PRODUCTION AND OPERATIONS MANAGEMENT
Vol. 12, No. 1, Spring 2003
Printed in U.S.A.

AN IMPROVED HEURISTIC FOR STAFFING TELEPHONE CALL CENTERS WITH LIMITED OPERATING HOURS*

LINDA V. GREEN, PETER J. KOLESAR, AND JOÃO SOARES
Columbia Business School, 408 *Uris Hall, New York, New York* 10027, *USA*
University of Coimbra, Departamento de Matematica, 3000 *Coimbra, Portugal*

Many telephone call centers that experience cyclic and random customer demand adjust their staffing over the day in an attempt to provide a consistent target level of customer service. The standard and widely used staffing method, which we call the *stationary independent period by period* (SIPP) approach, divides the workday into planning periods and uses a series of stationary independent Erlang-c queuing models—one for each planning period—to estimate minimum staffing needs. Our research evaluates and improves upon this commonly used heuristic for those telephone call centers with limited hours of operation during the workday. We show that the SIPP approach often suggests staffing that is substantially too low to achieve the targeted customer service levels (probability of customer delay) during critical periods. The major reasons for SIPP's shortfall are as follows: (1) SIPP's failure to account for the time lag between the peak in customer demand and when system congestion actually peaks; and (2) SIPP's use of the planning period average arrival rate, thereby assuming that the arrival rate is constant during the period. We identify specific domains for which SIPP tends to suggest inadequate staffing. Based on an analysis of the factors that influence the magnitude of the lag in infinite server systems that start empty and idle, we propose and test two simple "lagged" SIPP modifications that, in most situations, consistently achieve the service target with only modest increases in staffing.

(CALL-CENTERS; STAFFING; QUEUING; NON-STATIONARY; SIPP)

1. Introduction

This paper shows that the algorithm commonly used to determine staffing levels for telephone call centers often fails to meet targeted customer service levels. We identify why this happens, and when the shortfalls are worst. We propose a simple modification that eliminates the problem at modest increase in staffing cost.

Telephone call centers play an increasingly important role in the economy. Observers of the industry estimate that there are over 100,000 call centers in the U.S. with more than 3 million customer service agents. Telephone call-center management has benefited greatly from the use of management science models for decision support—particularly in the areas of forecasting, capacity planning, and staffing and scheduling (Mehrothra 1997).

Most telephone call centers experience a time-varying stream of customer calls (arrivals) that have one or more peaks and valleys over the day. In an attempt to economically provide a consistent level of service at all times of the day, managers typically adjust the staffing

* Received December 2000; revisions received October 2001 and January 2002; accepted February 2002.

1059-1478/03/1201/046$1.25

levels through the workday to match the anticipated demand pattern. This is often done with the assistance of commercially available software packages that incorporate queuing analysis. Unfortunately, although telephone call centers can in principle be modeled as queues, the mathematical models that capture both the randomness and the time-varying pattern of the demand process do not admit of solutions that are easy to use. Hence, the commercial packages that support call-center staffing decisions typically use approximations. The classical Erlang-b and Erlang-c queuing models are the basis of most of the products used to help make these decisions.

The primary focus of this paper is to evaluate and improve upon the most commonly used staffing heuristic, which we call the *stationary independent period by period* or SIPP approach, in the context of telephone call centers with limited operating hours. Our main finding is that the SIPP approach frequently suggests staffing levels that are so low that customer delays substantially exceed management's service target. The contributions of this paper are as follows: (1) to demonstrate the magnitude and domain of SIPP's understaffing; (2) to provide an understanding of the structural reasons for SIPP's understaffing, particularly regarding the magnitude of congestion lags in systems starting empty and idle; and (3) based on this understanding, to propose two simple modifications of SIPP that, in most situations, provide more reliable staffing with little or modest increases in costs.

The SIPP approach begins by dividing the workday into planning periods, e.g., hours, half-hours, or quarter-hours. Then a series of stationary queuing models, most often $M/M/s$ (Erlang)-type models, are constructed, one for each planning period. Each of these period-specific models is independently solved for the minimum number of servers needed to meet the service target in that period. In some implementations, workforce schedules are then constructed by managers without the benefit of additional models, while in others, the planning period staffing requirements generated by SIPP become the right-hand sides of key constraints in an integer programming model that derives the actual staffing schedule. (For examples of the linear programming approach to workforce scheduling, see Segal 1974 or Kolesar et al 1975.) An approach that integrates the optimization and queuing steps is given in Ingolfsson, Haque, and Umnikov (2002) and Ingolfsson and Cabral (2002). While the use of the SIPP approach is common in industry, there has been little published research exploring the conditions when using it provides reasonable minimum staffing levels. (See Green, Kolesar, and Soares 2001 for a review of the extant literature and some proposed remedies for SIPP's shortcomings.)

The research reported on here builds on our earlier work on staffing in service systems that operate continuously over time, so-called 24/7 systems (Green, Kolesar, and Soares 2001, hereafter referred to as GKS). In that research, we found that the SIPP method frequently fails to provide adequate staffing for 24/7 systems. As this paper will show, the understaffing problem may be substantially worse when SIPP is applied to systems which shut down for some part of each day.

Our research indicates that a major reason for SIPP's unreliability is that it implicitly assumes that the system's congestion in a given planning period is a result of the demand arising in that period alone. The fallibility of this assumption is evident even in the case when, though demand is time-varying, staffing levels are kept constant. We have shown that in such systems the time-varying pattern of system congestion is out-of-phase with the demand pattern (see, e.g., Green, Kolesar, and Svoronos 1991). In particular, there is a time lag between the epoch of peak demand and the epoch of peak system congestion. For certain system parameter values, this lag effect can be significant and if ignored, results in decreased staffing levels just as delays are peaking, that is, somewhat after a peak in the arrival rate. As a consequence, unacceptably long delays for service may persist for hours after the arrival peak.

Thus, we postulated that understanding lags is central to understanding how to staff. However, there are no analytical results for congestion lags in finite server systems with cyclic demands. However, Eick, Massey, and Whitt (1993) derived closed form expressions in the case of infinite server, steady-state systems with Poisson arrivals and sinusoidal arrival rates. In GKS, we demonstrated that two simple modifications of SIPP, which incorporate an estimate of the time lag based on the work of Eick, Massey, and Whitt, perform far better than the simple SIPP approach in a broad range of practical situations. Our use of the lag to improve SIPP's performance is related to the modified offered load (MOL) approximation proposed by Jagerman (1975) to estimate blocking probabilities in the nonstationary Erlang loss model. Jennings et al. (1996) suggested that the MOL could be used to determine server staffing in nonstationary, multiserver queues. In Green and Kolesar (1997), we showed that what we called the lagged point-wise stationary approximation (PSA), a precursor to the Lag SIPP approach proposed in GKS, was far more accurate in identifying the staffing levels needed to keep the peak probability of delay low. The lag approach is also considerably simpler computationally than MOL.

Extending the Lag SIPP approach to systems with limited operating hours is not trivial for it requires developing a foundation of knowledge about time lags in such systems. Queues with limited operating hours and cyclic demand patterns have not been extensively studied. Though there is fairly extensive literature on queues with service disruptions, most of it deals with the case of random interruptions such as those due to breakdowns (see e.g., Bardhan 1993) rather than scheduled on- and off-periods. Federgruen and Green (1986, 1989) considered a system with a single server that alternates between fixed on- and off-periods, but with a constant arrival rate which is independent of the state of the server. Furthermore, there is no literature on congestion lags in the first cycle of cyclic demand service systems that start empty and idle—the problem faced here. Unlike the continuously operating case, there are no closed form results available that characterize this lag even for infinite server models.

Therefore, a major focus of this paper is to identify the factors that affect the magnitude of congestion lags in infinite server systems with cyclic demand patterns and limited operating hours. In our analysis of Section 4, we show that in some cases these lags are longer in the first cycle of a system starting empty and idle than in steady state and, more importantly, that they are non-decreasing with the frequency of the arrival rate process. Since the demand patterns of many call centers with limited hours have a higher frequency than those for continuously operating centers, both of these behaviors contribute to SIPP's lesser reliability in these limited operating hours situations.

The paper is organized as follows. We describe the models and our methodology in more detail in Section 2. In Section 3, we demonstrate how SIPP reliability depends upon the system parameters and we identify situations in which SIPP performance is unacceptable. We show that there is a broad range of parameter values and, by implication, actual telephone call centers for which the SIPP approach suggests inadequate staffing. In Section 4, we derive an expression for the number of busy servers in an infinite server system with sinusoidal arrival rate that starts empty and idle, and we use this expression to obtain results on the behavior of lags in the first cycle. In Section 5 we test two modifications of the SIPP method and find that they produce reliable staffing levels in the limited operating window case as well. Using both theoretical models and empirical data, we also explore the impact of these SIPP modifications on total staffing requirements. In Section 6, we offer concluding remarks.

2. Model and Methodology

We study $M(t)/M/s(t)$ queuing systems with $\lambda(t)$, the arrival rate at time t given by the sinusoid

$$\lambda(t) = \lambda + A \sin(2\pi t/T), \qquad 0 \leq t \leq W \tag{1}$$

where T is the period (or cycle length) of the sinusoid, λ is the average arrival rate over period T, $A > 0$ is the amplitude, and W is the length of the operating window. (All times are measured in hours.) To capture the shape of demand patterns that we have seen in applications, we focus on both "single-peak" and "double-peak" models. For modeling an operating window of length W with a single arrival peak, we use the first half of the sinusoid given in equation (1) by setting $T = 2W$ so that the arrivals during the operating window are symmetric around the peak. For the double-peak case we set $T = 2/3W$, which results in two symmetric peaks and one trough during the operating window. To draw comparisons with the continuously operating systems we studied in GKS, we also consider an arrival process modeled by a full sinusoid by setting $T = W$. The other model parameters are μ, the service rate, and $s(t)$, the number of servers on duty at time t.

We observe that, $\bar{\lambda}$, the average arrival rate over the operating window, W, is given by

$$\bar{\lambda} = \frac{1}{W}\int_0^W \lambda(t)\mathrm{d}t, \tag{2}$$

resulting in $\bar{\lambda} = \lambda + 2A/\pi$ for single-peak models, $\bar{\lambda} = \lambda + 2A/3\pi$ for the double-peak models, and $\bar{\lambda} = \lambda$ for the full sinusoid model.

Let $p_n(t)$ be the periodic steady-state probability that n customers are in the system at time t. We obtain these functions numerically by solving the following standard set of differential equations that describe the system (see Gross and Harris 1985):

$$\begin{aligned}
p_0'(t) &= -\lambda(t)p_0(t) + \mu p_1(t), \\
p_n'(t) &= \lambda(t)p_{n-1}(t) + (n+1)\mu p_{n+1}(t) - (\lambda(t) + n\mu)p_n(t), \qquad 1 \le n < s(t), \\
p_n'(t) &= \lambda(t)p_{n-1}(t) + s(t)\mu p_{n+1}(t) - (\lambda(t) + s(t)\mu)p_n(t), \qquad n \ge s(t).
\end{aligned} \tag{3}$$

As we are interested in systems which start each day empty and idle, $p_0(0) = 1$. Details on our numerical analysis methods are given in Green et al. (1991). We focus on the probability of delay as the main measure of customer service. Let $p_D(t)$ be the instantaneous probability that a customer arriving at time t is delayed. This is also the probability that all servers are busy at time t and is given by

$$p_D(t) = 1 - \sum_{n=0}^{s(t)-1} p_n(t) \tag{4}$$

The principal output from our differential equation solver (simulator) is a vector of $60W$ estimates of $p_D(t)$ made at 1-minute intervals over the operating window.

3. SIPP Reliability

An Empirical Example

Before developing our SIPP reliability analysis, we present an empirical example for which SIPP fails. Figure 1 shows the arrival rate curve derived from an insurance company's incoming telephone call center that operates daily over an 8-hour window. (This two-peak pattern is typical of many other call-center data we have seen; see Agnihotri and Taylor 1991.) We tested the performance of SIPP by modeling this call center using a Poisson arrival process with this empirical time-varying pattern and parameter values that reflect the actual call-center operations. The historical service rate was about eight calls per hour; half-hour planning periods were used and the service performance target was a 10% probability of delay.

The staffing levels suggested by SIPP are also shown in Figure 1—a total of 719.5 staff-hours over the 8-hour operating window. Figure 2, the concomitant probability of delay curve, illustrates that the SIPP staffing is clearly inadequate. Specifically, the instantaneous

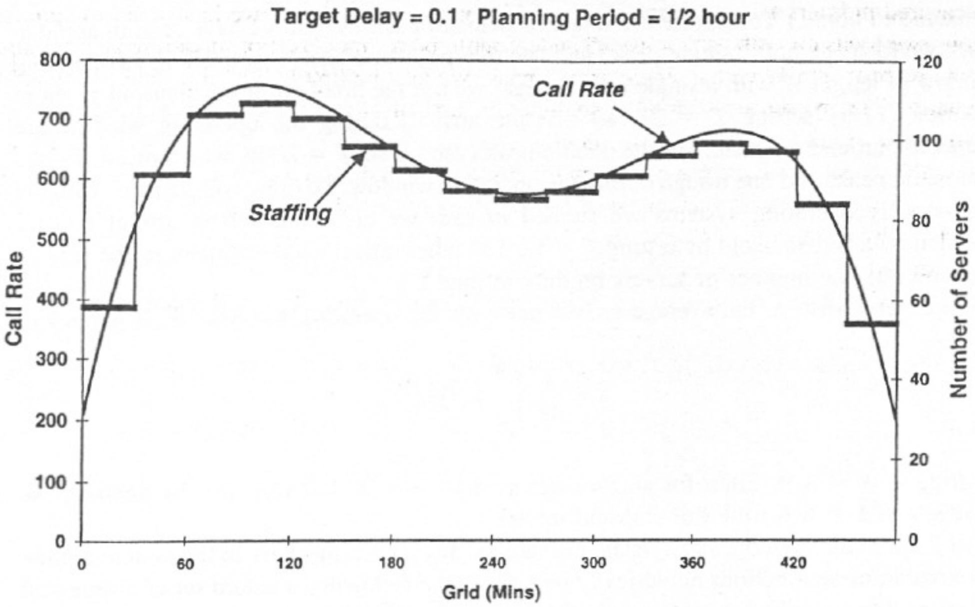

FIGURE 1. Staffing and Call Rates in the Insurance Company Model with SIPP Staffing.

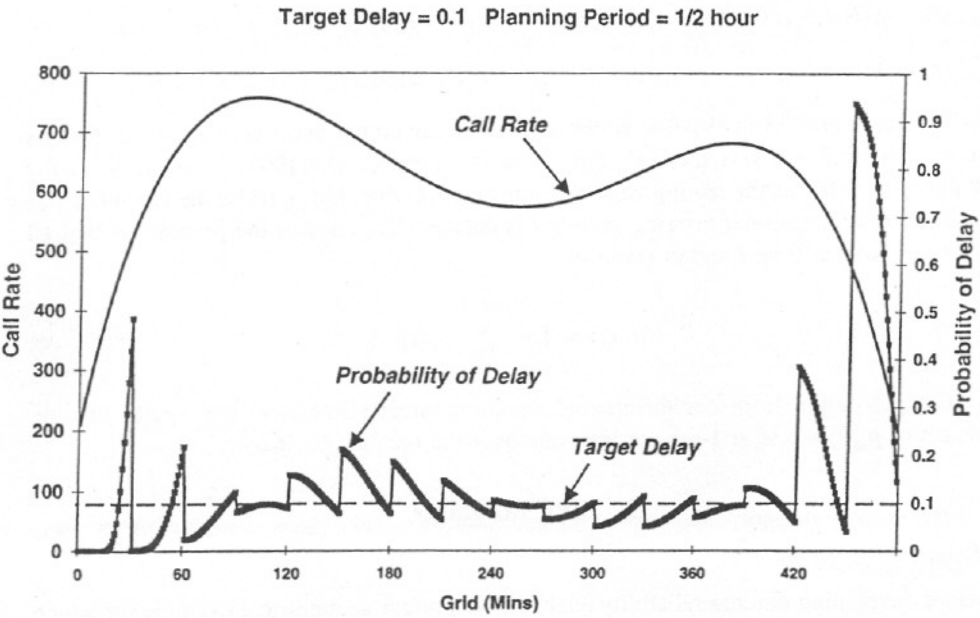

FIGURE 2. Delay Probability in the Insurance Company Model with SIPP Staffing.

peak probability of delay is over 93.5% and the service target is exceeded in 7 of the 16 half-hour planning periods. In 5 of these half-hours, the probability of delay is more than 110% of the target. Thus, if staffed per the SIPP suggestions, actual customer service would be considerably worse than desired.

A Framework for Analysis of SIPP Reliability

Our example illustrates that SIPP can be unreliable in a specific actual scenario. Now, to explore SIPP's reliability more broadly, we analyzed models of service systems with parameter values that span those of many actual call centers that we have experienced personally or have encountered in the literature. A scenario (model) is characterized by the following seven parameters:

- The length of the sinusoidal period of the arrival rate function, T
- The length of the operating "window," i.e., the number of hours the system operates during the day, W
- The average arrival rate over the period T, λ
- The relative amplitude of the arrival rate function, $RA = A/\lambda$
- The service rate, μ
- The target probability of delay, τ
- The length of the planning period, PP

An important derived measure is $\rho = \bar{\lambda}/\mu$, an important measure of system "size." (Recall that $\bar{\lambda}$ is given by equation (2).) Our core set of models has service rates starting at a low of $\mu = 2$ per hour, that is, with average service times as long as 30 minutes, doubling up to 64 per hour ($\mu = 2, 4, 8, 16, 32, 64$). We used average customer arrival rates starting at a low of $\bar{\lambda} = 32$ customers per hour and doubling up to 4,096 per hour ($\bar{\lambda} = 32, 64, 128, 256, 512, 1{,}024, 2048, 4{,}096$). Not all of the 48 ($\mu$, $\bar{\lambda}$) combinations implied by the above were either computationally feasible (when $\bar{\lambda} \gg \mu$ the system of equations (3) becomes too large to solve in any reasonable amount of time) or interesting (e.g., if $\rho = 1$, much of the time there will be only one or two servers). So we limited most of our runs to 18 core (μ, $\bar{\lambda}$) combinations that correspond to ρ values of 16, 32, and 64. Our examination of SIPP reliability used three relative amplitudes: $RA = 0.1$, 0.5, and 1.0; three probability delay targets: $\tau = 0.05$, 0.10, and 0.20; four planning period lengths: $PP = 0.25$, 0.05, 1.0, and 2.0 hours; and four operating window lengths: $W = 8$, 12, 18, and 24 hours. (The inclusion of 24 hours enables comparisons with the 24/7 models studied in GKS.) We also considered three cyclic arrival patterns as described in the last section: one with a single demand peak, one with two peaks, and a full sinusoid for a total of 7,776 scenarios. While these parameter combinations define a broad experimental range, which includes many real call-center scenarios, we do not contend that it covers all regions of possible interest. It is also important to note that the scenarios are not spread uniformly over the experimental region.

The analytic sequence for each scenario is as follows:

1. Fix the scenario's exogenous parameters: T, λ, RA, and μ, and fix the managerial parameters: τ, W, and PP.

2. Divide the cycle into non-overlapping intervals of length PP. For each planning period compute the average arrival rate by using equation (2). Then use this average arrival rate and the service rate in an iterative version of the Erlang delay equation (Cooper 1972, p. 100) to find the minimum staffing needed in the period to achieve the target delay probability, τ. This produces a vector of staffing levels $\{s(n), n = 1, W/PP\}$.

3. Run the simulator with the exogenous parameters specified as in (1) and the $\{s(n)\}$ as determined in (2). This produces the output vector of delay probabilities $\{p_D(t), t = 1, 60W\}$.

4. Using the vector $\{p_D(t)\}$, compute various summary performance measures.

Our analysis focuses primarily on the following performance measure: *the number of half-hours in which the target is exceeded by at least* 10%.

Results

Table 1 summarizes the results of our 648 simulations for the case when the operating window is 12 hours long and there are two peak demand epochs. For each scenario the table contains our main reliability measure—the count of the number of half-hours in which the probability of delay exceeds 110% of the target. By our standard, SIPP is "reliable" for a scenario—a cell in the table—if that count is zero. Conversely, any scenario for which the count exceeds zero will be called an "error."

Overall, we see in Table 1 that SIPP is reliable for only 133 of the 648 scenarios or about

TABLE 1: Counts of half hours in which delay exceeds 110% of target.
Two-peak, 12-hour Model: SIPP Results

			Rho, Mu,																	
					16						32						64			
RA	Plan. Pd.	Target	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64
0.1	0.25	0.05	1	0	0	0	0	0	5	0	0	0	0	0	8	7	1	0	0	0
		0.1	1	0	0	0	0	0	5	0	0	0	0	0	9	6	0	0	0	0
		0.2	0	0	0	0	0	0	5	0	0	0	0	0	7	2	0	0	0	0
	0.5	0.05	2	1	0	0	0	0	7	4	0	0	0	0	7	7	1	0	0	0
		0.1	1	0	0	0	0	0	7	2	0	0	0	0	7	5	0	0	0	0
		0.2	1	0	0	0	0	0	3	2	0	0	0	0	7	6	0	0	0	0
	1	0.05	6	3	2	4	4	4	4	1	1	0	0	0	8	4	4	7	7	8
		0.1	0	0	0	0	0	0	3	2	1	0	0	0	6	4	3	4	2	2
		0.2	4	2	2	0	0	0	5	3	2	2	4	4	8	5	3	6	6	6
	2	0.05	1	0	0	0	0	0	2	5	4	6	6	6	10	9	11	11	11	11
		0.1	2	2	2	2	2	2	2	3	5	4	2	2	8	11	11	10	10	10
		0.2	0	0	0	0	0	0	4	8	8	8	10	10	7	10	9	10	10	10
0.5	0.25	0.05	9	8	7	0	0	0	11	10	10	5	0	0	12	10	10	9	3	0
		0.1	10	10	7	0	0	0	9	10	7	3	0	0	12	11	10	5	0	0
		0.2	9	8	1	0	0	0	10	9	6	3	0	0	10	11	10	7	0	0
	0.5	0.05	9	8	5	1	1	0	11	10	9	5	3	3	12	11	10	8	9	9
		0.1	9	8	2	1	0	0	11	10	10	4	2	3	12	11	10	8	7	8
		0.2	10	8	3	1	0	0	10	8	7	4	1	1	12	11	10	9	4	3
	1	0.05	8	6	7	9	12	12	10	10	11	12	12	12	10	11	12	12	12	12
		0.1	8	4	5	7	7	7	10	9	9	11	12	12	10	10	13	12	12	12
		0.2	9	7	7	6	6	6	10	8	8	9	10	10	10	10	12	12	12	12
	2	0.05	11	11	11	11	11	11	13	13	13	12	12	12	13	14	14	12	12	12
		0.1	12	12	11	12	12	12	13	12	13	12	12	12	14	13	15	13	12	12
		0.2	10	10	9	9	11	12	12	12	14	12	12	12	14	14	14	15	15	12
1	0.25	0.05	12	11	8	8	2	0	13	11	10	8	5	3	14	12	11	10	8	12
		0.1	12	11	8	5	0	0	13	11	10	8	4	1	14	12	10	10	10	8
		0.2	11	11	8	7	1	1	14	12	11	8	3	1	14	13	11	10	8	5
	0.5	0.05	12	11	9	7	9	6	13	11	10	9	14	15	14	12	10	15	18	20
		0.1	13	11	10	8	8	11	13	11	10	10	10	14	14	12	11	14	18	20
		0.2	11	11	10	7	2	2	14	12	11	11	10	13	14	13	11	12	17	16
	1	0.05	10	11	11	12	12	12	10	12	14	12	12	12	12	16	16	14	14	12
		0.1	10	11	14	12	12	12	12	13	14	13	12	12	12	14	16	14	14	13
		0.2	10	11	10	12	12	12	12	12	12	14	12	12	12	13	16	16	14	14
	2	0.05	13	13	13	12	12	12	14	15	15	14	14	14	17	15	16	16	16	14
		0.1	13	15	14	13	12	12	14	14	15	15	14	14	18	15	16	17	17	16
		0.2	14	13	14	15	13	13	14	13	15	15	16	16	18	17	18	17	17	17

20% of the cases. SIPP is generally reliable in the upper left corner of the table where relative amplitude and presented load are both low, while it is extremely unreliable in the lower right corner where the converse is true. We also see that SIPP reliability tends to get worse as service rates decrease and planning-periods increase. These results are illustrative of the patterns we found in our analyses of the other operating windows and arrival patterns. Furthermore, these directional conclusions about SIPP reliability are essentially the same as those we reached in our earlier study of continuously operating systems (see GKS for details and interpretation).

We get a broader understanding of SIPP reliability by examining Table 2, which contains the summary of results for all the scenarios we analyzed. Specifically, this summary demonstrates that SIPP reliability is impacted by the frequency of the arrival rate curve. This is seen first, by noting that reliability is greatest for the single-peak models where the operating window is one-half the sinusoidal period and is worst for the double-peak models

TABLE 2

SIPP Reliability: Percentage of Cases with No Errors

	8-Hour	12-Hour	18-Hour
Single-peak	39.5	45.5	53.2
Double-peak	13.3	20.5	25.2

which operate for one and one-half periods; and second, by observing that SIPP reliability increases with the operating window length. Both of these observations are at least partially the result of SIPP's implicit assumption of stationarity during a planning period. As the rate of change in the arrival rate decreases, this assumption is better and hence SIPP reliability increases.

Table 2 also shows that overall SIPP reliability is almost identical for the comparable continuously operating (steady-state) systems as for those with limited hours. Therefore, in actual call centers, SIPP is more likely to be unreliable for systems with limited operating hours than for those with continuous operations, primarily because of the shorter operating windows.

4. Congestion Lags in Infinite Server Queues

To exploit our conjecture that utilizing the lag concept is a key to improving SIPP performance, we study the transient behavior of infinite server Markovian systems with sinusoidal arrival rates starting with the system empty and idle. Such a system is described by equations (1) and (3) but with $s(t) = \infty$ for $t \geq 0$. We focus on the time lag between the peak in the arrival rate and the peak in the number of busy servers. Our goal is to identify a simple estimate for this time lag which can be used to modify and improve the standard SIPP method for service systems with a fixed length operating window. The basic idea is that since the delay curve lags the arrival curve in cyclical demand systems, the staffing level required at a given time can be more accurately determined by basing it on the arrival rate that occurred approximately a lag period earlier.

In GKS, we proposed the Lag SIPP approach for 24/7 systems which uses $1/\mu$ as an estimate of the lag. This was motivated by the exact, closed form results in Eick et al. (1993) for the steady-state lag in infinite server systems with sinusoidal arrival rate and exponential service times. Eick et al. (1993) showed that the time lag L between the epoch of the maximum arrival rate and the epoch of the maximum server occupancy is a function solely of the service rate μ and the period of the sinusoid T and is given by:

$$L = (\cot^{-1}(\mu/\gamma))/\gamma \tag{5}$$

where

$$\gamma = 2\pi/T. \tag{6}$$

For $\mu \geq 2\pi/T$, this can be expanded in a Taylor series:

$$L = 1/\mu - \gamma^2/3\mu^3 + \gamma^4/5\mu^5 - \cdots \tag{7}$$

indicating that $1/\mu$ is the dominant term affecting the lag. In Green and Kolesar (1998), we numerically confirmed that $1/\mu$ is a very good approximation for L in steady-state, infinite server systems when $\mu \geq 2$ and $T \geq 8$. Our purpose here is to find an equally simple and effective estimate of the lag in the first cycle of an infinite server system that can be used to improve SIPP's reliability for call centers with limited operating windows without using an excessive number of staff.

Let $N(t)$ denote the system occupancy at time t, that is, the expected number of busy servers at time t. From equation (3) it can be shown (Feller 1968) that $N(t)$ satisfies the differential equation

$$N'(t) = \lambda(t) - \mu N(t), \tag{8}$$

with $N(0) = 0$ and with $\lambda(t)$ given by equation (1). Equation (8) is solved by

$$N(t) = \frac{\lambda}{\mu}\left\{1 - \left(\frac{RA\mu\gamma}{\mu^2 + \gamma^2}\right)e^{-\mu t} + \frac{RA}{\mu^2 + \gamma^2}\left(\mu^2 \sin(\gamma t) - \mu\gamma\cos(\gamma t)\right)\right\}, \tag{9}$$

where $\gamma = 2\pi/T$ and T is the period of the sinusoid. We are interested in determining the value of $t > 0$ for which $N(t)$ first reaches a local maximum. The difference between this value and the t where the arrival rate function first reaches its maximum defines the time lag. An analytic solution for the maximum of (9) not being available, we solved for the time lags numerically.

Our results show that as with steady-state lags, the magnitude of the first-cycle lag is unaffected by the mean arrival rate. However, unlike the steady-state case, first-cycle lags are affected by relative amplitude in addition to being affected by the service rate and the period of the sinusoid. This is illustrated in Table 3, which also shows that, as in the steady-state situation, the service rate is the dominant factor influencing first-cycle lags, with the lags decreasing as the service rate increases. At low service rates, i.e., $\mu = 2$, the lag also decreases as the period length T increases. This observation helps explain SIPP's lesser reliability for shorter operating windows, which correspond to smaller values of T. At such low service rates, the lag also decreases as the relative amplitude increases. We believe that this is due to the faster approach to steady-state because of the increased number of transitions that occur at higher amplitudes. It is important to note that for values of the service rate that are likely to correspond to many telephone call-center situations, i.e., $\mu \geq 4$, the impact of both RA and T on lags is slight and the service rate remains the dominant factor.

Table 4 shows how the first-cycle lags compare with the lags in steady-state. Though the first-cycle lags are greater for low service rates, particularly for shorter cycle lengths, these discrepancies virtually disappear for $\mu \geq 4$, a parameter range which is likely to be pertinent for many call centers. The finding that the transient lags are essentially the same as the

TABLE 3

Lags in Hours for the First Cycle

		Service Rate				
Rel. Amp.	Cycle Length (hrs.)	1	2	4	8	16
0.1	8	1.48	0.65	0.25	0.12	0.07
	12	1.40	0.55	0.25	0.12	0.07
	18	1.23	0.50	0.25	0.12	0.07
	24	1.10	0.50	0.25	0.12	0.07
0.25	8	1.17	0.55	0.25	0.12	0.07
	12	1.15	0.52	0.25	0.12	0.07
	18	1.08	0.50	0.25	0.12	0.07
	24	1.03	0.50	0.25	0.12	0.07
0.5	8	1.00	0.52	0.25	0.12	0.07
	12	1.03	0.50	0.25	0.12	0.07
	18	1.02	0.50	0.25	0.12	0.07
	24	1.00	0.50	0.25	0.12	0.07
1	8	0.90	0.50	0.25	0.12	0.07
	12	0.97	0.50	0.25	0.12	0.07
	18	0.98	0.50	0.25	0.12	0.07
	24	0.98	0.50	0.25	0.12	0.07

TABLE 4: Counts of half hours in which delay exceeds 110% of target
First Cycle vs. Steady-state Lags (Relative Amplitude = 0.1)

Cycle Length (hrs.)	Cycle	Service Rate				
		1	2	4	8	16
8	1	1.48	0.65	0.25	0.12	0.07
	infinite	0.85	0.48	0.25	0.12	0.07
12	1	1.40	0.55	0.25	0.12	0.07
	infinite	0.92	0.48	0.25	0.12	0.07
18	1	1.23	0.50	0.25	0.12	0.07
	infinite	0.97	0.50	0.25	0.12	0.07
24	1	1.10	0.50	0.25	0.12	0.07
	infinite	0.98	0.50	0.25	0.12	0.07

steady-state lags for these cases suggests that the Lag SIPP approaches that we had previously proposed for continuously operating systems are likely to be reliable in the limited operating hours case as well. We explore this in the next section.

5. Reliability of a Lagged SIPP Approach

In our study of 24-hour continuously operating systems in GKS, we proposed two modifications of SIPP and guidelines for their use that produce reliable staffing levels in a broad range of applications. Both use a stationary *M*/*M*/*s* model for each planning period but whereas the SIPP method uses a model based on the average arrival rate during the planning period, the proposed alternatives use arrival rates based on a "lagged" arrival rate curve. Specifically, the "Lag Avg" method bases the staffing for each planning period on a model where the $\lambda(t)$ curve is advanced by $1/\mu$ time-units and then averaged over the period. So, for example, if $\mu = 2$ customers per hour, the average arrival rate used to determine the staffing requirement for a planning period starting at t_0 and one hour in length, i.e., $[t_0, t_0 + 1]$, would be calculated using the arrival rates during the interval $[t_0 - 0.5, t_0 + 0.5]$. The "Lag Max" method also advances the $\lambda(t)$ curve by $1/\mu$ time-units but then uses the maximum arrival rate during the planning period instead of the average as input to the queuing model. We propose Lag Max for planning periods in which the arrival rate changes substantially, to compensate for what we believe would be an underestimation of actual delays due to use of the average arrival rate. This is based on our earlier research (Green et al. 1991) which showed that for fixed server queuing systems with sinusoidal Poisson input streams, the average probability of delay is monotone increasing in relative amplitude.

We tested the two lag methods on our scenarios that correspond to actual call centers with limited operating windows, i.e., we excluded from this analysis the full sinusoid demand pattern and the 24-hour operating window. Since, in the limited operating window situation, the arrival rate is assumed to be zero before the first planning period of the day, we modify the lag method for the beginning of the operating window as follows: do not lag the arrival rate in planning period 1 or in planning period *n* if $1/\mu \geq n{*}PP$.

In virtually all scenarios, the lag methods produce fewer errors than the SIPP approach. This is illustrated by Tables 5 and 6, which show the performance of Lag Avg and Lag Max, respectively, for the same 12-hour two-peak scenarios as shown in Table 1. Comparison of these results with the SIPP results in Table 1 shows the extent of the improvement. Focusing first on Table 5, we see that though Lag Avg generally produces fewer errors than SIPP, it is consistently reliable (i.e., no half-hours over 110% of the target) only when $RA = 0.1$ and $PP = 0.25$ or 0.5. However, the Lag Max results in Table 6 show a dramatic improvement. Whereas SIPP is unreliable in over 83% of the 648 scenarios, Lag Max is unreliable in only 5.4%. Moreover, in all scenarios, there are far fewer errors with Lag Max than with SIPP. Lag

TABLE 5: Counts of half hours in which delay exceeds 110% of target
Two-peak, 12-hour Model: Lag Avg Results

			Rho,	Mu,																
					16						32						64			
RA	Plan. Pd.	Target	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64
0.1	0.25	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.5	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0.05	1	0	0	2	2	2	1	1	2	2	0	0	2	5	4	4	7	8
		0.1	0	0	1	0	0	0	1	1	2	1	0	0	3	4	4	4	2	3
		0.2	0	0	0	0	0	0	0	0	0	0	2	4	1	2	1	3	3	3
	2	0.05	4	1	1	0	0	0	3	6	6	7	6	6	7	10	8	11	11	11
		0.1	0	3	3	4	4	4	3	4	7	6	2	2	4	9	11	8	10	10
		0.2	1	1	0	0	0	0	3	4	4	8	10	10	3	9	9	10	10	10
0.5	0.25	0.05	2	0	0	0	0	0	5	1	0	0	0	0	6	0	0	0	0	0
		0.1	1	0	0	0	0	0	4	0	0	0	0	0	6	0	0	0	0	0
		0.2	1	0	0	0	0	0	5	0	0	0	0	0	6	0	0	0	0	0
	0.5	0.05	3	0	0	0	0	1	7	0	0	2	3	2	9	8	7	11	11	9
		0.1	2	0	0	0	0	0	7	1	0	0	1	1	8	5	1	6	6	9
		0.2	4	0	0	0	0	0	6	0	0	0	0	0	6	0	0	3	1	6
	1	0.05	9	8	8	9	12	12	9	11	11	12	12	12	10	11	12	12	12	12
		0.1	6	5	8	9	9	9	7	8	8	12	12	12	9	9	12	12	12	12
		0.2	5	5	9	6	6	6	9	9	12	11	10	10	9	10	12	12	12	12
	2	0.05	6	10	12	12	12	12	11	11	12	12	12	12	11	12	12	12	12	13
		0.1	7	11	11	12	11	12	11	11	12	12	12	12	11	13	12	12	12	12
		0.2	8	11	10	10	11	12	8	12	12	12	12	12	13	14	15	14	13	12
1	0.25	0.05	6	1	1	0	0	0	7	5	2	1	0	4	8	9	10	13	8	11
		0.1	5	3	2	0	1	0	7	6	0	1	1	1	8	8	8	7	6	7
		0.2	6	1	0	0	0	0	6	5	1	0	0	0	8	7	4	1	1	6
	0.5	0.05	7	6	11	5	10	7	9	12	12	13	15	15	11	13	17	19	18	20
		0.1	9	7	2	4	3	6	9	10	13	14	10	15	11	13	17	17	18	18
		0.2	7	4	3	2	2	1	9	6	7	9	9	11	10	10	14	17	19	18
	1	0.05	9	9	10	12	12	12	11	11	12	12	12	12	11	12	13	12	13	12
		0.1	9	9	10	10	12	12	11	11	12	12	12	12	12	12	13	13	13	12
		0.2	9	8	7	10	12	12	10	12	12	12	12	12	12	12	14	14	15	14
	2	0.05	11	12	12	12	12	12	12	14	13	14	14	14	13	15	16	16	16	14
		0.1	11	13	13	12	12	12	13	13	15	16	14	14	13	15	16	16	16	16
		0.2	11	13	15	12	13	13	12	14	15	15	16	16	13	15	15	16	16	17

Max never produces more than three errors and is perfectly reliable for $RA = 0.1$, and whenever $\mu > 4$, a parameter range that we believe is likely to be found in many actual call centers.

Results for the other window lengths are very similar and are summarized in Table 7. For two-peak systems, Lag Max never produces more than three errors and is perfectly reliable for $RA = 0.1$ and $RA = 0.5$ and whenever $\mu > 4$ for the 18-hour window, and for $RA = 0.1$ and whenever $\mu > 4$ for the 8-hour system. In all of these cases, Lag Avg again is generally much less reliable, particularly for larger relative amplitudes and longer planning periods. For the single-peak systems, Lag Max is always perfectly reliable. Lag Avg is perfectly reliable for the 12- and 18-hour systems whenever planning periods are short, i.e., $PP = 0.25$ or 0.5. For the 8-hour system, Lag Avg is reliable for short planning periods, but not when relative amplitude is high, i.e., $RA = 1$. As for the simple SIPP method, Table 7 demonstrates that

TABLE 6: Counts of half hours in which delay exceeds 110% of target.

Two-peak, 12-hour Model: Lag Max Results

			rho	mu																
					16						32						64			
RA	Plan. Pd.	Target	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64
0.1	0.25	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.5	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.5	0.25	0.05	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
		0.1	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	0.5	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0.25	0.05	3	0	0	0	0	0	3	0	0	0	0	0	3	1	0	0	0	0
		0.1	2	0	0	0	0	0	3	1	0	0	0	0	3	1	0	0	0	0
		0.2	2	0	0	0	0	0	3	0	0	0	0	0	3	0	0	0	0	0
	0.5	0.05	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0
		0.1	1	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
		0.2	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	1	0.05	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
		0.1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
		0.2	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

reliability of the Lag methods is generally better for the single-peak models and for longer operating windows.

How well do the Lag methods perform on the empirical data from the insurance company that we considered earlier? Given the above numerical results for sinusoidal models, we

TABLE 7

Reliability of Lag Methods: Percentage of Cases with no Errors

	Lag Avg			Lag Max		
	8-Hr.	12-Hr.	18-Hr.	8-Hr.	12-Hr.	18-Hr.
Single-peak	62.2	65.7	68.8	100	100	100
Double-peak	26.1	34	42.1	92.7	94.6	96.8

predicted that Lag Avg would not perform very well due to the two-peak arrival pattern, while Lag Max should be very reliable. Our prediction was correct. The results for the insurance company model show that Lag Avg results in only three half-hours in which the target is exceeded by more than 10% with a maximum probability of delay of 0.63. (Recall that for SIPP these figures were seven half-hours and 93.5%, respectively.) Figure 3, the $p_D(t)$ curve for the Lag Max solution for the empirical demand, shows that the target is met for the entire operating window.

What is the economic impact of using the Lag approach on total staffing levels? Use of Lag Avg usually produces fewer errors than SIPP does while using the same number of staff-hours. Lag Max does use more staffing hours than SIPP. For the insurance model, the SIPP solution uses 719.5 staff-hours while Lag Max uses 763.5 staff-hours or about a 6% increase. Table 8 contains a summary of the economics and performance of the several staffing methods in the case of 12-hour two-peak models. The table shows that SIPP is optimal (no half-hours above 110% of target and lowest staffing among these methods) for only of 96 of the 648 scenarios and that none of the three staffing methods is optimal in 35 of the scenarios. Lag Avg is optimal in 82 scenarios and, when used in these cases, corrects the SIPP average deficiency of 3.5 half-hours over 110% of target while reducing staffing modestly. Max Lag

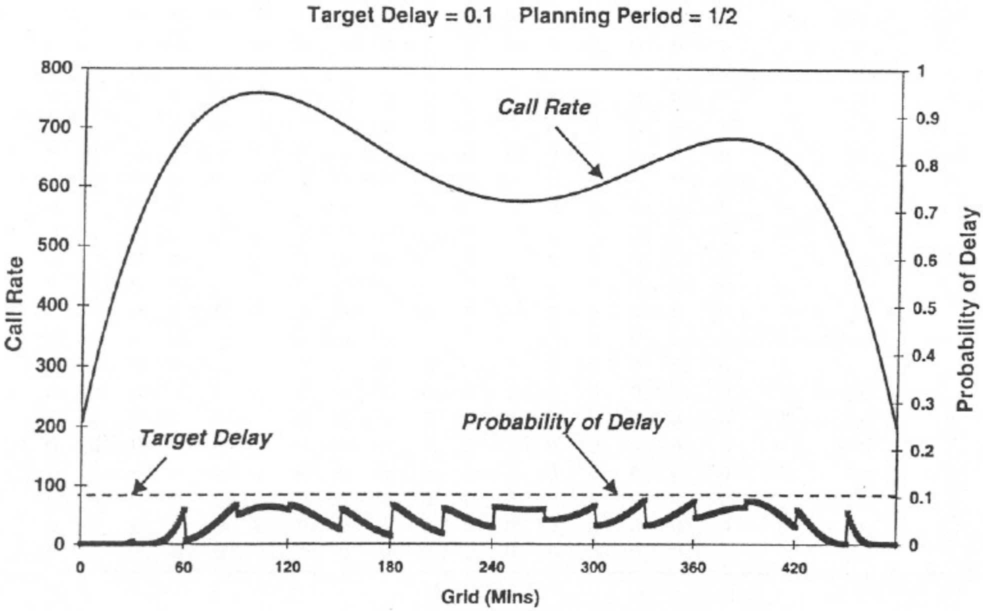

FIGURE 3. Delay Probability in the Insurance Company Model with Max Lag Staffing.

TABLE 8

Staffing Requirements and Errors for the 12-Hour Two Peak Models

	Optimal Staffing Method			
	SIPP	LagAvg	MaxLag	None[1]
Average staffing cost over SIPP (%)	0.0	−0.1	11.8	8.7
Maximum staffing cost over SIPP (%)	0.0	0.0	35.5	18.8
Average SIPP errors (1/2 Hours)	0.0	3.5	9.7	12.2
Maximum SIPP errors (1/2 Hours)	0	11	20	14
Number of scenarios	96	82	435	35

[1] Note: For these 35 scenarios no method was optimal, but MaxLag had the least errors.

is optimal in 435 scenarios where it corrects the SIPP average deficiency of 9.7 half-hours over 110% of target while increasing staffing by 11.8% on average. Finally, examining all 648 scenarios of these two-peak models, we have calculated that the very conservative and simple to implement policy of always using Max Lag increases staffing by 9% on average over SIPP while reducing the number of errors by 98%.

6. Summary and Conclusions

The results in this paper have practical implications for the design and management of many types of telephone call centers. First, our findings show that SIPP frequently fails to provide targeted delay probabilities and by enough to be of concern to managers of actual telephone call centers. Results from both this research as well as our past work support the thesis that SIPP's poor performance is largely due to its failure to account for congestion lags. Second, our findings indicate that congestion lags due to time-varying demands may, in some cases, be significantly longer for call centers that shut down every day than those that operate continuously over 24 hours. This is particularly true when operating windows are short or demand patterns have a higher arrival frequency. An additional problem in these cases is that the increased arrival rate variability runs counter to SIPP's implicit assumption of stationarity during a planning period. The combination of these effects make SIPP even more unreliable for these limited hours systems than in the continuously operating case. SIPP is particularly unreliable for systems with a two-peak demand pattern. Of course, the degree to which actual performance will be as bad as indicated by our results depends on how the proposed staffing requirements are translated into actual work schedules, as well as on how rigidly worker behavior adheres to the suggested schedules. For example, when SIPP-based staffing requirements are used as constraints in an LP-based scheduling model, the LP-generated schedule frequently adds staff (slack) in some periods due to other constraints. However, it is far from certain that such slack would be added where needed most to compensate for SIPP's shortcomings. In summary, our results on SIPP accuracy give call-center managers and designers fair warning that staffing levels suggested by the industry standard approach are quite likely to be inadequate.

Third, our results indicate that the same approximation for the lag used in the 24/7 case is good for many limited hours systems as well and that the Lag approach is a reliable and simple-to-implement alternative for those scenarios in which SIPP is likely to be unreliable. Our findings suggest the following implementation guidelines:

1. SIPP can be reliably used for single-peak systems when RA is low, i.e., $RA = 0.1$ planning periods are short, i.e., $PP = 0.25$ or 0.5, and service rate is high, i.e., $\mu > 4$. For longer operating windows, i.e., $W = 18$, the area of reliability can be extended to include $PP = 1$ and $\mu > 2$.

2. For two-peak systems, SIPP can also be reliably used when RA is low, planning periods are short, and service rate is high, except when the operating window is short, i.e., $W = 8$. In this latter case, SIPP is only reliable if, in addition, the offered load is not too high, i.e., $\rho \leq 32$.

3. Lag Avg is generally reliable for single-peak systems when planning periods are short and the operating window is longer, i.e., $W \geq 12$. For shorter windows, Lag Avg is not reliable if RA is high, i.e., $RA = 1$. In the case of two-peak demand patterns, Lag Avg is only reliable if RA is low and planning periods are short.

4. In all other cases, Lag Max will be significantly more reliable.

As a simple and safe guideline for practitioners, we recommend that, as in the continuously operating case, Lag Avg be used for low values of RA and short planning periods, while Lag Max be used for all other situations. Of course, Lag Max will typically use more staff-hours than SIPP or Lag Avg. However, our results indicate that the required increase in staffing is modest in most cases. A call-center manager may want to consider the tradeoff between higher labor costs and what may sometimes be infrequent or small violations of a target service level. In situations in which even a small percentage increase in staff hours may be considered too costly, managers would be well advised to closely examine the tradeoffs between using the Lag Avg and Lag Max methods. The methodology described in this paper could be easily adapted to assess almost any real situation.

Finally, we observe that while this work clarifies and solves an important aspect of the dynamics of call-center staffing, there are other serious call-center staffing issues that merit more research. We mention two that have attracted our attention:

FORECASTING. Call-center staffing models, including SIPP and our lagged SIPP alternatives, presume that there is, in effect, a perfect forecast of the periodic mean customer arrival rate and that all deviations from that forecasted mean arrival rate are encompassed by the variability of a Poisson process. The staffing levels are chosen as insurance against the upper tails of the resulting delay distribution. Clearly, this is not true in practice. Examination of the deviations of actual customer arrivals from forecasted mean arrival rates in many call centers shows that the variability is substantially above that predicted by a Poisson process. Note that we do not speak here about deviations caused by extraordinary events such as new product offerings, a stock market crash, public emergencies, or the like. Rather, this is "ordinary" variation from forecasts. So, what is needed for adequate call-center staffing are models that contain both the deviations of the actual mean demand from forecasts as well as the variation inherent in the Poisson process itself.

WORK FLOW COMPLEXITY. The standard queuing models used in telephone call-center analysis assume that all customers, all jobs, and all servers are statistically identical. Increasingly, this is not true in modern telephone call centers. Customers are broken into classes depending on their importance or on the nature of the service they require. The trend to "mass customization" in which customers are offered more variety and choice is increasing this tendency (Pine, 1993). In modern call centers, these customers are then steered, according to their class and requirements, to banks of specialized servers. Service rates typically vary by customer-server class and class-related priority schemes are frequently imposed. When these workflow complexities are ignored, as they are in simple Erlang models, there is again a tendency for the model to understaff. (The disaggregated real world system is less efficient than the aggregated model, see e.g., Kolesar and Green 1998.) The usual analytic prescription for such complex systems is simulation, but even given the tremendous advances in computing speed and the increased ease in building simulation models, such an approach remains an awkward tool for tactical and strategic decision-making. The call center industry would benefit from a modeling approach that is intermediate between the two extremes of Erlang's simplicity and simulation's complexity. What is needed is a set of staffing models that are able to capture at least the essence of this workflow of complexity. Some work that has addressed some of these complexities under the assumption of a stationary arrival stream include Gans and van Ryzin (1997), Shumsky (1999), and Aksin and Harker (2000).

References

AGNIHOTRI, S. R. AND P. F. TAYLOR (1991), "Staffing a Centralized Appointment Scheduling Department in Lourdes Hospital," *Interfaces,* 21, 5, 1–11.

AKSIN, O. Z., AND P. T. HARKER (2000), "Computing Performance Measures in a Multi-Class Multi-Resource Processor-Shared Loss System," *European Journal of Operational Research,* 123, 61–72.

BARDHAN, I. (1993), "Diffusion Approximations for GI/M/s Queues with Service Interruptions," *Operations Research Letters,* 13, 3, 175–182.

COOPER, R. B. (1972), *Introduction to Queueing Theory*, Macmillan Co., New York.

EICK, S. G., W. A. MASSEY, AND W. WHITT (1993), "$M_t/G/\infty$ Queues with Sinusoidal Arrival Rates," *Management Science,* 39, 2, 241–252.

FEDERGRUEN, A. AND L. GREEN (1986), "Queueing Systems with Service Interruptions," *Operations Research*, 34, 5, 752–768.

——— AND ——— (1989), "Queueing Systems with Service Interruptions II," *Naval Research Logistics*, 35, 3, 345–358.

FELLER, W. (1968), *An Introduction to Probability Theory and Its Applications, Volume I, third ed.*, John Wiley & Sons, New York.

GANS, N. AND G. VAN RYZIN (1997), "Optimal Control of a Multi-Class, Flexible Queueing System," *Operations Research*, 45, 5, 677–693.

GREEN, L. AND P. KOLESAR (1997), "The Lagged PSA for Estimating Peak Congestion in Markovian Queues With Periodic Arrival Rates," *Management Science*, 43, 1, 1234–5678.

——— AND ——— (1998), "A Note on Approximating Peak Congestion in M/G/∞ Queues With Sinusoidal Arrivals," *Management Science*, 44, 11, Part 2 of 2, S137–S143.

———, ———, AND A. SVORONOS (1991), "Some Effects of Nonstationarity on Multiserver Markovian Queuing Systems," *Operations Research*, 39, 3, 502–511.

———, ———, AND J. SOARES (2001), "Improving the SIPP Approach for Staffing Service Systems that have Cyclic Demands," *Operations Research*, 49, 4, 549–564.

GROSS, D., AND C. M. HARRIS (1985), Fundamentals of Queueing Theory, 2nd ed., John Wiley & Sons, New York.

INGOLFSSON, A., M. A. HAGUE, AND A. UMNIKOV (2002), "Accounting for Time-Varying Queueing Effects in Tour Scheduling," *European Journal of Operational Research*, forthcoming 2002.

INGOLFSSON, A. AND E. CABRAL (2002). "Combining Integer Programming and the Randomization Method to Schedule Employees," Research Report Number 02-1, Faculty of Business University of Alberta, 2002.

JAGERMAN, D. L. (1975), "Nonstationary Blocking in Telephone Traffic," *Bell System Technical Journal*, 54, 625–661.

JENNINGS, O. B., A. MANDELBAUM, W. A. MASSEY, AND W. WHITT (1996), "Server Staffing to Meet Time-Varying Demand," *Management Science*, 42, 10, 1383–1394.

KOLESAR, P. J., K. L. RIDER, T. B. CRAYBILL, AND W. W. WALKER (1975), "A Queuing-Linear Programming Approach to Scheduling Police Patrol Cars," *Operations Research*, 23, 6, 1045–1062.

KOLESAR, P. J. AND L. V. GREEN (1998), "Insights on Service System Design from a Normal Approximation to Erlang's Delay Formula," *Production and Operations Management*, 7, 3, 282–293.

MEHROTRA, V. (1997), "Ringing Up Big Business: A Look Inside the Call Center Industry," *ORMS Today*, 24, 4, 18–24.

PINE, B. J. II (1993), *Mass Customization: The New Frontier in Business Competition*, Harvard Business School Press, Boston, MA.

SEGAL, M. (1974), "The Operator Scheduling Problem: A Network Flow Approach," *Operations Research*, 22, 4, 808–823.

SHUMSKY, R. (1999), "Approximation and Analysis of a Queueing System with Flexible and Specialized Servers," Simon School Working Paper, OP-99-2.

Linda V. Green is the Armand G. Erpf Professor at the Graduate School of Business, Columbia University. She is the author of dozens of papers primarily dealing with the design and management of stochastic service systems. Her early path-breaking work led to the development of a model for the dispatching and allocation of emergency vehicles which became the foundation for a patrol car allocation model that is used by most major cities in the U.S. as well as several other countries. More recently, she has been the co-author of numerous publications on service systems with time-varying arrivals. This paper is one of a series coauthored with Peter Kolesar and others on the use of simple stationary approximations for queueing systems with cyclic demand patterns. Her other current major area of research is examining the efficiency and effectiveness of capacity planning and management policies in health care delivery systems, particularly hospitals.

Peter J. Kolesar is Professor of Operations Management in the Columbia Business School where he is also the Research Director of the Deming Center for Quality Management. He won the 1975 Lanchester Prize of ORSA and the 1976 NATO Systems Science Prize for his research on the deployment of police patrol cars and fire engines. Professor Kolesar's current technical research is on efficient staffing in service systems that experience time varying and random customer demand patterns. His managerial research studies the effectiveness of total quality management programs, particularly of the "six sigma" type in U.S. industry and he has twice been an examiner for the Malcolm Baldrige National Quality Award. He has been a consultant to many firms in the U.S. and abroad on quality and productivity, frequently at the CEO level and has served on the editorial boards of *Operations Research, Management Science, and Interfaces.* He is the author of some 60 papers on a range of topics in management science and operations management.

João Soares is assistant professor in the Departamento de Matematica of the University of Coimbra, Portugal. He completed his Ph.D. at Columbia University where his research focused on combinatorial optimization. His current research includes work on the optimal staffing of service systems and on the numerical solution of differential equations.