Service Level Variability of Inbound Call Centers

Alex Roubos[†], Ger Koole[†] & Raik Stolletz[‡]

[†]Department of Mathematics, VU University Amsterdam,

De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

[‡]Chair of Production Management, University of Mannheim,

L13 9, 68163 Mannheim, Germany

a.roubos@vu.nl, ger.koole@vu.nl, stolletz@bwl.uni-mannheim.de

June 7, 2011

Abstract

In practice, call center service levels are reported over periods of finite length that are usually no longer than 24 hours. In such small periods the service level has a large variability. It is therefore not sufficient to base staffing decisions only on the expected value of the service level. In this paper we consider the classical $M/M/s$ queueing model that is often used in call centers. We develop accurate approximations for the service level distribution by means of extensive numerical experimentation based on simulations. This distribution is used for a service level variability-controlled staffing approach to circumvent the shortcomings of the traditional staffing based on the expected service level.

Keywords: call centers, service level, normal distribution, simulations, staffing.

1 Introduction

The hierarchical planning in call centers is usually divided into forecasting, requirements planning for short intervals and staff scheduling, see Gans et al. (2003). For the requirements planning stationary queueing models are applied to derive the minimum number of agents to fulfill a specific performance measure. In call centers the Erlang C model is often used to provide an estimate for the fraction of calls that wait less than Z seconds. This service level estimate Y can be interpreted as the long run fraction of calls that waits less than Z seconds. However, in call centers we are never interested in the long run: service level realizations are considered at 30-minute intervals, and sometimes aggregated over full days, but seldom over

longer periods (see, e.g., Stolletz, 2003). The goal of call center managers is often to meet an aggregated Y/Z service level for a high fraction X of periods.

Service levels fluctuate. The reason for service level deviations is that call centers operate in a highly volatile environment, with possibly erroneous forecasts, staffing levels which are not as planned, etc. But the actual service level will deviate from the service level prediction, even if all other parameters (such as arrival rate and number of agents) are correct. Simulations show that this difference can be considerable, 5% over a whole day is not exceptional (see Section 3). Managers are aware that the actual service level can differ from the expected service level. However, they do not realize that part of the fluctuations are completely due to randomness. It is our personal experience that managers are surprised to learn this and are willing to think in new solutions, such as the one we propose.

Call center managers deal with all other fluctuations by *traffic management*, the activity that consists of rescheduling the workforce on a short notice as to obtain the required service level (see, e.g., Mehrotra et al., 2010). A higher than necessary service level is generally not a problem, but managers might be penalized for failing to meet the target in too many periods. To this end some managers deliberately opt for a higher expected service level $\overline{Y} > Y$ or a lower target time $\overline{Z} < Z$ in order to meet the original target Y/Z with higher likelihood. Such a behavior is also observed in other research areas, such as inventory management (Thomas, 2005). Both approaches are based on the experience of the call center manager, because the influence of $\overline{Y}$ and $\overline{Z}$ on the probability X to reach the target Y/Z is not described in the literature yet.

Of course there are costs involved in deciding on the staffing level. It is imperative to make a trade-off between staffing costs and costs for not reaching the target service level. For example, when staffing according to the expected value of the service level, it can happen that the target service level will only be met 50% of the time intervals (see Section 4). However, one additional staffed agent can already improve this probability to 80%. Is it better to risk not reaching the target service level 50% of the time, or to schedule one additional agent and accept a risk of 20%? This is a trivial decision now, once the costs are quantified. Related to this is the work of Baron and Milner (2009), where approximations are constructed for the expected penalties for failing to meet the target service level for impatient customers.

A challenging task in call center planning is to consider variable arrival rates. For problems to forecast time-varying rates we refer to Steckley et al. (2009) and Akşin et al. (2007). The stationary independent period by period (SIPP) approach (and variants of it) are widely used for time-dependent requirements planning (staffing) in call centers, (see Green et al., 2001, 2003). Ingolfsson et al. (2007) and Stolletz (2008) review these and other evaluation methods for time-dependent systems and compare them in numerical experiments. To the best of our knowledge, there is no method that considers different lengths of the staffing period and the aggregation interval for performance measurement.

The contribution of this paper is twofold. First, we analyze the variability of the service level dependent on the length of the aggregation interval. For such a finite length interval the actual service level is a random variable, and the service level estimate given by the Erlang C formula

is the *expected* service level. We give an accurate closed-form approximation for the complete distribution of the service level and validate it extensively. Second, in contrast to decisions about staffing levels at the basis of the expected service level we propose a new approach for variability-controlled staffing. The approximated distribution of the service level is used to set the staffing level to meet the service level Y/Z with a targeted probability X. We integrate this variability-controlled staffing approach in the traditional SIPP approach for time-dependent rates. With this method the staffing period and the aggregation interval could be different, which is important for highly volatile rates in call centers.

Related to our first contribution, Steckley et al. (2009) provide a descriptive analysis to compute the service level distribution, for a special case only. Their approach works if, upon a customer arrival, it can be determined from the state of the system whether that customer will receive service before Z. In case $Z = 0$, a customer will receive satisfactory service if at least one server is available. So the state can be chosen as the number of customers in the system. For $Z > 0$, a way could be to keep track of the remaining time until becoming idle, for each server. Unfortunately, this will turn out to be computationally infeasible due to the high dimensionality of the state space.

The remainder of the paper is organized as follows. We start in Section 2 with the model description, where the basic notation and definitions are introduced for the queueing model we consider. Section 3 contains a thorough description of the approximations based on numerical experiments. Several performance evaluations are presented as well. The approximations of Section 3 are used in Section 4, where we present a new way to do staffing calculations. We do this in such a way that we have desired control over the variability. In Section 5 we show how our staffing approach could be used to address the issue of non-homogeneous systems. Finally, conclusions and directions for further research are given in Section 6.

2 Model Description

We model a call center by the $M/M/s$ queueing system. Arrivals occur according to a Poisson process with parameter λ. The service times are exponentially distributed with parameter μ. There are s identical independent servers. Arriving customers that find all servers occupied line up in an infinite buffer queue. Arrivals are served in a first-come first-served order. The service level is defined as the fraction of customers with a waiting time in the queue less than τ time units. On the long run, in a stationary situation, the service level can be interpreted as the probability that the waiting time in the queue, W_Q, is less than τ. This probability is given by the Erlang C formula as follows

$$\mathbb{P}(W_Q < \tau) = 1 - C(s, a)e^{-(s\mu-\lambda)\tau},$$

where $a = \lambda/\mu$. The constant $C(s, a)$ can be seen as the probability of delay. This result can be found in many standard books on queueing theory, e.g., Kleinrock (1976). Perhaps the easiest way to compute the probability of delay is by relating it to the probability of blocking in the

$M/M/s/s$ queue, where customers are blocked if upon arrival all servers are occupied. Cooper (1981) gives the following relation

$$C(s,a) = \frac{sB(s,a)}{s - a(1 - B(s,a))},$$
$$B(s,a) = \mathbb{P}(N = s)/\mathbb{P}(N \leq s),$$

where $N \sim \text{Poisson}(a)$, for $s > a$. The necessary and sufficient condition for stability is that the offered load per server defined by $\rho = \lambda/(\mu s)$ is less than one. We will denote $\mathbb{P}(W_Q < \tau)$ by $\mathbb{E}$SL, that is, the expected service level. The expected service level depends on λ, μ, s and τ. Traditionally, service level objectives have been notated as Y/Z, which means that at least Y% of the customers has to wait less than Z seconds. While this steady-state performance measure will be met in the long run, we are interested in the service level aggregated over intervals of finite length t. The realized average service level could be lower or higher than the expected one. The distribution of the realized average service level strongly depends on the length t.

Throughout the paper we assess the accuracy of the approximations and our staffing approach on several examples. We mainly consider two call centers modeled by the $M/M/s$ queueing system, with parameters that could be found in practice. These systems are defined as follows.

Large system $\lambda = 40$, $\mu = 0.2$ and $s = 210$.

Small system $\lambda = 3$, $\mu = 0.2$ and $s = 19$.

Unless specified otherwise the time scale is expressed in minutes and we take the acceptable waiting time equal to $\tau = 1/3$. This means that the expected service level for the large and small systems are 80.7% and 81.3%, respectively.

3 Numerical Approximations

To demonstrate the effect of the aggregation length t on the service level distribution, we have performed straightforward simulations of the large system. The results are shown in Figure 1. The simulations are performed 10,000 times, each starting after a warming-up period of 24 hours (such that the transient effects of starting from an empty system are gone) and continuing for 3 hours and 24 hours, respectively. After each run, one realization of the service level is obtained. The histograms depict what percentage of the runs fall into each of the bins. In both cases the average service level is 80.7%, which is equal to the outcome of the Erlang C formula. However, let us now consider the complete distribution of the service level. The shape of the distribution depends on the level of aggregation. For a short aggregation length (e.g., 3 hours) the distribution is asymmetric and has a large variability, whereas for a longer aggregation length (e.g., 24 hours) the variability decreases and the distribution becomes more like a normal distribution. This can be explained by the central limit theorem (see Baron and

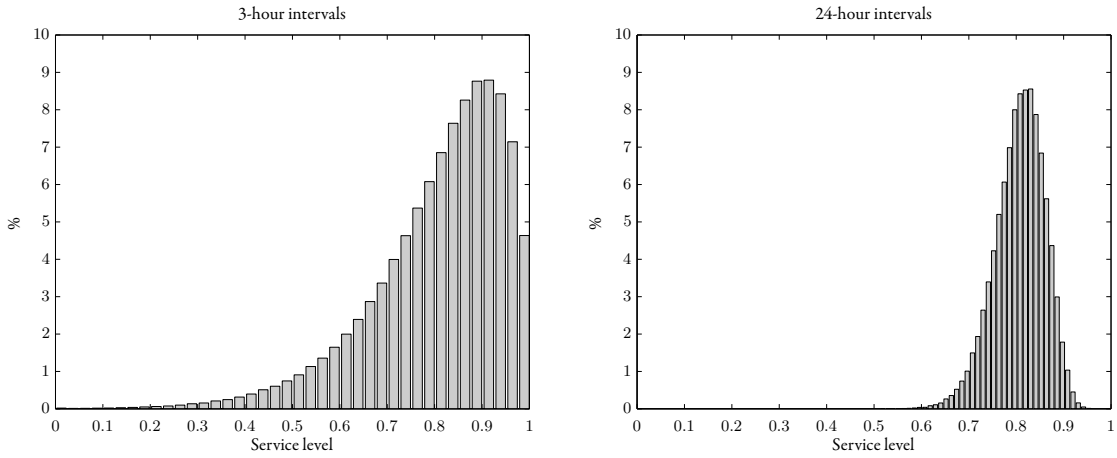

Figure 1. Histograms of the service level aggregated over 3-hour intervals on the left and aggregated over 24-hour intervals on the right.

Milner, 2009, Corollary 2). What is remarkable is that the variability, even when aggregated over the whole day, still is huge: 35% of the realizations deviate more than 5% from the average in this example (i.e., a service level outside [75.7%, 85.7%]).

To account for the significant variability of the service level in intervals of finite length, staffing decisions should not only be made on the basis of the expected value but also on the variability. To be able to do this, we need to quantify this variability. In this section we show that we can accurately approximate the distribution of the service level by the normal distribution. In the normal distribution the variability is characterized by the standard deviation. To this end, we develop an approximation for the standard deviation.

3.1 Standard Deviation Approximation

In the limit $t \to \infty$ the service level distribution approaches the normal distribution. It is intuitively clear (and can also be observed from Figure 1) that the standard deviation goes to zero in this limit. On the other hand, the standard deviation is positive for finite t. Furthermore, if t is large enough, the service level distribution cannot be distinguished from the normal distribution, according to statistical tests for normality (see Section 3.2 for a description of such a test). As a first step we therefore consider large t and express the estimate $\hat{\sigma}$ of the unknown standard deviation σ in the system parameters λ, μ, s, τ and t. We denote that $\hat{\sigma}$ is a function of these parameters by $\hat{\sigma}(\cdot)$. As a next step, we show the results of this approximation for shorter intervals.

The central limit theorem can be used to derive the functional form of σ. The central limit theorem states that the distribution of the average of n independent and identically distributed random variables, each having mean $\mathbb{E}\mathrm{SL}$ and standard deviation ς, converges to the normal distribution with mean $\mathbb{E}\mathrm{SL}$ and standard deviation $\sigma = \varsigma/\sqrt{n}$. Baron and Milner (2009,

Corollary 2) prove that the central limit theorem also holds for a stochastic number of random variables. The contributions of the individual customers to the service level are not independent. However, the contributions of renewal cycles are independent.

Consider a renewal process with as renewal moments the epochs at which an arriving customer initiates a busy period. The time between consecutive renewal moments consists of a busy period B and an idle period I, so that the mean time between renewals is $\mathbb{E}B + \mathbb{E}I$. Then by Asmussen (2003, Proposition 1.4) in the interval of length t the number of renewal cycles converges to $n = t/(\mathbb{E}B + \mathbb{E}I)$ as $t \to \infty$.

Result for $\tau = 0$

For $\tau = 0$ it is possible to derive the standard deviation ς of the service level in a renewal cycle. In this case only the customers that arrive during the idle period are successfully served. In Daley and Servi (1998) the mean and variance are given for the number of arrivals in a busy period, N_B, and in an idle period, N_I. They are

$$\mathbb{E}N_B = \frac{1}{1-\rho}, \qquad \operatorname{var} N_B = \frac{\rho(1+\rho)}{(1-\rho)^3},$$
$$\mathbb{E}N_I = \frac{P_{s-1}}{\pi_{s-1}}, \qquad \operatorname{var} N_I = 2\sum_{i=1}^{s-1} \frac{P_i P_{i-1}}{\pi_i \pi_{s-1}} + \frac{P_{s-1}}{\pi_{s-1}} - \left(\frac{P_{s-1}}{\pi_{s-1}}\right)^2,$$

where π is the steady-state number of customers in the system and $P_i = \sum_{j=0}^{i} \pi_j$. The service level is then given by $N_I/(N_I + N_B - 1)$. (The -1 comes from the fact that the arrival that initiates the busy period is included in both periods.) The expected value of the service level follows immediately from the renewal process and is given by

$$\mathbb{E}\mathrm{SL} = \frac{P_{s-1}/\pi_{s-1}}{P_{s-1}/\pi_{s-1} + \rho/(1-\rho)},$$

which is also equal to the outcome of the Erlang C formula. The variance of the service level in a renewal cycle can be obtained from the multivariate delta method (Casella and Berger, 2002), i.e., a Taylor series expansion. Using the most important terms in the series expansion, the variance simplifies to

$$\varsigma^2 \approx \frac{\operatorname{var} N_I}{(\mathbb{E}N_I + \mathbb{E}N_B - 1)^2} - 2\frac{\mathbb{E}N_I \operatorname{var} N_I}{(\mathbb{E}N_I + \mathbb{E}N_B - 1)^3} + \frac{(\mathbb{E}N_I)^2(\operatorname{var} N_I + \operatorname{var} N_B)}{(\mathbb{E}N_I + \mathbb{E}N_B - 1)^4}.$$

Finally, the mean length of the renewal cycle equals $(\mathbb{E}N_I + \mathbb{E}N_B - 1)/\lambda$ and hence

$$n = \frac{t\lambda}{\mathbb{E}N_I + \mathbb{E}N_B - 1},$$

as $t \to \infty$. The standard deviation is then approximately given by $\sigma = \varsigma/\sqrt{n}$.

A special case is the $M/M/1$ queue, for which these expressions can be simplified to $\mathbb{E}\text{SL} = 1-\rho, \varsigma^2 \approx \rho(1+\rho)(1-\rho), n = t\lambda(1-\rho)$ and $\sigma^2 \approx (1+\rho)/(\mu t)$.

In Steckley et al. (2009) a descriptive analysis is provided to approximate the standard deviation in case $\tau = 0$. We have extended their results by providing a closed-form solution. Both methods give exactly the same standard deviation. This follows from an analytical comparison in case $s = 1$, and from a numerical comparison in case $s > 1$.

Although this standard deviation approximation for $\tau = 0$ has been analytically derived, numerical results show that it is not accurate for a high utilization. For example, if ρ goes to one in the $M/M/1$ queue, the standard deviation goes to $\sqrt{2/(\mu t)}$, a positive number. However, one would expect that the standard deviation goes to zero, because there is no variability when the expected service level is zero. Differences are clearly noticeable for $\rho > 0.5$ for the $M/M/1$ queue. The accuracy increases for systems with more agents. For instance, a system with $s = 10$ has a perfect accuracy for $\rho < 0.9$. Since this approach can only be applied to systems with $\tau = 0$, we take the following alternative approach to approximate the standard deviation for $\tau \geq 0$.

Method for $\tau \geq 0$

The method consists of generating the 'real' standard deviation σ by means of simulations, for different parameter combinations. Then, we try to find an approximation $\hat{\sigma}$, such that the approximation is very accurate on all generated instances. The parameter combinations used in the simulations are obtained by the following steps.

1. We varied the target service level from the set {0.25, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95} and the acceptable waiting time τ from the set {1/6, 1/3, 1/2, 1, 2}.

2. We varied the offered load ρ within the interval $[0.5, 1)$ in step sizes of 0.001 and we fixed μ equal to 0.25.

3. For given values of ρ and μ there exists a unique combination of the pair (λ, s) such that the expected value of the service level is as close as possible to the Y/Z service level chosen in step 1. After this step the s remains fixed.

4. Due to the integrality constraint of s, however, the expected service level might not be close enough to the target. For given values of ρ and s we generally can get arbitrarily close by changing μ and hence λ. To be precise, we increase μ by a step size until the expected service level is greater than the target. In this case we half the step size and start decreasing μ until the expected service level is lower than the target. We continue until we reach the Y/Z service level within the desired accuracy of 0.001. The only exception is that for very lightly loaded systems the s computed in step 3 might already be too high to ever reach the target. We ignored these instances.

	λ	μ	s	τ	t
Lower bound	0.1	0.2	1	1/6	6000
Upper bound	200	2	750	2	6000

Table 1. Bounds of the parameter combinations used for approximating σ.

Table 1 lists the bounds of the parameter combinations that we have obtained using this scheme. Note that we have a value of $t = 6000$ for the aggregation interval, which is large enough for the normal distribution to be justified. In total we have performed well over 20,000 different simulations. Each simulation is independently executed 1,000 times out of which one simulated standard deviation of the service level is obtained. Again, the warming-up period is 24 hours.

In this way we have for a wide range of parameter combinations the standard deviation. The goal is to construct a function $\hat{\sigma}$ that can very accurately fit the data.

Result for $\tau \geq 0$

Looking at the simulated values, it became apparent that, for a fixed service level and acceptable waiting time, the data can be completely described by the following simple function

$$\hat{\sigma}(\lambda, \mu, s, \tau, t) = \frac{\alpha(\mathbb{E}\mathrm{SL}, \tau)}{\sqrt{\mu s}(1-\rho)\sqrt{t}}, \tag{1}$$

where α is a parameter that depends on the system parameters only through the expected service level and the acceptable waiting time. To approximate α we impose the functional form given by $\alpha(\mathbb{E}\mathrm{SL}, \tau) = (1 - \mathbb{E}\mathrm{SL})^{a_1 + a_2\tau} \cdot \mathbb{E}\mathrm{SL}^{b_1 + b_2\tau} \cdot (c_1 + c_2\tau)$. This specific form is motivated by our observations in the data and the requirement that the standard deviation is zero in case the expected service level is either zero or one. The constants are determined by the least-squares regression over all experiments. In the end, α is given by

$$\alpha(\mathbb{E}\mathrm{SL}, \tau) = (1 - \mathbb{E}\mathrm{SL})^{0.4348 + 0.0132\tau} \cdot \mathbb{E}\mathrm{SL}^{1.0708 + 0.0776\tau} \cdot (1.6271 + 0.0339\tau). \tag{2}$$

The corresponding mean squared error then is $4.4 \cdot 10^{-6}$. In addition, the mean absolute percentage error is only 3.4%, despite the divisions by very small numbers. The value of the coefficient of determination, defined by $R^2 = 1 - \sum_i (\sigma_i - \hat{\sigma}_i)^2 / \sum_i (\sigma_i - \bar{\sigma})^2$, is 0.98.

Figure 2 shows how the value of α depends on the expected service level and the acceptable waiting time. If the expected service level is close to its bounds of zero or one, i.e., a really bad or an excellent customer service, the value of the parameter α is close to zero. Also, for increasing values of the acceptable waiting time, the parameter α decreases.

Validation

To validate Equation (1), we simulated 200 new instances that are shown in Figure 3. The left plot shows the simulated and approximated standard deviation σ for the small and large

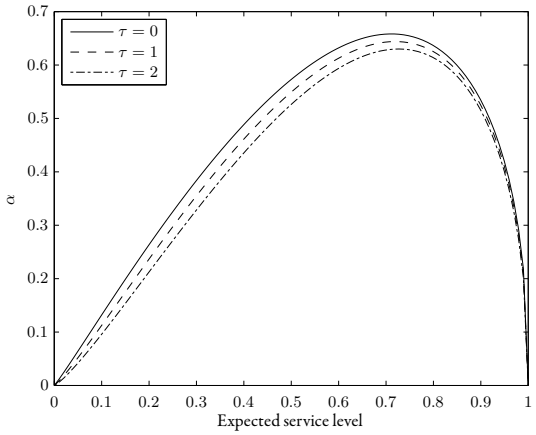

Figure 2. Plot of the function α dependent on the expected service level, for different values of τ (in minutes).

system, dependent on the utilization ρ. The arrival rate λ is changed from the base examples such that the pre-specified ρ is obtained. This plot shows that the standard deviation is well approximated for a broad range of ρ. Only in case of an unrealistically high utilization the standard deviation is overestimated. In these cases ($\rho > 0.98$) the expected service level is way below 50%, so there are more important concerns other than a well approximated standard deviation. The standard deviations increase if ρ increases up to a very high utilization before it starts to diminish. If we compare the standard deviation for the small and the large system, we see that the large system has a lower standard deviation than small system, for $\rho < 0.98$. Also, the standard deviation for the large system is zero up to a utilization of $\rho = 0.80$. This can be explained by the realized service levels for the different utilizations. The expected service level in the large call center is always greater than the one in the small call center according to the economies of scale. The expected service level of the large call center is 100% up to a utilization of $\rho = 0.80$. It is astonishing to see that the large system has a higher standard deviation than the small system for $\rho \geq 0.98$.

Next we show how generalizable the approximations are. We consider parameter combinations chosen at random uniformly between the lower and upper bounds as displayed in Table 1. The interval length t is chosen from $[600, 6000]$ instead, as to allow other large intervals as well. Randomly chosen parameter combinations can result in unstable system. Therefore, we only considered stable systems. In addition, we considered systems with an expected service level less than 1 only. Otherwise, the standard deviation will be zero since there is no variability. After 500 randomly selected instances we get that the mean value of the simulated standard deviation is $1.1 \cdot 10^{-3}$. Moreover, we obtain a mean absolute error of $6.3 \cdot 10^{-5}$ and the maximum absolute error is $3.0 \cdot 10^{-3}$. The absolute percentage error corresponding to this maximum then is only 1.9%. We see that under all circumstances the accuracy of the approximation is very good.

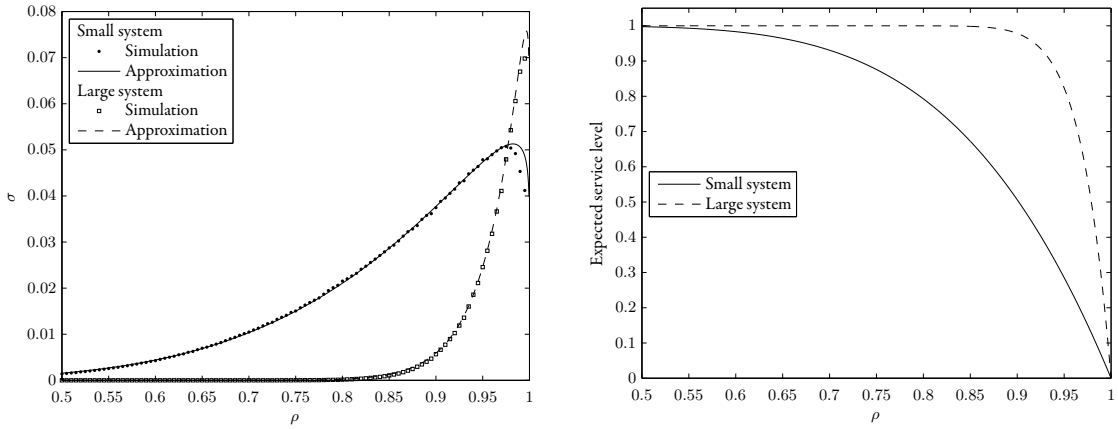

Figure 3. Comparison of the simulated and approximated standard deviation, for the large and small system. The plot on the right shows the corresponding expected service level.

Shorter Intervals

So far, we have considered large values of the interval length t. We have developed an approximation for the standard deviation of the service level that shows to have an excellent accuracy in these cases. Moreover, the distribution of the service level is indistinguishable from the normal distribution.

In shorter intervals the distribution will be different from the normal distribution (see, e.g., Figure 1). This is because there are too few busy periods in order for the central limit theorem to provide a good approximation. Our approximation of the standard deviation is motivated by the applicability of the central limit theorem. Since we are looking at a stochastic number of busy periods, n, the standard deviation will also be different from $\sigma = \varsigma/\sqrt{n}$ in shorter intervals. Consequently, our standard deviation approximation will have a lower accuracy.

To assess the accuracy of the standard deviation approximation in shorter intervals, we have performed additional experiments. In Table 2 the results are shown on the two examples, for intervals ranging from 30 minutes up to 1440 minutes. The table shows the simulated standard deviation σ, the approximated standard deviation $\hat{\sigma}$ and the relative difference between these two. There can be made two observations. First, as the intervals become smaller, the standard deviation becomes larger. Second, as the intervals become smaller, the accuracy of the approximation becomes less. Both observations have been explained already. There is also a difference between the large and the small system. The approximation of the standard deviation is more accurate on the small system. This is likely the result of a smaller busy period length, since the offered load is less.

	Large system			Small system		
t (minutes)	σ	$\hat{\sigma}$	Δ%	σ	$\hat{\sigma}$	Δ%
30	0.260	0.372	43.248	0.218	0.278	27.750
60	0.214	0.263	22.887	0.173	0.197	13.709
120	0.166	0.186	11.785	0.131	0.139	6.309
180	0.140	0.152	8.114	0.109	0.114	3.810
360	0.103	0.107	4.546	0.079	0.080	1.295
720	0.074	0.076	2.879	0.057	0.057	0.003
1440	0.053	0.054	2.243	0.040	0.040	0.291

Table 2. Accuracy assessment of the standard deviation approximation, for several interval lengths t.

3.2 Normal Approximation

While the relative differences of the standard deviation approximation can be quite large for small intervals, what is more important is the accuracy of the normal approximation that uses this standard deviation approximation. As will be shown in this subsection, the accuracy of the resulting normal approximation is good. In total we get that the service level distribution can be approximated as

$$\mathrm{SL} \sim \mathcal{N}(\mathbb{E}\mathrm{SL}, \hat{\sigma}^2). \tag{3}$$

The mean of the service level distribution is equal to the outcome of the Erlang C formula and the standard deviation is defined by Equations (1) and (2).

There are two possible sources of error in this approximation. First, the standard deviation might not be estimated correctly. We have assessed the accuracy of the standard deviation approximation in the previous subsection. Second, the normal distribution itself might not be a good distribution for the service level. This we can test.

To test the null hypothesis that a sample from the unknown service level distribution comes from a distribution in the normal family, we perform the Lilliefors test (Lilliefors, 1967). This is a goodness-of-fit test similar to the Kolmogorov-Smirnov test, with the difference that the mean and variance of the sample are used in the null hypothesis. The test statistic is

$$D = \max_x |G(x) - F(x)|,$$

where G is the empirical cumulative distribution function estimated from the sample and F is the normal cumulative distribution function with mean and standard deviation equal to the mean and standard deviation of the sample. The null hypothesis is rejected if the test statistic is larger than the critical value.

If we perform the Lilliefors test on the two examples, we find the test statistics as shown in Table 3. The values D are decreasing in the interval length t. This suggests that the normal

	Large system				Small system			
t (minutes)	D	Sim	App	Δ%	D	Sim	App	Δ%
30	0.220	0.405	0.330	18.449	0.206	0.506	0.456	9.741
60	0.180	0.503	0.470	6.533	0.147	0.578	0.561	2.981
120	0.123	0.580	0.569	1.934	0.082	0.638	0.635	0.497
180	0.089	0.617	0.613	0.730	0.069	0.667	0.667	0.024
360	0.066	0.669	0.670	0.060	0.049	0.708	0.710	0.284
720	0.047	0.709	0.710	0.122	0.036	0.738	0.740	0.266
1440	0.032	0.738	0.738	0.096	0.025	0.760	0.761	0.159

Table 3. Test statistic of the normal approximation and comparison of the 0.1-quantile between the simulation and the approximation, for several interval lengths t.

distribution becomes an appropriate distribution for the service level as the intervals become larger. However, for all intervals shown in the table, the null hypothesis is rejected at a 5% significance level.

Given that we make an error in the approximation of the standard deviation and in the approximation by the normal distribution, we are interested in the accuracy of Equation (3). Therefore, we compare the 0.1-quantiles of our approximated service level distribution with the empirical distribution based on simulations. If we denote by F^{-1} the quantile function, then we have in the former case, for $x = 0.1$,

$$F^{-1}(x) = \mathbb{E}\mathrm{SL} + \Phi^{-1}(x)\hat{\sigma},$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function. Table 3 lists the results of the comparison between the simulation and the approximation, together with the relative error. From these results, we can observe that the error is decreasing in the interval length t. This is as expected since both the standard deviation approximation and the approximation by the normal distribution become more accurate when the interval length is increased. We can also see that the approximation becomes really good starting from an interval length of 120–180 minutes. In conclusion we can say that, even in the case where the service level is clearly not normally distributed, the normal approximation for this unknown distribution performs very well.

4 Variability-Controlled Staffing

Staffing decisions that are made solely at the basis of the expected value suffer from the variability in the service level. Depending on a couple of factors, it can very well be the case that the target service level will only be met 50% of the time. These factors include, for instance, the level of aggregation and the expected service level. Improved decisions are possible such that these kind

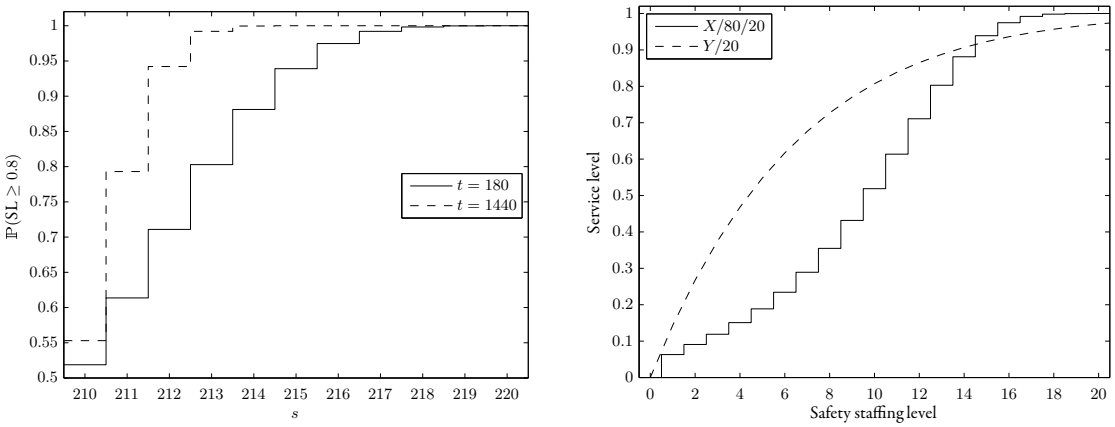

Figure 4. Left plot: Stairs plot of the probability that the 80/20 service level will be met as a function of the number of agents, for two values of t. Right plot: Service level as a function of the safety staffing level. Examples based on the large system with $s^{\min} = 200$.

of situations are prevented. By taking the distribution of the service level into account, one can control how often the target service level will be met.

The left plot in Figure 4 shows the probability that the service level objective will be met depending on the number of agents. From a managerial point of view this figure is useful in two different ways. First, for a given staffing level it could be used to show with what probability the target service level will be met. Second, for a given target it shows the optimal staffing level. This staffing decision is based on a new service level objective. Instead of an Y/Z service level we now get an $X/Y/Z$ service level. This means that in X% of the intervals the target service level of Y/Z will be met. The variability-controlled staffing level $\hat{s}$ can be calculated as follows, taking 90/80/20 as an example,

$$\hat{s} = \min\left\{s \in \mathbb{N} \,|\, \mathbb{P}(\mathcal{N}(\mathbb{E}\text{SL}, \hat{\sigma}^2) \geq 0.8) \geq 0.9\right\}. \tag{4}$$

Remark. The new way to do the staffing calculations in Equation (4) generalizes the way it is done in the Erlang C formula. When we take $t \to \infty$ we have $\hat{\sigma} \to 0$ and the approximation of the service level by the normal distribution becomes deterministic with value $\mathbb{E}\text{SL}$. Then in Equation (4) the probability $\mathbb{P}(\mathbb{E}\text{SL} \geq 0.8)$ is either 1 or 0. So the staffing level corresponding to the $X/Y/Z$ service level is the same as that of the Y/Z service level for $t \to \infty$. Also, the $50/Y/Z$ service level results in the same staffing level, for all t, as the Y/Z service level. This is because the normal distribution is symmetric and if the probability must be at least a half that a normally distributed random variable is greater than some value y, then this reduces to whether or not the mean is greater than y.

A planning according to the variability-controlled staffing level comes at higher staffing costs. The minimum number of agents needed to handle all calls in a deterministic system is $s^{\min} = \lceil \lambda/\mu \rceil$. The planning according to the traditional Y/Z service level leads to a higher

number of agents $s^{Y/Z}$. The difference $s^{Y/Z} - s^{\min}$ could be interpreted as a safety staffing level to provide a higher service to the customers and is further increased to the safety staffing level $\hat{s} - s^{\min}$ according to the variability-controlled staffing of Equation (4). The right plot in Figure 4 shows the expected service level $Y/20$ and the probability X to reach the 80/20 service level as a function of the safety staffing level. To reach an 80/20 service level a safety staffing level of 10 agents is needed. To reach this service level with a probability of 90% in an interval of $t = 180$ the safety staffing level increases to 15 agents. If the call center management include an $X/Y/Z$ service level in their contracts, they have to consider the additional costs for these increased staffing level in their pricing schemes.

We demonstrate the implications of our staffing approach on the staffing levels for the large and small call center. The default staffing levels are 210 and 19 agents, respectively. Due to the observed deviation in the service level, the traditional 80/20 service level will be met only in, respectively, 55.3% and 62.6% of the intervals of length 24 hours. For different interval lengths and for different target service levels the variability-controlled staffing levels are displayed in Table 4. The optimal values derived via time-consuming simulations are given in parentheses. From the table a couple of observations can be made. Firstly, it is not surprising to see that the staffing levels increase if the traditional target service level must be met with higher probability. Secondly, the smaller the intervals, the more uncertainty there is in service level. Hence generally more agents are needed as well. However, this does not hold for the 50/80/20 service level, because the 80/20 service level will be met with a probability higher than 50% with the default staffing levels. Thirdly, the absolute increase in staffing levels is larger for the larger call center, than it is for the smaller call center. This is because of the *law of diminishing returns* (see, e.g., Koole and Pot, 2010) which states that the marginal increase in service level declines in the number of agents. An increase in expected service level is needed to ensure that the target service level is satisfied with the specified probability. Verification with simulations shows that a good amount of these staffing levels are indeed optimal. The staffing levels for the examples with $X < 99$ are optimal for $t \geq 180$, because our approximation of the service level distribution is very accurate. In the cases $t \leq 120$, there is a slight over- or understaffing of at most two in our examples, except for $X = 99$. This justifies the applicability of the approximations once more.

5 Staffing for Non-Homogeneous Systems

The SIPP approach is a traditional approach that helps to determine performance measures and staffing levels for time-dependent systems. In these systems the parameters (essentially the arrival rate and number of agents) are dependent on the time. This is for instance denoted by the $M(t)/M/s(t)$ queueing system. From a practical point of view the staffing levels $s(t)$ are to remain constant within a planning period, which typically has a duration of 30 minutes. In the SIPP approach a stationary queueing model, e.g., the $M/M/s$ model, is constructed for each planning period. Each model is then independently solved for the minimum number of agents needed to meet the target service level.

	Large system				Small system			
t (minutes)	50/80/20	90/80/20	95/80/20	99/80/20	50/80/20	90/80/20	95/80/20	99/80/20
30	210 (208)	219 (217)	220 (220)	223 (226)	19 (18)	22 (22)	23 (23)	23 (25)
60	210 (208)	217 (216)	218 (219)	220 (224)	19 (19)	22 (21)	22 (22)	23 (24)
120	210 (209)	216 (215)	217 (217)	218 (221)	19 (19)	21 (21)	21 (22)	22 (23)
180	210 (210)	215 (215)	216 (216)	217 (220)	19 (19)	21 (21)	21 (21)	22 (22)
360	210 (210)	214 (214)	214 (214)	216 (217)	19 (19)	20 (20)	21 (21)	21 (21)
720	210 (210)	213 (213)	213 (213)	214 (215)	19 (19)	20 (20)	20 (20)	21 (21)
1440	210 (210)	212 (212)	213 (213)	213 (214)	19 (19)	20 (20)	20 (20)	20 (20)

Table 4. Variability-controlled staffing levels for different target service levels and interval lengths. (Optimal staffing levels are in parentheses.)

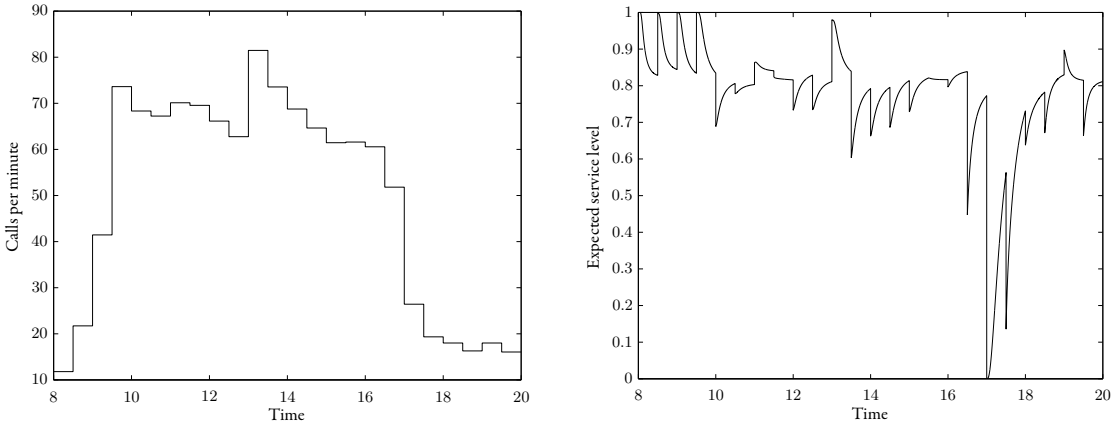

Figure 5. Left plot: incoming call volume by 30-minute intervals. Right plot: transient expected service level.

In this section we show how our variability-controlled staffing approach can be integrated in the SIPP approach. To this end we consider a real-life example of a large banking call center. Available data consist of call detail records out of which, among other things, the call volumes and average service time can be extracted. The call volumes are shown in the left plot in Figure 5, from 8.00 until 20.00. The call volumes outside this time period are negligible. The average service time turns out to be 2.5 minutes ($\mu = 0.4$).

In Tables 5 and 6 we compare the traditional approach with the variability-controlled staffing approach for different lengths of the aggregation period, equal to 30 minutes, 6 hours and 12 hours. For each approach we report the number of staffed agents in each 30-minute interval, and the expected service level and the probability to meet the service level, aggregated over 30 minutes, 6 hours and 12 hours.

When we apply the SIPP approach to this call center, and model each 30-minute planning period by the $M/M/s$ queueing system, we can find the optimal staffing levels such that in each period the 80/20 target service level will be met. These staffing levels are displayed in

the columns labeled 80/20 in Table 5. We assess the performance of this staffing approach by means of simulations. In the simulations we modeled the change in staffing levels from one period to the next by the so-called exhaustive discipline (see Ingolfsson, 2005). This means that agents, that are still serving customers, will only leave as soon as they finish the call. This discipline is beneficial to the service level in periods in which the staffing level is lower than in the previous period. That the expected service level is not reached in each 30-minute period is due to the assumption of independent periods in the SIPP approach (see Stolletz, 2008), where waiting customers at the end of one period are not carried over to the next period. This effect is visible in the example in Table 5 for periods with a significant decrease in the arrival rate compared to the former period, for example in the period 17.00 – 17.30. Potentially larger queues at the end of the former period with more agents are carried over into a period with less agents. This leads to longer waiting times in the period with less agents. We can also observe this from the right plot in Figure 5, which shows the transient expected service level $\mathbb{E}\mathrm{SL}(t)$ for a customer arriving at time t (see Ingolfsson et al., 2007). Even though there are periods with a good average service level, the probabilities that the target service level will be met in the 30-minute periods are very low. Overall there are 1566.5 agent hours needed for the traditional SIPP approach without taking the variability of the service level into account.

The second part of Table 5 shows the results of the variability-controlled staffing according to 90/80/20 for 30-minute aggregation intervals in each 30-minute planning period, i.e., the length of the staffing period equals the length of the aggregation interval. This results in higher staffing levels and higher expected service levels. Moreover, the probabilities of reaching the desired target service level are brought to an acceptable level. For the same reason as in the 80/20-SIPP approach, the variability-controlled SIPP approach does not reach the desired probability to reach the service level in each interval.

Usually call center managers are more interested in aggregated service levels over several hours. To integrate the length of the interval for performance measurement, we apply the variability-controlled staffing approach for the different 30-minute periods. Assume that the service levels are reported over 6-hour intervals. For each 30-minute planning period we staff according to the 90/80/20 target service level for 6-hour intervals with the arrival rate of the respective 30-minute period. That is, this planning results in staffing decisions for short periods due to the dynamics in call volumes, but take into account the longer intervals for performance aggregation. For aggregation intervals of 6 and 12 hours, Table 6 reports for each 30-minute period the staffing levels and simulation results of the expected service level and the probability that the 80/20 service level will be met. Since the staffing levels are higher than the 80/20 case and lower than the 90/80/20 0.5-hour case, the results are also in between the two cases of Table 5.

Furthermore, in Tables 5 and 6 the results of the aggregated performance assessment are shown. For the aggregation of performance measures over periods with different arrival rates and staffing levels, we consider calls which start the service in the respective periods. The aggregated results show that for staffing according to 80/20 the probability to meet the 80/20 service

level over the whole day is very low, with a value just above 50%. On the other hand, staffing according to 90/80/20 for 30-minute aggregation intervals gives an excessive probability. Again, the results for staffing according to 90/80/20 for 6-hour and 12-hour aggregation intervals are in between. More importantly, the probabilities to reach the service level are closer to the desired values.

The last row shows the overall agents hours needed. The shorter the aggregation interval, the more agents are needed. In our example, the difference between the traditional approach and a 30-minute period is 61 agent hours, i.e., working with service goals for short intervals would need 3.89% more agent hours. When we compare the traditional approach with the 6-hour and 12-hour periods, we find an increase of 1.53% and 1.12% agent hours, respectively. Such analysis of additional costs is valuable in contract negotiations, where the call center management now knows the costs for a shorter aggregation interval for service level goals.

6 Conclusions and Further Research

In this paper we have considered the service level distribution beyond its expectation. When aggregated over intervals of finite length, the service level has a non-negligible variability. Motivated by the central limit theorem, we have approximated the service level distribution by the normal distribution. In the normal distribution the variability is characterized by the standard deviation. By means of extensive numerical experimentation based on simulations, we have developed an accurate closed-form approximation for the standard deviation, dependent on the length of the aggregation interval. These approximations for the service level distribution turn out to be quite accurate, also for relatively short intervals.

Using the complete distribution of the service level, it is possible to make improved staffing decisions. Our variability-controlled staffing approach offers the possibility to control the probability that the traditional target service level is met. This results in an $X/Y/Z$ service level objective. This means that in $X\%$ of the aggregation intervals the Y/Z target service level will be met.

Finally, we have shown, by means of an example, how our variability-controlled staffing approach could be integrated in the traditional SIPP approach to deal with time-dependent arrival rates. Since the service levels are often aggregated over several hours, we apply our approach in each small planning period, but for a longer aggregation interval. Although the assumptions of the SIPP approach are not justified, it is clear that our approach adds value for the call center management.

A possible direction for further research could be to consider more realistic models, instead of the basic $M/M/s$ queueing system. In reality customers are impatient and will abandon if their waiting time in the queue exceeds some (stochastic) threshold. This introduces the patience distribution as another parameter where the service level depends on. Maybe abandoned customers will redial at a later time, giving rise to two more parameters: the redial probability and the redial time distribution. Furthermore, it has been shown (see, e.g., Jongbloed and

Koole, 2001) that the Poisson process cannot explain all variability in the arrival process. The arrival rate itself could therefore be modeled by a random variable. In addition, the service time distribution is in practice different from the exponential distribution (the lognormal distribution would be more appropriate). It would be valuable if the dependence of all these characteristics on the service level distribution could be quantified.

References

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

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

Baron, O., J. Milner. 2009. Staffing to maximize profit for call centers with alternate service-level agreements. *Operations Research* **57** 685–700.

Casella, G., R.L. Berger. 2002. *Statistical Inference*. 2nd ed. Duxbury Press.

Cooper, R.B. 1981. *Introduction to Queueing Theory*. 2nd ed. North Holland.

Daley, D.J., L.D. Servi. 1998. Idle and busy periods in stable $M/M/k$ queues. *Journal of Applied Probability* **35** 950–962.

Gans, N., G.M. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5** 79–141.

Green, L.V., P.J. Kolesar, J. Soares. 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* **49** 549–564.

Green, L.V., P.J. Kolesar, J. Soares. 2003. An improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management* **12** 46–61.

Ingolfsson, A. 2005. Modeling the $M(t)/M/s(t)$ queue with an exhaustive discipline. Working paper.

Ingolfsson, A., E. Akhmetshina, S. Budge, Y. Li, X. Wu. 2007. A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline. *INFORMS Journal on Computing* **19** 201–214.

Jongbloed, G., G.M. Koole. 2001. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* **17** 307–318.

Kleinrock, L. 1976. *Queueing Systems, Volume I: Theory*. Wiley.

Koole, G.M., S.A. Pot. 2010. A note on profit maximization and monotonicity for inbound call centers. *Operations Research* To appear.

Lilliefors, H.W. 1967. On the Komogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* **62** 399–402.

Mehrotra, V., O. Ozlük, R. Saltzman. 2010. Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management* **19** 353–367.

Steckley, S.G., S.G. Henderson, V. Mehrotra. 2009. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences* **23** 305–332.

Stolletz, R. 2003. *Performance Analysis and Optimization of Inbound Call Centers*. Springer.

Stolletz, R. 2008. Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queuing models: The stationary backlog-carryover approach. *European Journal of Operational Research* **190** 478–493.

Thomas, D.J. 2005. Measuring item fill-rate performance in a finite horizon. *Manufacturing & Service Operations Management* **7** 74–80.

	80/20						90/80/20 0.5-hour					
Interval	s	$\mathbb{E}$SL			$\mathbb{P}(\text{SL} \geq 0.8)$		s	$\mathbb{E}$SL			$\mathbb{P}(\text{SL} \geq 0.8)$	
8.00 – 8.30	34	0.896	0.828	0.797	0.812	0.675	37	0.971	0.959	0.946	0.970	0.999
8.30 – 9.00	60	0.898			0.820		64	0.975			0.975	
9.00 – 9.30	110	0.906			0.832		115	0.974			0.965	
9.30 – 10.00	191	0.914			0.845		198	0.983			0.980	
10.00 – 10.30	178	0.773			0.612		184	0.946			0.916	
10.30 – 11.00	175	0.796			0.653		181	0.954			0.931	
11.00 – 11.30	183	0.849			0.739		189	0.968			0.955	
11.30 – 12.00	181	0.816			0.682		187	0.959			0.941	
12.00 – 12.30	173	0.799			0.654		178	0.945			0.912	
12.30 – 13.00	164	0.786			0.637		170	0.951			0.924	
13.00 – 13.30	211	0.901			0.820		218	0.980			0.972	
13.30 – 14.00	191	0.739			0.566		197	0.929			0.884	
14.00 – 14.30	179	0.753			0.588		185	0.945			0.913	
14.30 – 15.00	169	0.777			0.621		175	0.953			0.927	
15.00 – 15.30	161	0.791			0.645		166	0.946			0.917	
15.30 – 16.00	161	0.820			0.685		167	0.962			0.943	
16.00 – 16.30	159	0.828			0.702		164	0.957			0.934	
16.30 – 17.00	136	0.697	0.751		0.493	0.382	142	0.918	0.928		0.869	0.984
17.00 – 17.30	72	0.376			0.070		76	0.657			0.291	
17.30 – 18.00	54	0.625			0.381		57	0.875			0.786	
18.00 – 18.30	50	0.768			0.596		54	0.953			0.935	
18.30 – 19.00	46	0.796			0.626		49	0.941			0.914	
19.00 – 19.30	50	0.844			0.715		54	0.965			0.959	
19.30 – 20.00	45	0.782			0.597		48	0.932			0.897	
Agent hours	1566.5						1627.5					

Table 5. Simulation results of staffing according to 80/20 and 90/80/20 for 30-minute aggregation intervals.

	90/80/20 6-hour						90/80/20 12-hour					
Interval	s	$\mathbb{E}$SL		$\mathbb{P}(\text{SL} \geq 0.8)$			s	$\mathbb{E}$SL		$\mathbb{P}(\text{SL} \geq 0.8)$		
8.00 – 8.30	35	0.926	0.903	0.883	0.936	0.933	35	0.926	0.886	0.883	0.888	0.860
8.30 – 9.00	61	0.928		0.882			61	0.928		0.882		
9.00 – 9.30	112	0.940		0.903			112	0.940		0.903		
9.30 – 10.00	194	0.954		0.924			193	0.941		0.902		
10.00 – 10.30	180	0.858		0.754			180	0.849		0.735		
10.30 – 11.00	178	0.893		0.818			177	0.873		0.781		
11.00 – 11.30	185	0.919		0.859			184	0.893		0.811		
11.30 – 12.00	184	0.908		0.842			183	0.881		0.793		
12.00 – 12.30	175	0.883		0.801			174	0.854		0.747		
12.30 – 13.00	166	0.871		0.774			166	0.863		0.762		
13.00 – 13.30	214	0.951		0.919			213	0.937		0.895		
13.30 – 14.00	194	0.858	0.879	0.753			193	0.827	0.860	0.702		
14.00 – 14.30	182	0.879		0.791			181	0.846		0.731		
14.30 – 15.00	171	0.873		0.779			170	0.839		0.719		
15.00 – 15.30	163	0.877		0.787			162	0.844		0.727		
15.30 – 16.00	163	0.895		0.821			163	0.882		0.796		
16.00 – 16.30	161	0.894		0.817			160	0.878		0.788		
16.30 – 17.00	138	0.804	0.845	0.659	0.768		138	0.793	0.823	0.636	0.671	
17.00 – 17.30	73	0.470		0.111			73	0.471		0.118		
17.30 – 18.00	55	0.739		0.533			55	0.735		0.524		
18.00 – 18.30	52	0.889		0.803			51	0.848		0.727		
18.30 – 19.00	47	0.877		0.781			47	0.865		0.753		
19.00 – 19.30	52	0.925		0.881			51	0.893		0.817		
19.30 – 20.00	46	0.857		0.735			46	0.853		0.731		
Agent hours	1590.5						1584					

Table 6. Simulation results of staffing according to 90/80/20 for 6-hour and 12-hour aggregation intervals.