})$ is given either through the simulation or by the CLT approximation in (16). In Table 4, we present the results where the penalty is based on excessive delay, $K_i = \{W_i > d\}$. For each case, we simulated 10,000 periods of length T. We let $\lambda = 100$, $\mu = 1$, $\theta = 1$, $d = 1/3$, $SL = 80\%$, and $c_p = 1$. Letting $T = 20$, we vary N from 75 to 85, observing that the simulated penalty rate decreases from approximately 74% to less than 1%. Similarly, letting $N = 78$, we vary T from 10 to 1,000, observing that the penalty rate decreases from 36% to about 1%. We observe that the simulated and approximate penalty rate using the simulated $P(\text{Penalty})$ are very close. However, we observe moderate absolute and relative errors between the simulated penalty rate and that using the CLT approximation in (16). For example, when about 20% of the periods are penalized, we observe about 19% relative error when varying N and an 8% relative error when varying T. Thus, the error of our approximation for $E[\text{Penalty Rate}^{\text{PB}}]$ is primarily a result of the approximation of $P(W > d)$ and the CLT, rather than the approximation in (3), i.e., letting the penalty rate be given by $c_p \lambda P(\text{Penalty})$. We observe that as the period length increases, the penalty rate approaches zero for $N = 78$. For $N = 78$, the steady-state probability of excessive delay is less than $1 - SL$. This is true also for $N = 77$ though, in this case $P(W_i > d) \approx 1 - SL$, so that the penalty rate approaches zero at a slower rate as T increases. We note that $N = 77$ is the solution for the HB case and is given by (7) (see Corollary 3 below). For $N < 77$, the probability of a penalty approaches 100%. That is, as T increases, the probability of a penalty approaches either zero or one. In the next section, we show that for the Fixed λ case, this implies that the staffing for the PB case approaches that

Table 4. The $E[\text{Penalty Rate}^{\text{PB}}]$ for excessive delay ($K_i = \{W_i > d\}$) given by simulation, $\lambda \times$ simulated $P(\text{Penalty})$, and $\lambda \times$ CLT approximation for $\lambda = 100$, $SL = 0.8$ and (a) $T = 20$, varying N, and for (b) $N = 78$, varying T, displaying the error and relative error.

(a)

N	Sim.	$\lambda \times$ sim. $P(\text{penalty})$	$\lambda \times$ CLT approx.	Error	Rel. error (%)
75	73.41	73.00	68.31	−4.69	−6.42
76	60.12	59.67	58.29	−1.38	−2.32
77	47.15	46.70	47.68	0.98	2.09
78	34.47	34.04	37.06	3.02	8.86
79	23.00	22.65	27.03	4.38	19.33
80	14.62	14.37	18.19	3.82	26.58
81	8.30	8.12	11.03	2.91	35.81
82	4.63	4.52	5.83	1.31	28.93
83	2.18	2.12	2.56	0.44	20.96
84	1.10	1.07	0.88	−0.19	−17.54
85	0.33	0.32	0.22	−0.10	−31.97

(b)

T	Sim.	$\lambda \times$ sim. $P(\text{penalty})$	$\lambda \times$ CLT approx.	Error	Rel. error (%)
10	35.94	35.34	40.76	5.42	15.35
15	34.79	34.29	38.74	4.45	12.97
20	34.47	34.04	37.06	3.02	8.86
40	29.89	29.59	32.02	2.43	8.20
50	28.29	28.03	30.07	2.04	7.28
75	24.91	24.71	26.12	1.41	5.69
100	21.46	21.30	23.00	1.70	7.99
200	14.43	14.34	14.81	0.47	3.25
500	4.66	4.64	4.93	0.29	6.19
1,000	1.02	1.02	0.97	−0.05	−4.51

Table 5. The $E[\text{Penalty Rate}^{\text{PB}}]$ for abandonment ($K_i = \{V_i > X_i\}$) given by simulation, $\lambda \times$ simulated $P(\text{Penalty})$, and $\lambda \times$ CLT approximation for $\lambda = 100$, $SL = 0.95$ and (a) $T = 20$, varying N, and for (b) $N = 99$, varying T, displaying the error and relative error.

(a)

N	Sim.	$\lambda \times$ sim. $P(\text{penalty})$	$\lambda \times$ CLT approx.	Error	Rel. error (%)
90	99.54	99.56	99.67	0.11	0.11
91	98.72	98.68	99.28	0.60	0.61
92	97.20	97.13	98.48	1.35	1.39
93	94.56	94.41	96.89	2.48	2.63
94	89.75	89.49	93.91	4.42	4.94
95	81.73	81.37	88.68	7.31	8.98
96	71.84	71.38	80.20	8.82	12.35
97	60.96	60.39	67.80	7.41	12.26
98	47.22	46.67	51.87	5.20	11.14
99	35.65	35.12	34.47	−0.65	−1.85
100	25.22	24.77	18.92	−5.85	−23.64

(b)

T	Sim.	$\lambda \times$ sim. $P(\text{penalty})$	$\lambda \times$ CLT approx.	Error	Rel. error (%)
10	37.54	36.81	38.87	2.06	5.61
15	36.57	35.96	36.46	0.50	1.40
20	35.65	35.12	34.47	−0.65	−1.85
40	32.00	31.64	28.60	−3.04	−9.61
50	30.36	30.05	26.37	−3.68	−12.24
75	27.61	27.37	21.95	−5.42	−19.80
100	25.17	24.97	18.58	−6.39	−25.60
200	17.79	17.67	10.32	−7.35	−41.62
500	8.36	8.32	2.28	−6.04	−72.54
1,000	2.42	2.41	0.24	−2.17	−90.22

of the HB case. We contrast this with the results of the Uncertain λ case.

Penalty Approximation for Abandonment. In Table 5, we present the results for the case of excessive abandonment, $K_i = \{V_i > X_i\}$. We let $SL = 95\%$, i.e., a penalty is incurred if more than 5% of arrivals abandon. We keep the other parameters as above. Varying N from 90 to 100 for $T = 20$, and T from 10 to 1,000, holding $N = 99$, we observe that the CLT approximation again has moderate absolute and relative errors. The relative error grows large as T increases, indicating that the approximation of the PB case approaches the HB case faster than the simulated process, with small absolute errors leading to large relative errors. (As above, we chose $N = 99$ so that the steady-state probability of abandonment is just less than $1 - SL$.)

6. Determining Optimal Staffing Levels

We next apply the approximations to the problem of determining the staffing level for an outsourced call center to maximize (1). We address two questions: How accurate are the staffing decisions made using these approximations? How do the staffing policies differ for the three SLAs?

Let $z_\lambda^{\text{IB}}(N)$ be the profit rate function (1) for a fixed value of λ for the IB-SLA and let N_λ^{IB} be the associated optimal staffing level. Similarly, let $z_\Lambda^{\text{IB}}(N)$ and N_Λ^{IB} denote these for the Uncertain λ case. Also, define similar zs and Ns for the PB- and HB-SLAs. Let the gross profit rate (not including penalties) be $\Pi(N) = E[Rev(\lambda, N)] - F(N)$. To aid in determining the optimal staffing values, we establish:

PROPOSITION 2. *Let N_λ^* be the smallest N such that $E[I\{K_i\}|\lambda, N] < (1 - SL)$.*

(a) *If $\Pi(N)$ is decreasing for $N > N_\lambda^*$, there exists $\hat{c}_p$ such that $N_\lambda^{\text{HB}} = N_\lambda^*$ for $c_p \geqslant \hat{c}_p$. If $\Pi(N)$ is decreasing or convex for $N \geqslant 0$, $\hat{c}_p = -z_\lambda^{\text{HB}}(N_\lambda^*)/\lambda$.*

(b) *For $N \geqslant N_\lambda^*$, $z_\lambda^{\text{HB}}(N) > z_\lambda^{\text{IB}}(N)$ and $z_\lambda^{\text{HB}}(N) > z_\lambda^{\text{PB}}(N)$.*

(c) *Let N_Λ^* be the smallest N such that $\int_0^\infty \lambda E[I\{K_i\} \mid \lambda, N]\, \mathrm{d}H_\Lambda(\lambda) < (1 - SL)E[\Lambda]$. Then, analogous results to* (a) *and* (b) *hold for the Uncertain λ case.*

As a direct consequence, we have:

COROLLARY 3. *The square root safety staffing rule, $N \approx Re^{-\theta d} + \beta\sqrt{Re^{-\theta d}}$, provides the optimal staffing for the HB-SLA for the case of excessive delay and Fixed λ under the conditions of Proposition* 2(a).

The optimal HB-SLA staffing policy is defined through constraint satisfaction (see, e.g., Borst et al. 2004). That is, assuming the regularity conditions on the revenue and operating costs, the HB-SLA will staff exactly the number of servers to just satisfy the service-level constraint. In this case, one would expect $\Pi(N)$ to be decreasing in N near N^*. Because this might lead to negative profit for the call center, we allow for transfer payments to the call center. (Ren and Zhou 2008 consider such payments in their decentralized model.) The value of $\hat{c}_p$ given for $\Pi(N)$ decreasing or convex implies that if the penalty exceeds the average loss per customer, the constraint is binding. Part (b) of Proposition 2 provides upper bounds on the profit rate in the IB- and PB-SLA cases. These bounds restrict the search range for the optimal N for each.

We also observe in our numerical experiments that $N_\lambda^{\text{PB}} > N_\lambda^{\text{HB}}$ (under common cost parameters). An argument supporting this observation relies on considering how the penalty in the PB case given in (16) behaves as the natural period length, T, increases. Observe that this penalty given by the CLT approximation approaches the step function penalty given by the law of large numbers. Therefore, the PB penalty approaches the HB penalty from below for $N > N_\lambda^{\text{HB}}$. This implies that more servers are hired in the PB case. A formal argument would require understanding how ν_f^2, the variance constant in the CLT, changes with N, which is beyond the scope of this paper.

Test Cases. We investigate the accuracy of the staffing decisions for the SLAs for Fixed and Uncertain λ.

In our test cases, we assume that the revenue is linear in the number of customers served, i.e., $Rev(\lambda, N) = r\lambda P(\text{Customer is Served} \mid \lambda, N)$. Also we assume that the capacity cost is linear in the number of servers, i.e., $F(N) = c_o N$. In general, the objective functions for the PB- and HB-SLA cases are not necessarily convex. We can show that if $P(Ab)$ is decreasing and convex in N, then $\Pi(N)$ is decreasing if $r < c_o$, so that for a sufficiently large penalty, Proposition 2(a) holds. It is easy to show that $P(Ab)$ is decreasing in N; Armony et al. (2009) indicates that it is also convex $\theta \leqslant \mu$. (Koole 2005 discusses the convexity of the service level measured by excessive delay and finds a similar result.)

We consider a base case (Case 1) with $\mu = 1$, $\theta = 1$, $SL = 80\%$, $d = 1/3$, and natural period length $T = 20$. We consider four additional cases varying one parameter at a time as follows: Case 2: $SL = 90\%$; Case 3: $d = 0.1$; Case 4: $\theta = 0.5$; and Case 5: $T = 100$. We consider two distributions for the demand rate, $H(\Lambda)$: $H(\Lambda) \sim$ a discretized normal c.d.f. with mean 100 and standard deviation 6.66 truncated on the integers between 80 and 120, and $H(\Lambda) \sim$ a discretized normal c.d.f. with mean 200 and standard deviation 6.66 truncated on the integers between 180 and 220. In each of the five cases, we let the revenue rate be $r = 0$ and $r = 0.9$. We let $c_p = c_o = 1.0$. Alternate costs give qualitatively similar results. Throughout, we assume that the penalty is based on excessive delay ($K_i = \{W_i > d\}$). Note that because the firm only receives revenue for served calls, an IB-SLA penalty of the form $\lambda E[I\{V_i > X_i\} \mid N, \lambda]$ is implicitly imposed in all three SLA's profit functions as well.

Comparing Approximate and Exact/Simulated Staffing Levels. For the case of Fixed λ, N_λ^{HB} is given by the square root staffing rule (7). The value of N is found letting $\alpha = (1 - SL)$ and $\beta = -\sqrt{\theta/\mu}\Phi^{-1}(\alpha \exp(\theta d))$. By Proposition 2(c), N_Λ^{HB} is found by searching for the smallest N such that $\int_0^\infty \lambda E[I\{K_i\} \mid \lambda, N] \mathrm{d}H_\Lambda(\lambda) < (1 - SL)E[\Lambda]$, e.g., using a bisection search. By Proposition 2(b), the optimal value for the IB-SLA for the Fixed λ case is found by searching on $[0, \hat{N}]$, where $\hat{N}$ is the smallest $N > N_\lambda^{\text{HB}}$ such that $z_\lambda^{\text{HB}}(N) \leqslant z_\lambda^{\text{IB}}(N_\lambda^{\text{HB}})$. Analogous results hold for the Uncertain λ case and for the PB-SLA for both Fixed and Uncertain λ.

In Table 6, we present the optimal staffing levels. We also present the error of the staffing with respect to the exact optimum for the IB- and HB-SLAs (e.g., $N^{\text{IB}} - N^{\text{Exact}}$), and the error with respect to a simulated optimum for the PB-SLA. (We simulated 10,000 periods of length T varying N in the neighborhood of N_Λ^{PB}; we present the error based on the profit maximizing N.) We also tested the PB-SLA

Table 6. Optimal staffing levels, deviation from optimal (exact or simulated) for Fixed and Uncertain λ. Also, mean and standard deviation of simulated error from optimal for PB-SLA in the Fixed λ case.

			Fixed λ								Uncertain λ					
			N^*			Error			Mean	Std. dev.	N^*			Difference		
Case	$E[\Lambda]$	r	IB	PB	HB	IB	PB	HB	error	error	IB	PB	HB	IB	PB	HB
1	100	0	85	84	77	0	0	−1	−0.273	0.647	86	89	79	−1	0	0
		0.9	92	86	77	1	0	−1	−1.091	0.701	93	92	79	0	0	0
	200	0	165	162	151	0	0	−1	−0.091	0.539	166	166	152	0	0	0
		0.9	177	164	151	1	−1	−1	−1.182	1.079	179	171	152	1	0	0
2	100	0	85	89	82	0	0	0	−0.455	0.82	86	94	84	−1	1	0
		0.9	92	90	82	1	−2	0	−0.818	1.328	93	96	84	0	0	0
	200	0	165	169	157	0	0	−1	−0.2	0.422	166	173	158	0	0	−1
		0.9	177	171	157	1	−2	−1	−2.182	0.874	179	178	158	1	0	−1
3	100	0	106	104	99	0	−1	1	−0.545	1.44	108	111	100	0	0	−1
		0.9	108	104	99	2	−1	1	−0.545	1.44	110	112	100	0	1	−1
	200	0	206	199	192	0	2	−1	0.182	1.834	208	207	193	0	−3	−1
		0.9	208	200	192	2	3	−1	0.091	2.3	210	208	193	0	−2	−1
4	100	0	97	95	90	−1	1	0	−1.182	1.25	99	102	92	−1	0	0
		0.9	99	96	90	0	2	0	−0.909	1.446	102	103	92	0	−1	0
	200	0	188	186	177	0	−1	0	−2.182	0.603	191	192	178	1	0	−1
		0.9	193	187	177	0	−4	0	−3	1.414	195	195	178	0	0	−1
5	100	0	85	81	77	0	0	−1	−0.091	0.302	86	88	79	−1	0	0
		0.9	92	82	77	1	0	−1	0.091	0.539	93	90	79	0	0	0
	200	0	165	157	151	0	1	−1	0.3	0.483	166	163	152	0	−1	0
		0.9	177	158	151	1	0	−1	0.182	0.405	179	167	152	1	0	0
Average						0.45	−0.15	−0.5	−0.695	0.993				0	−0.25	−0.4
Exact						11	8	6						12	14	12
Off by 1						7	6	14						8	4	8
Off by >1						2	6	0						0	2	0

for each λ from 95 to 105, and from 195 to 205. We present the mean and standard deviation of the error for these cases in the appropriate rows in the table. These may be used to estimate the accuracy of the PB-staffing for the Fixed λ case.

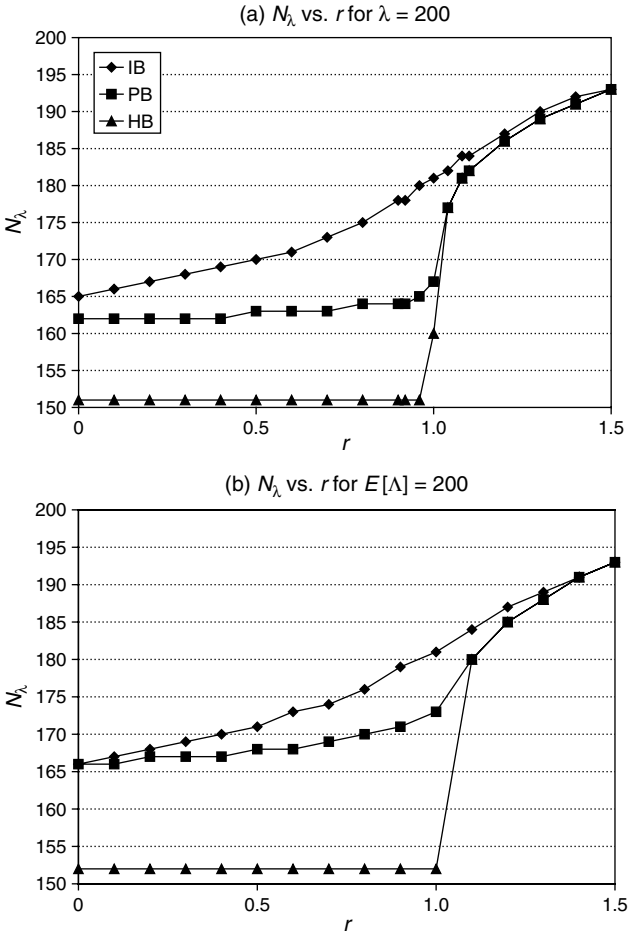
We observe that in the vast majority of cases, the staffing level given by the approximations is either exact or within one of the optimum. There is a slight bias to overstaff with the IB-SLA approximation and to understaff for the PB- and HB-SLA approximations. The greatest deviations occur for the PB-SLA with $E[\Lambda] = 200$ under Case 3 ($d = 0.1$), and Case 4 ($\theta = 0.5$) for Fixed $\lambda = 200$. We note that the profit function is flat in the neighborhood of the optimal N so that even an error of 2% or 3% in N will result in a small change in profit. In general, the results are in keeping with the observations of Borst et al. (2004) that such approximate staffing levels are very good. The approximations for the probability of abandonment and the penalty rate introduce additional error in our results. Based on this accuracy, we turn next to deriving insights based on the experimental outcomes.

Comparing the Fixed and Uncertain λ Cases. We now use the approximations to compare the staffing levels for the cases of Fixed and Uncertain λ for the three SLAs while changing the revenue rate, r, and the period length, T. In Figure 1, we present the staffing levels for the base case with Fixed $\lambda = 200$ and Uncertain λ with $E[\Lambda] = 200$. When $r > c_o = 1$, the value of N_λ is sensitive to changes in r for all three SLAs. Here, the call center trades off capacity cost with additional revenue resulting from less abandonment; the penalty plays a minor role in determining the staffing level. For $r < c_o = 1$, N_λ^{HB} and N_Λ^{HB} are determined as in Proposition 2(a), satisfying the constraint to achieve the service level. Observe that, similarly, N_λ^{PB} and N_Λ^{PB} are generally robust to changes in r for $r < 1$. In comparison, the IB-SLA responds to the changes in r throughout the range. Thus, for the PB-SLA, when the penalty cost is relevant, the staffing level is primarily determined by the trade-off between the penalty and the cost of capacity.

Figure 2(a) presents the staffing levels with $r = 0$ and Fixed $\lambda = 200$ for the base case and the case with $SL = 90\%$ as T changes from 10 to 1,000. Figure 2(b) presents the staffing levels for the Uncertain λ case with $E[\Lambda] = 200$. Letting $r = 0$ represents a call center with a fixed revenue that must trade off potential penalties with operating cost. We observe that N_λ^{PB} and N_Λ^{PB} are responsive to changes in T over the whole range. As T increases, N_λ^{PB} approaches N_λ^{HB}. This is a consequence of the CLT in (16) as we discussed above. The responsiveness of the staffing with respect to changes in T highlights the importance of using the CLT rather than the strong law of large numbers for evaluating the PB-SLA.

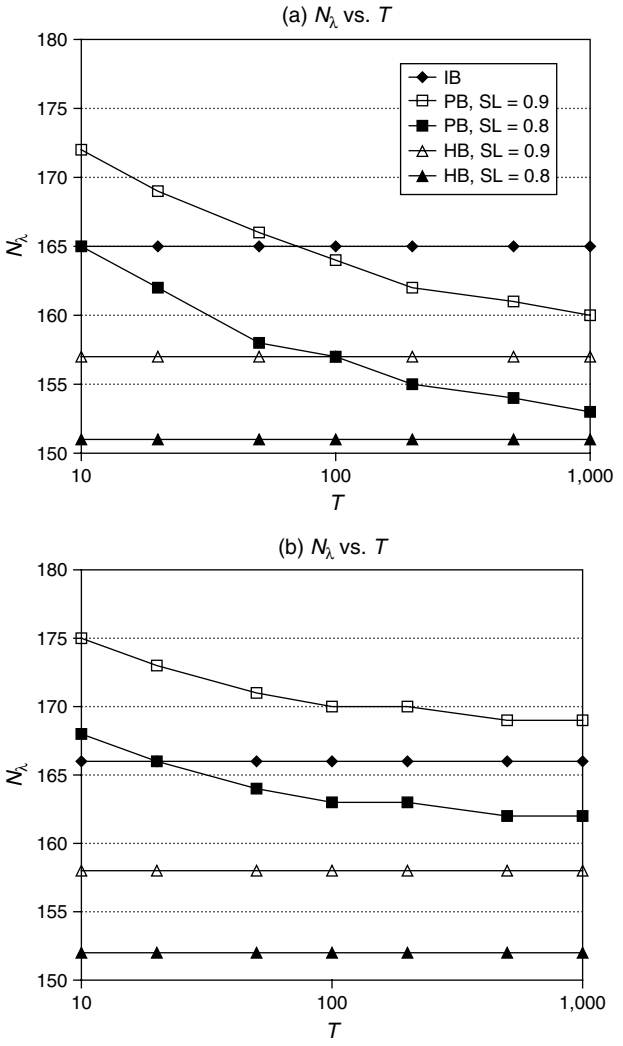
We observe that for Uncertain λ, the staffing of the PB-SLA does not approach that of the HB-SLA. Rather, although the solution to the IB- and HB-SLA's Uncertain λ case is only slightly higher than that for the Fixed λ case, we observe that N_Λ^{PB} is much higher than N_λ^{PB}. The result leads to our conclusion that the PB-SLA responds to Uncertain λ in a more conservative way than either the IB-SLA or HB-SLA. We next measure this difference and then provide an explanation.

Figure 1. Staffing as revenue rate r changes for (a) N_λ for Fixed $\lambda = 200$ and (b) N_Λ for Uncertain $E[\Lambda] = 200$.



We measure how each SLA responds to Uncertain versus Fixed λ, using the cases in Table 6, and present the results in Table 7. To do so, we find values for the acceptable delay, d^{IB} and d^{HB}, for the IB-SLA and HB-SLA cases, respectively, that equate the number of servers with that of the PB-SLA for the Fixed λ case. (This ensures that all three SLAs provide the same level of service in the Fixed λ case.) Using these ds, we find N_Λ^{IB} and N_Λ^{HB}. We then let $\lambda^{\mathrm{IB}}(N_\Lambda^{\mathrm{IB}})$ be the maximum integral λ for the Fixed λ case that results in the staffing level N_Λ^{IB}, i.e., $\lambda^{\mathrm{IB}}(N_\Lambda^{\mathrm{IB}}) = \max\{\lambda\colon N_\lambda^{\mathrm{IB}} = N_\Lambda^{\mathrm{IB}}\}$. (Typically, there is only one such λ.) We analogously define $\lambda^{\mathrm{PB}}(N_\Lambda^{\mathrm{PB}})$ and $\lambda^{\mathrm{HB}}(N_\Lambda^{\mathrm{HB}})$. If the realized demand in the Uncertain λ case is less than or equal to $\lambda^{\cdot B}(N_\Lambda^{\cdot B})$, there would be at least the optimal number of servers. That is, $H_\Lambda(\lambda^{\cdot B}(N_\Lambda^{\cdot B}))$ is the probability that there is sufficient staff in the Uncertain λ case for a

Figure 2. Staffing as natural period length T changes for (a) N_λ, Fixed $\lambda = 200$ and (b) N_Λ, Uncertain $E[\lambda] = 200$.



given $N_\Lambda^{\cdot B}$. Because the SLAs define $N_\Lambda^{\cdot B}$ relative to the same Fixed λ solution, N_λ^{PB}, the values of $H_\Lambda(\lambda^{\cdot B}(N_\Lambda^{\cdot B}))$ may be compared. We observe that the PB-SLA consistently responds more conservatively to Uncertain λ than the other two SLAs, staffing on average to the equivalent of approximately 90% of demand, whereas the IB-SLA and the HB-SLA staff to approximately 65% and 63% of demand, respectively.

The result highlights the differences in the assumptions of the models. For the HB-SLA, from Proposition 2(a), the staffing level for Uncertain λ requires only that the constraint $\int_0^\infty \lambda/E[\Lambda]E[I\{K_i\} \mid \lambda, N]\,\mathrm{d}H_\Lambda(\lambda) < (1 - SL)$ be satisfied, whereas for the Fixed λ case, N_λ^{HB} is given by the constraint $E[I\{K_i\}|\lambda, N] < (1 - SL)$. Because periods of high demand are balanced with periods of low demand in the Uncertain λ case, penalties are only incurred if the staffing cannot meet the service level on average. Therefore, if $\lambda = E[\Lambda]$ for the Fixed λ case, one would expect that their optimal values, N_Λ^{HB} and N_λ^{HB}, would not differ greatly. That is, the response to Uncertain λ is muted because the service level must only be achieved in the long run.

In contrast, the PB-SLA incurs a penalty in each period that the value of λ is too high for the staffing level to support, and this penalty is proportional to the number of customers in the period. When λ is uncertain, the staffing level must be determined to balance the operating cost $F(N)$ with the expected penalty incurred from the tail of the distribution. Thus, assuming $\lambda = E[\Lambda]$, we would expect a much higher level of staffing under the Uncertain λ case than under the Fixed λ case. Further, this increase is more pronounced as T increases (as seen in Figure 2) because, first, there are more customers in a period, and second, the probability of a penalty approaches zero or one. That is, each period for the PB-SLA becomes more like an entire horizon. However, the difference with respect to the HB-SLA case is that the value of λ is fixed over the length of the long period and is not balanced out with other values of λ.

Similarly, the IB-SLA faces the entire distribution of Λ, and so should react to the Uncertain λ case by increasing its staffing. In contrast to the PB-SLA case, under an IB-SLA the call center is penalized only for those customers whose service is not acceptable. Because this may be a significantly smaller number, the increase in staffing is smaller.

7. Conclusions

In this paper, we introduce the PB-SLA and compare its performance to that of the IB- and HB-SLAs for outsourced call centers. We generalize the QED staffing regime to reflect a positive acceptable delay, and provide a square root safety staffing rule that is effective for solving the HB-SLA. We develop an approximation for the probability of abandonment that is generally an order of magnitude better than a previously given one. We show how a central limit theorem may be used to accurately approximate the expected penalty rate for the PB-SLA. We use these approximations to find profit-maximizing staffing levels that differ from exact or simulated optimal levels by at most one in the vast majority of cases.

The problem we study is of interest because outsourced call centers may be expected to operate to maximize their profits. By introducing the PB-SLA, we provide a more representative model of the service expected by outsourcing firms. Because the PB-SLA measures the call center over a natural period where the mean demand rate can be considered constant, it induces the call center to staff so that the expected service level is approximately achieved in every period. Under an HB-SLA, a call center may adopt policies that are not congruent with the desires of the outsourcing firm. For example, if demand varies over time, a call center may adopt differing service levels depending on the expected demand rate in a period. Detection

Table 7. Comparison of the response of the solutions to the three SLAs to Uncertain λ.

				d			Uncertain N_Λ			$\lambda(N_\Lambda)$			$H_\Lambda(\lambda(N_\Lambda))$ (%)		
Case	$E[\Lambda]$	r	Fixed N_λ^{PB}	IB	PB	HB	IB	PB	HB	IB	PB	HB	IB	PB	HB
1	100	0	84	0.35	0.33	0.25	85	89	86	102	106	102	64.6	83.6	64.6
		0.9	86	0.44	0.33	0.23	87	92	87	102	108	101	64.6	90.0	58.9
	200	0	162	0.35	0.33	0.27	164	166	163	202	206	201	64.6	83.6	58.9
		0.9	164	0.43	0.33	0.26	166	171	165	203	209	202	70.1	92.4	64.6
2	100	0	89	0.29	0.33	0.25	90	94	91	102	106	102	64.6	83.6	64.6
		0.9	90	0.37	0.33	0.24	91	96	92	102	107	103	64.6	87.0	70.1
	200	0	169	0.31	0.33	0.26	170	173	171	202	205	202	64.6	79.6	64.6
		0.9	171	0.38	0.33	0.25	172	178	172	201	209	201	58.9	92.4	58.9
3	100	0	104	0.12	0.10	0.05	106	111	106	102	108	102	64.6	90.0	64.6
		0.9	104	0.15	0.10	0.05	106	112	106	102	108	102	64.6	90.0	64.6
	200	0	199	0.14	0.10	0.07	200	207	200	202	208	201	64.6	90.0	58.9
		0.9	200	0.155	0.10	0.06	202	208	201	202	208	201	64.6	90.0	58.9
4	100	0	95	0.38	0.33	0.23	97	102	97	103	108	102	70.1	90.0	64.6
		0.9	96	0.41	0.33	0.21	99	103	98	103	108	102	70.1	90.0	64.6
	200	0	186	0.36	0.33	0.23	188	192	188	203	207	202	70.1	87.0	64.6
		0.9	187	0.43	0.33	0.22	189	195	189	203	209	202	70.1	92.4	64.6
5	100	0	81	0.39	0.33	0.29	82	88	82	102	109	101	64.6	92.4	58.9
		0.9	82	0.51	0.33	0.28	83	90	83	102	111	102	64.6	95.9	64.6
	200	0	157	0.38	0.33	0.30	158	163	158	201	209	201	58.9	92.4	58.9
		0.9	158	0.48	0.33	0.29	160	167	159	203	212	202	70.1	97.1	64.6

of this behavior would require monitoring on a periodic basis as we suggest for the PB-SLA. (The study of the PB-SLA in the context of a time-varying arrival rate is a potential extension of this work.) By penalizing each unacceptable service, the IB-SLA aims to treat all customers equally. Because there is natural variability in arrival rates and service times, individual customers may not experience acceptable service even when, over a period, a stated service level is achieved. Thus, the IB-SLA may result in penalties when the PB-SLA would not. The imposition of penalties for failing to acceptably serve an individual is contrary to the spirit of SLAs used in practice where firms are more interested in ensuring average service standards. Under the PB-SLA, for a typical large-scale call center, the penalty is not based on a failure to adequately serve an individual, but rather on a failure to adequately serve hundreds or thousands of customers during a period.

We show that the PB-SLA possesses several salient features. First, it has a more conservative response to uncertain demand than either of the other two SLAs. Second, in contrast to the IB-SLA, it is robust to specification of the revenue rate. Thus, the focus of the PB-SLA is the trade-off between service and cost to the call center. Third, it is applicable over a wide range of parameter values implying it is adaptable to many operating environments. Finally, we note that many standard call center data reports provide aggregate data that support measurement of a PB-SLA.

8. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Endnotes

1. A similar representation has been shown to be a Markov chain in the classic paper by Kiefer and Wolfowitz (1955), and a similar Markov chain representation for the more restricted $M/M/N+M$ case was discussed by Steckley et al. (2005).
2. Geometric decay of the covariance is true for finite-state Markov chains and Markov chains that satisfy Doeblin's criteria, e.g., Meyn and Tweedie (1993, p. 389).

Acknowledgments

The authors thank the editors and referees for their many helpful comments that have improved this work. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Akşin, O. Z., F. de Vericourt, F. Karaesmen. 2008. Call center outsourcing contract analysis and choice. *Management Sci.* **54**(2) 354–368.

Anscombe, F. J. 1952. Large-sample theory of sequential estimation. *Proc. Cambridge Philos. Soc.* **48** 600–607.

Armony, M., E. Plambeck, S. Seshadri. 2009. Sensitivity of optimal capacity to customer impatience in an unobservable M/M/S queue (Why you shouldn't shout at the DMV). *Manufacturing Service Oper. Management* **11**(1) 19–32.

Asmussen, S. 2003. *Applied Probability and Queues*, 2nd ed. Springer, New York.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* **54**(3) 419–435.

Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd ed. Wiley, New York.

Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Brandt, A., M. Brandt. 1999. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Eval.* **35** 1–18.

Brandt, A., M. Brandt. 2002. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s$+GI system. *Queueing Systems* **41**(1–2) 73–94.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

Chen, B. P. K., S. G. Henderson. 2001. Two issues in setting call center staffing levels. *Ann. Oper. Res.* **108**(1–4) 175–192.

de Vericourt, F., O. Jennings. 2006. Nurse-to-patient ratios in hospital staffing: A queuing perspective. Working paper, Duke University, Durham, NC.

Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54**(2) 324–338.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–587.

Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.* **52**(2) 243–257.

Jackson, K. 2002. Thinking beyond the old 80/20 rule. *Call Center Magazine* **15**(1) 54–56.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42**(10) 1383–1394.

Jongbloed, G., G. Koole. 2001. Managing uncertainty in call centers using Poisson mixtures. *Appl. Stochastic Models Bus. Indust.* **17** 307–318.

Kiefer, J., J. Wolfowitz. 1955. On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78** 1–18.

Koole, G. 2005. Convexity properties of queueing systems with applications to call centers. Presentation, Ecole Centrale de Paris, February 10, and INSEAD, Fontainebleau, France, February 11. http://www.math.vu.nl/ koole/presentations/2005insead/pres.html.

Mandelbaum, A. 2004. Call centers: Research bibliography with abstracts. Version 6. Technical report, Technion, Haifa, Israel. http://iew3.technion.ac.il/serveng/References/ccbib.pdf.

Mandelbaum, A., S. Zeltyn. 2006. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. Working paper, Technion, Haifa, Israel.

Meyn, S. P., R. L. Tweedie. 1993. *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Milner, J. M., T. L. Olsen. 2008. Service level agreements in call centers: Perils and prescriptions. *Management Sci.* **54**(2) 238–252.

OutsourcingBestPractices.com. 2001. Ten key questions for developing effective service level agreements. http://www.outsourcing-best-practices.com/ten.html.

Puhalskii, A. 1994. On the invariance principle for the first passage time. *Math Oper. Res.* **19**(4) 946–954.

Ren, J., Y.-P. Zhou. 2008. Call center outsourcing: Coordinating staffing level service. *Management Sci.* **54**(2) 369–383.

Ross, A. M. 2001. Queueing systems with daily cycles and stochastic demand with uncertain parameters. Ph.D. thesis, University of California, Berkeley.

Steckley, S. G., S. G. Henderson, V. Mehrotra. 2005. Service system planning in the presence of a random arrival rate. Working paper, Cornell University, Ithaca, NY.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Sci.* **38**(5) 708–723.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.

Whitt, W. 2005a. Engineering solution of a basic call-center model. *Management Sci.* **51**(2) 221–235.

Whitt, W. 2005b. Two fluid approximations for multi-sever queues with abandonments. *Oper. Res. Lett.* **33** 363–372.

Whitt, W. 2006a. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.

Whitt, W. 2006b. Staffing a call center with uncertain arrival rate and absenteeism. *Production Oper. Management* **15**(1) 88–102.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queuing Systems* **51**(3–4) 361–402.