Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Manufacturing & Service Operations Management

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

Service-Level Variability of Inbound Call Centers

Alex Roubos, Ger Koole, Raik Stolletz,

To cite this article:
Alex Roubos, Ger Koole, Raik Stolletz, (2012) Service-Level Variability of Inbound Call Centers. Manufacturing & Service Operations Management 14(3):402-413. http://dx.doi.org/10.1287/msom.1120.0382

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 14, No. 3, Summer 2012, pp. 402–413
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.1120.0382

Service-Level Variability of Inbound Call Centers

Alex Roubos, Ger Koole
Department of Mathematics, VU University Amsterdam, 1081 HV Amsterdam, The Netherlands
{a.roubos@vu.nl, ger.koole@vu.nl}

Raik Stolletz
Business School, University of Mannheim, 68131 Mannheim, Germany, stolletz@bwl.uni-mannheim.de

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 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 based on extensive 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.

Key words: call centers; service level; normal distribution; simulations; staffing
History: Received: January 26, 2010; accepted: December 28, 2011. Published online in *Articles in Advance* May 4, 2012.

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 used to determine 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 no more than Z seconds. This service-level estimate Y can be interpreted as the long-run fraction of calls that wait no more 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 service-level target that Y% of the calls are answered within Z seconds is commonly expressed in the form Y/Z. 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. One of the reasons for service-level deviations is that call centers operate in a highly volatile environment, with possibly erroneous forecasts, staffing levels that are not as planned, etc. Even if all parameters are correct, the realized service level will still deviate from the service-level prediction, because of the intrinsic randomness in the call center environment. Simulations show that this difference can be considerable, for example, 5% over a whole day is not exceptional (see §3). Managers are aware that the actual service level can differ from the expected service level. However, they do not realize the impact of the randomness on the amount of fluctuations. It is our personal experience that managers are surprised to learn this and are willing to consider new solutions, such as the one we propose.

Call center managers deal with fluctuations by *traffic management*, the activity that consists of rescheduling the workforce on short notice 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 $\bar{Y} > Y$ or a lower target time $\bar{Z} < Z$ to meet the original target Y/Z with higher likelihood. Such behavior is also observed in inventory management (Thomas 2005) and other fields. Both approaches are based on the experience of the call center manager, because the influence of $\bar{Y}$ and $\bar{Z}$ on the probability X to reach the target Y/Z is not yet described in the literature.

Costs play a crucial role in our analysis. For example, when staffing according to the expected value of the service level, the target service level may only be met 50% of the time intervals (see §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%? To make this trade-off, we have to quantify both the costs of staffing and the costs for not reaching the target service level. Finding the optimal trade-off then becomes equivalent to minimizing total costs. 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.

In call center planning, there are a number of challenging problems related to time-varying arrival rates. For forecasting problems with time-varying rates, we refer to Akşin et al. (2007) and Steckley et al. (2009). 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. In all these methods there is no distinction between 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 as a function of 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 a closed-form approximation for the complete distribution of the service level and validate it extensively. Second, in contrast to decisions about staffing levels that are based on 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 an analysis to compute the service-level distribution for the special case $Z = 0$ 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 on or before Z. In case $Z = 0$, a customer will receive satisfactory service if at least one server is available. Therefore, the state can be chosen as the number of customers in the system. Their approach cannot be generalized to $Z > 0$.

The remainder of this paper is organized as follows. We start in §2 with the model description, where the basic notation and definitions are introduced for the queueing model under consideration. Section 3 deals with the approximations that are based on numerical experiments. Several performance evaluations are presented as well. The approximations of §3 are used in §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 §5 we show how our staffing approach could be used to address the issue of nonhomogeneous systems. Finally, conclusions and directions for further research are given in §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 or equal to τ time units. In 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 or equal to τ. This probability is given by the Erlang C formula:

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

where $a = \lambda/\mu$. The constant $C(s, a)$ can be interpreted as the probability of delay. This result can be found in many standard books on queueing theory (e.g., Kleinrock 1976). 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 $\mathbb{P}(W_Q \le \tau)$ 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 have to wait less than or equal to Z seconds. Both τ and Z can be used to denote the acceptable waiting time, although the unit of Z is seconds and the unit of τ is arbitrary. There is a difference between $\mathbb{E}SL$ and Y: Y is used to denote the target service level, i.e., the minimum required service level; $\mathbb{E}SL$ is the service level that is expected to be obtained given all parameters. Although the steady-state performance measure Y/Z 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.

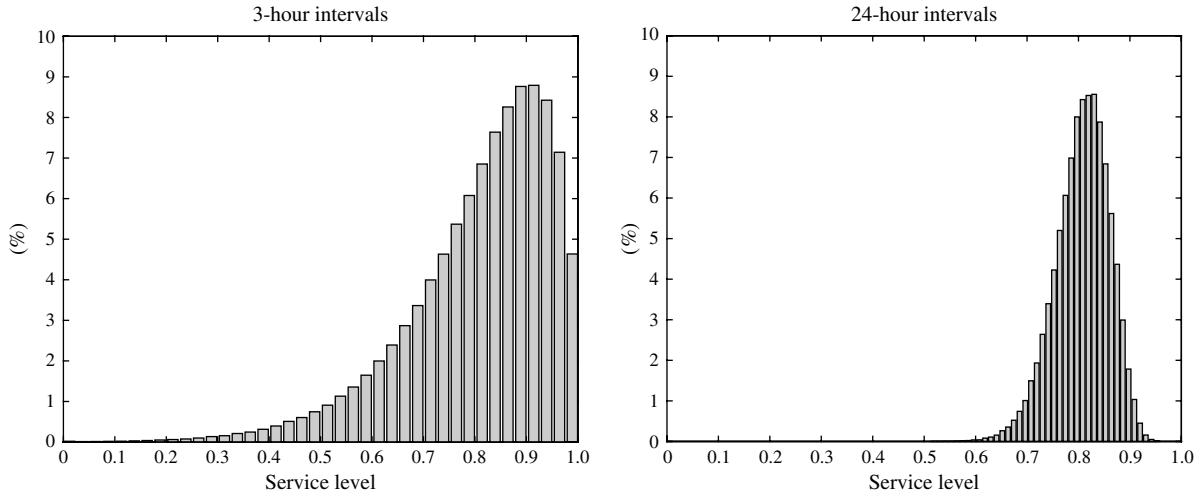
Throughout the paper, we assess the accuracy of the approximations and our staffing approach on several examples. We mainly consider two call centers modeled by the $M/M/s$ queueing system, with parameters that could be found in practice. These systems are defined as follows.

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

Figure 1 Histograms of the Service Level Aggregated over 3-Hour Intervals (Left) and 24-Hour Intervals (Right)



equal to $\tau = 1/3$. This means that the expected service level is 80.7% for the large system and 81.3% for the small system.

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 warm-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. 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 looks more like a normal distribution. This can be explained by the central limit theorem (see Baron and Milner 2009, Corollary 2). What is remarkable is that the variability, even when aggregated over the whole day, is still huge: 35% of the realizations deviate more than 5% from the average in this example (i.e., they have 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 service level, but should also reflect both the variability that is inherent in service levels and the level of desired confidence in achieving the service-level objective. 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 §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 $\mathbb{E}SL$ and standard deviation ς, converges to the normal distribution with mean $\mathbb{E}SL$ 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 the epochs at which an arriving customer initiates a busy period

as renewal moments. 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 $t\to\infty$.

3.1.1. 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

$$\mathbb{E}N_B=\frac{1}{1-\rho},\quad \operatorname{var}N_B=\frac{\rho(1+\rho)}{(1-\rho)^3},$$

$$\mathbb{E}N_I=\frac{P_{s-1}}{\pi_{s-1}},\quad \operatorname{var}N_I=2\sum_{i=1}^{s-1}\frac{P_iP_{i-1}}{\pi_i\pi_{s-1}}+\frac{P_{s-1}}{\pi_{s-1}}-\left(\frac{P_{s-1}}{\pi_{s-1}}\right)^2,$$

where π is the steady-state distribution of the 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 $\mathbb{E}\mathrm{SL}=1-\rho$, $\varsigma^2\approx\rho(1+\rho)(1-\rho)$, $n=t\lambda(1-\rho)$, and $\sigma^2\approx(1+\rho)/(\mu t)$.

In Steckley et al. (2009) an analysis is provided to approximate the standard deviation in case $\tau=0$. That approximation has to be obtained by solving multiple sets of equations. 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 $\sqrt{2/(\mu t)}$, which is nonzero. 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 $\rho<0.9$. That the accuracy decreases at high utilization can be attributed to the application of the central limit theorem. The length of a renewal cycle increases as the utilization increases, and therefore there are fewer independent renewal cycles in a fixed-length interval. Because this approach can only be applied to systems with $\tau=0$, we take the following alternative approach to approximate the standard deviation for $\tau>0$.

3.1.2. Method for $\tau>0$. The method consists of generating the "real" standard deviation σ by means of simulations for different parameter combinations. We then 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.

Step 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}.

Step 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.

Step 3. A unique combination of the pair (λ, s) exists for given values of ρ and μ 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.

Step 4. Because of 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 halve 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.

Table 1 Bounds of the Parameter Combinations Used for Approximating σ

	λ	μ	s	τ	t
Lower bound	0.1	0.2	1	1/6	6,000
Upper bound	200	2	750	2	6,000

Table 1 lists the bounds of the parameter combinations that we have obtained using this scheme. Note that we have a value of $t = 6{,}000$ 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 warm-up period is 24 hours.

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

3.1.3. Result for $\tau > 0$. Motivated by the simulation results for a fixed service level and acceptable waiting time, we deduce the following simple functional form to describe the data

$$\hat{\sigma}(\lambda, \mu, s, \tau, t) = \frac{\alpha(\mathbb{ESL}, \tau)}{\sqrt{\mu s}(1-\rho)\sqrt{t}}, \tag{2}$$

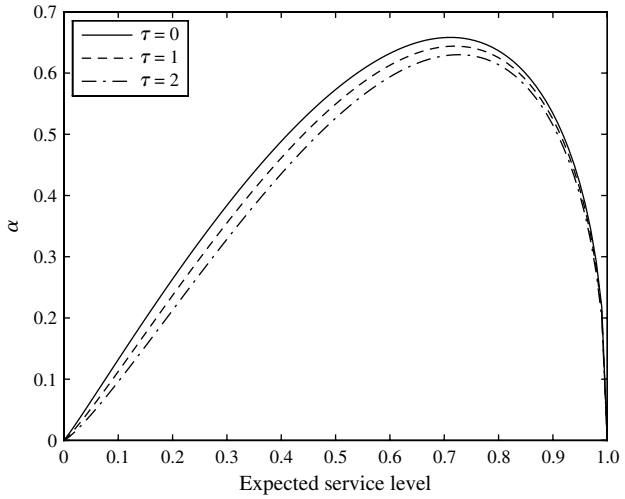
where α is a parameter that depends on the system parameters only through the expected service level and the acceptable waiting time. To approximate α, we impose the functional form given by $\alpha(\mathbb{ESL}, \tau) = (1-\mathbb{ESL})^{a_1+a_2\tau} \cdot \mathbb{ESL}^{b_1+b_2\tau} \cdot (c_1 + c_2\tau)$. This specific form is motivated by our observations in the data and the requirement that the standard deviation is zero in the case when 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

$$\alpha(\mathbb{ESL}, \tau) = (1-\mathbb{ESL})^{0.4348+0.0132\tau} \cdot \mathbb{ESL}^{1.0708+0.0776\tau} \cdot (1.6271 + 0.0339\tau). \tag{3}$$

The corresponding mean squared error is then $4.4 \cdot 10^{-6}$. 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.

Figure 2 Plot of the Function α Dependent on the Expected Service Level for Different Values of τ (in Minutes)

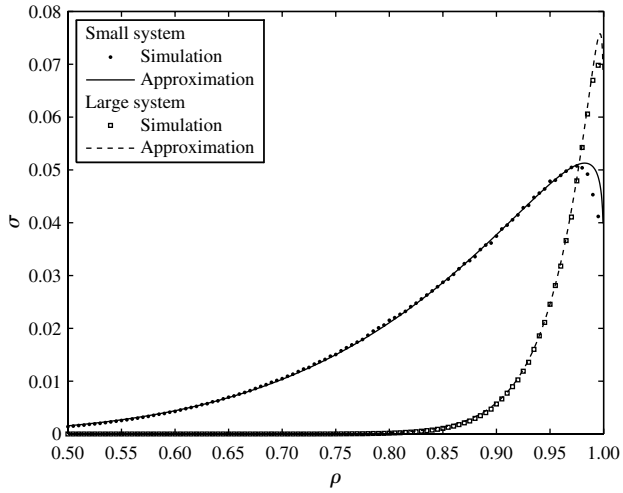


It is possible to apply this approximation to the case $\tau = 0$ as well. Then, we observe an increased accuracy at high utilizations, compared to the analytical approximation for $\tau = 0$ that loses accuracy at high utilizations. In that sense, this approximation is an important addition to the analytical one, even for $\tau = 0$.

3.1.4. Validation. To validate Equation (2), we simulated 200 new instances that are shown in Figure 3. This figure shows the simulated and approximated standard deviation σ for the small and the large system, dependent on the utilization ρ. The arrival rate λ is changed from the base examples such that the 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 is the standard deviation 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.

Next we show the generality of the approximations. 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 $[600, 6{,}000]$ instead to allow other large intervals as well. Randomly chosen parameter combinations can result in unstable systems. Therefore, we only considered stable systems. In addition, we considered systems with an expected service level less than one only. Otherwise, the standard deviation will be zero because 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,

Figure 3 **Comparison of the Simulated and Approximated Standard Deviation for the Large and Small System**



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. For parameter values outside the wide range of values in Table 1, our approximation has yet to be validated.

3.1.5. 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, and we have shown that it has 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. Because 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 1,440 minutes. The table shows the simulated standard deviation σ, the approximated standard deviation $\hat{\sigma}$, and the relative difference between these two. The simulations have been performed 10,000 times for an accurate measurement. Two observations can be made. First, as the intervals become smaller, the standard deviation becomes larger. Second, as the intervals become smaller, the accuracy of the approximation diminishes. Both observations were explained earlier. 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, because the offered load is less.

Table 2 **Accuracy Assessment of the Standard Deviation Approximation for Several Interval Lengths t**

	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
1,440	0.053	0.054	2.243	0.040	0.040	0.291

3.2. Normal Approximation

Although the relative differences of the standard deviation approximation can be quite significant for small intervals, what is more important is the accuracy of the normal approximation that uses this standard deviation approximation. As we show 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

$$\mathrm{SL} \sim \mathcal{N}(\mathbb{E}\mathrm{SL}, \hat{\sigma}^2). \tag{4}$$

The mean of the service-level distribution is equal to the outcome of the Erlang C formula. The standard deviation is defined by Equations (2) and (3).

There are two possible sources of error in this approximation. First, the standard deviation might not be estimated correctly. We 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. We can test this.

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.

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**

	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
1,440	0.032	0.738	0.738	0.096	0.025	0.760	0.761	0.159

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 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 (4). 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 $x = 0.1$,

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

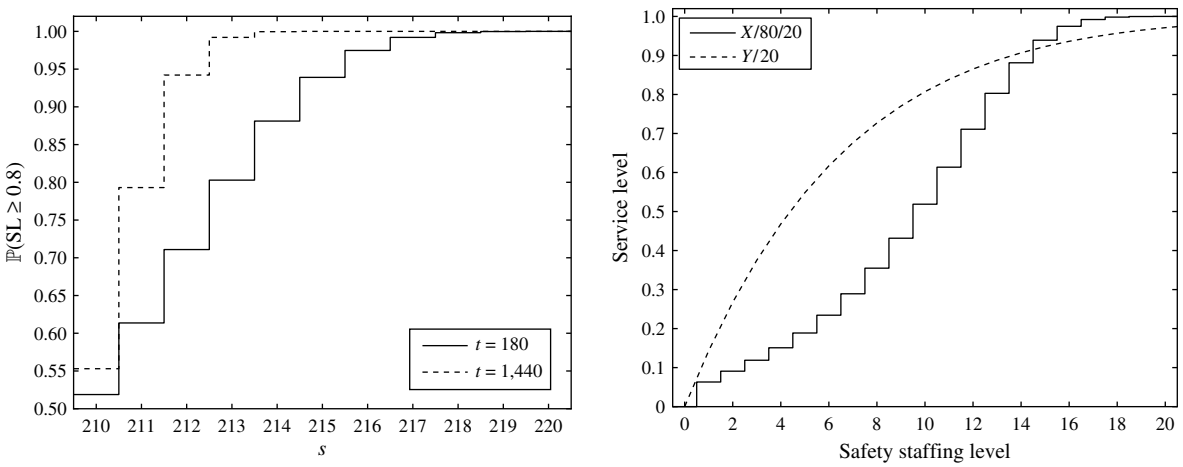
where Φ^{-1} is the inverse of the standard normal cumulative distribution function. Table 3 lists the results of the comparison between the simulation and the approximation, together with the relative error. From these results, we can observe that the error decreases in the interval length t. This is as expected because 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 performs very well starting from an interval length of 120–180 minutes. Therefore, the approximation is useful for relatively slow-changing demand. When arrival rates change significantly in a short period of time, call centers often have to divide time into smaller intervals, typically of 30 minutes. The approximation is not good in such cases.

4. Variability-Controlled Staffing

Staffing decisions that are made solely based on the expected value suffer from the variability in the service level. Depending on a couple of factors, it is possible that the target service level will be reached only 50% of the time. These factors include, for instance, the level of aggregation and the expected service level. By making better decisions, these kinds of situations can be prevented. By taking the distribution of the service level into account, one can control the likelihood that the target service level is 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 a 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

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**



Note. Examples based on the large system with $s^{\min} = 200$.

will be met. The variability-controlled staffing level $\hat{s}$ can be calculated as follows, taking 90/80/20 as an example:

$$\hat{s} = \min\{s \in \mathbb{N} \mid \mathbb{P}(\mathcal{N}(\mathbb{E}\mathrm{SL}, \hat{\sigma}^2) \geq 0.8) \geq 0.9\}. \quad (5)$$

Remark 1. The new way to do the staffing calculations in Equation (5) 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 $\mathbb{E}\mathrm{SL}$. Then in Equation (5), the probability $\mathbb{P}(\mathbb{E}\mathrm{SL} \geq 0.8)$ is either one or zero. Therefore, 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.

Remark 2. It should be noted that the staffing level corresponding to an $X/Y/Z$ target can also be achieved by a $\bar{Y}/\bar{Z}$ target, where $\bar{Y}$ or $\bar{Z}$ are possibly different from Y and Z. To illustrate this, consider, for example, the two cases in the left plot in Figure 4. To reach a 90/80/20 service level, we find $\hat{s} = 212$ for $t = 1{,}440$, and $\hat{s} = 215$ for $t = 180$. The same staffing levels correspond to an 84/20 and 91/20 target, respectively. The notation $X/Y/Z$ has several advantages. For instance, it is immediately clear from the target description what the likelihood is of reaching the Y/Z service level. In addition, the staffing level strongly depends on the interval length t. It is difficult to give an interpretation to $\bar{Y}/\bar{Z}$ on the likelihood X to reach Y/Z for different lengths t.

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 (5). 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 includes an $X/Y/Z$ service level in their contracts, they have to consider the additional costs for these increased staffing levels 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. Because of the observed deviation in the service level, the traditional 80/20 service level will be met only in 55.3% and 62.6% of 24-hour intervals, respectively. For different interval lengths and 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. First, it is not surprising to see that the staffing levels increase if the traditional target service level must be met with higher probability. Second, the smaller the intervals, the more uncertainty 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. Third, 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 2011), 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 $t \geq 180$,

Table 4 Variability-Controlled Staffing Levels for Different Target Service Levels and Interval Lengths

	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)
1,440	210 (210)	212 (212)	213 (213)	213 (214)	19 (19)	20 (20)	20 (20)	20 (20)

Note. Optimal staffing levels are in parentheses.

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 no more than two in our examples, except for $X = 99$. This justifies the applicability of the approximations once more.

5. Staffing for Nonhomogeneous 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.

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 from which we can extract, among other things, the call volumes and average service time. 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 we report the expected service level and the probability of meeting 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. The simulations are performed using 10,000 independent replications starting from an empty system, because the call center starts empty at the beginning of the day. 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 because of 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 fewer agents. This leads to longer waiting times in the period with fewer 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

Figure 5 **Left Plot: Incoming Call Volume by 30-Minute Intervals; Right Plot: Transient Expected Service Level**

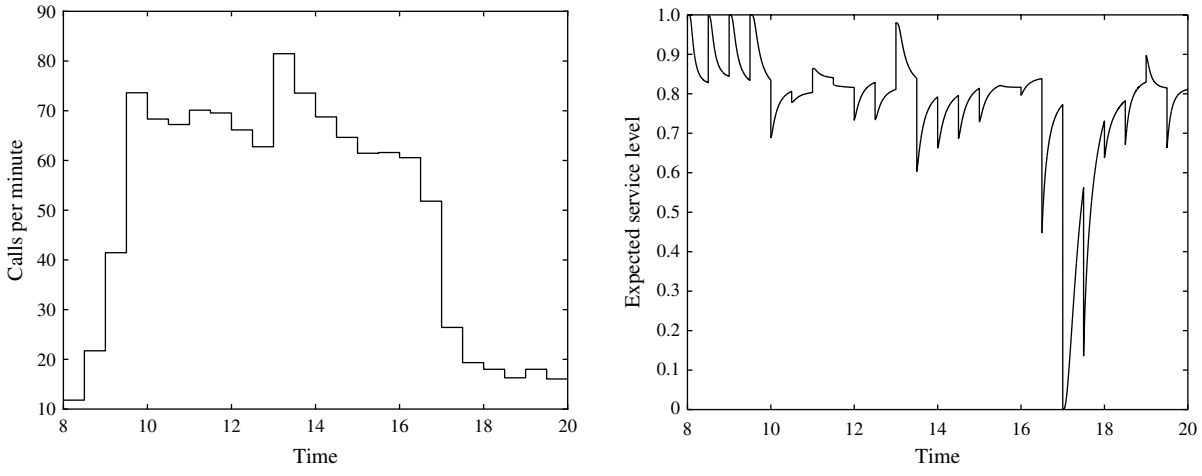


Table 5 Simulation Results of Staffing According to 80/20 and 90/80/20 for 30-Minute Aggregation Intervals

	80/20						90/80/20 30-minute					
Interval	s		$\mathbb{E}$SL		$\mathbb{P}(SL \geq 0.8)$		s		$\mathbb{E}$SL		$\mathbb{P}(SL \geq 0.8)$	
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	0.751		0.588	0.382	185	0.945	0.928		0.913	0.984
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			0.493		142	0.918			0.869	
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	1,566.5					1,627.5						

Table 6 Simulation Results of Staffing According to 90/80/20 for 6-Hour and 12-Hour Aggregation Intervals

	90/80/20 6-hour						90/80/20 12-hour					
Interval	s		$\mathbb{E}$SL		$\mathbb{P}(SL \geq 0.8)$		s		$\mathbb{E}$SL		$\mathbb{P}(SL \geq 0.8)$	
8.00–8.30	35	0.926	0.903	0.879	0.883	0.936	35	0.926	0.886	0.860	0.883	0.888
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			0.753		193	0.827			0.702	
14.00–14.30	182	0.879	0.845		0.791	0.768	181	0.846	0.823		0.731	0.671
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			0.659		138	0.793			0.636	
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	1,590.5					1,584						

level will be met in the 30-minute periods are very low. Overall, there are 1,566.5 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. This planning results in staffing decisions for short periods due to the dynamics in call volumes, but takes 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. Because the staffing levels are higher than the 80/20 case and lower than the 90/80/20 30-minute case, the results are also 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 that 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. The results for staffing according to 90/80/20 for 6- and 12-hour aggregation intervals lie between these two extreme cases. 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- 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 nonnegligible 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, we have developed an accurate closed-form approximation for the standard deviation, depending 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. Because the service levels are often aggregated over several hours, we apply our approach in each small planning period to a longer aggregation interval. Although the assumptions of the SIPP approach are not justified, it is clear that our approach adds value to 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 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, Avramidis et al. 2004, Brown et al. 2005) 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 differs in practice 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 OZ, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.

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

Avramidis AN, Deslauriers A, L'Ecuyer P (2004) Modeling daily arrivals to a telephone call center. *Management Sci.* 50(7): 896–908.

Baron O, Milner J (2009) Staffing to maximize profit for call centers with alternate service-level agreements. *Oper. Res.* 57(3):685–700.

Brown LD, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100(469):36–50.

Casella G, Berger RL (2002) *Statistical Inference*, 2nd ed. (Duxbury Press, Belmont, CA).

Daley DJ, Servi LD (1998) Idle and busy periods in stable $M/M/k$ queues. *J. Appl. Probab.* 35(4):950–962.

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

Green LV, Kolesar PJ, Soares J (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. *Oper. Res.* 49(4):549–564.

Green LV, Kolesar PJ, Soares J (2003) An improved heuristic for staffing telephone call centers with limited operating hours. *Production Oper. Management* 12(1):46–61.

Ingolfsson A (2005) Modeling the $M(t)/M/s(t)$ queue with an exhaustive discipline. Working paper, University of Alberta, Edmonton, AB, Canada.

Ingolfsson A, Akhmetshina E, Budge S, Li Y, Wu X (2007) A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline. *INFORMS J. Comput.* 19(2):201–214.

Jongbloed G, Koole GM (2001) Managing uncertainty in call centers using Poisson mixtures. *Appl. Stochastic Models Bus. Indust.* 17(4):307–318.

Kleinrock L (1976) *Queueing Systems, Volume I: Theory* (John Wiley & Sons, New York).

Koole G, Pot A (2011) A note on profit maximization and monotonicity for inbound call centers. *Oper. Res.* 59(5):1304–1308.

Lilliefors HW (1967) On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *J. Amer. Statist. Assoc.* 62(318):399–402.

Mehrotra V, Ozlük O, Saltzman R (2010) Intelligent procedures for intra-day updating of call center agent schedules. *Production Oper. Management* 19(3):353–367.

Pacheco A (1994) Some properties of the delay probability in $M/M/s/s+c$ systems. *Queueing Systems* 15(1/4):309–324.

Steckley SG, Henderson SG, Mehrotra V (2009) Forecast errors in service systems. *Probab. Engrg. Informational Sci.* 23(2):305–332.

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

Stolletz R (2008) Approximation of the non-stationary $M(t)/M(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach. *Eur. J. Oper. Res.* 190(2): 478–493.

Thomas DJ (2005) Measuring item fill-rate performance in a finite horizon. *Manufacturing Service Oper. Management* 7(1):74–80.