IMPROVING THE SIPP APPROACH FOR STAFFING SERVICE SYSTEMS THAT HAVE CYCLIC DEMANDS

LINDA V. GREEN

Columbia Business School, 3022 Broadway, Room 423, New York, New York 10027, Lvg1@columbia.edu

PETER J. KOLESAR

Columbia Business School, 3022 Broadway, Room 408, New York, New York 10027, pjk4@columbia.edu

JOÃO SOARES

Department of Mathematics, University of Coimbra, Coimbra, Portugal, jsoares@mat.uc.pt

(Received January 1998; revisions received December 1998, June 1999, October 1999; accepted November 1999)

This paper evaluates the practice of determining staffing requirements in service systems with random cyclic demands by using a series of stationary queueing models. We consider Markovian models with sinusoidal arrival rates and use numerical methods to show that the commonly used "stationary independent period by period" (SIPP) approach to setting staffing requirements is inaccurate for parameter values corresponding to many real situations. Specifically, using the SIPP approach can result in staffing levels that do not meet specified period by period probability of delay targets during a significant fraction of the cycle. We determine the manner in which the various system parameters affect SIPP reliability and identify domains for which SIPP will be accurate. After exploring several alternatives, we propose two simple modifications of SIPP that will produce reliable staffing levels in models whose parameters span a broad range of practical situations. Our conclusions from the sinusoidal model are tested against some empirical data.

Managers of service establishments in which the timing of customer demands for service is random and cyclic commonly adjust staffing levels in an attempt to provide a uniform level of service at all times. Examples include staffing of toll plazas (Edie 1954), airline ground services (Stern and Hersh 1980, Holloran and Byrne 1986, Brusco et al. 1995), tele-retailing (Andrews and Parsons 1989), banking (Brewton 1989), telecommunications (Segal 1974, Sze 1984), hospitals (Agnihothri and Taylor 1991), police patrol (Larson 1972, Kolesar et al. 1975, Green and Kolesar 1984, Taylor and Huxley 1989), and newspapers (Gopalakrishnan et al. 1993). The 800-number telephone call centers that increasingly provide a diversity of customer service and marketing functions are systems for which such staffing issues are important (Brigandi et al. 1994).

Developing specific staffing schedules in such service systems can also be difficult because implementations must take into account complex scheduling constraints. These include honoring employees' preferred start times, quitting times, and shift lengths; adhering to legal or policy limits on the number of consecutive hours and/or days worked; restricting the patterns of days off and on duty; providing required lunch and coffee breaks, and the like. Good schedules must also reflect the economic trade-offs that arise from shift-pay differentials, part-time pay, and overtime. A fundamental requirement is that there be enough staff on duty at all times to meet targeted service levels. In applications described in the literature, these staffing requirements are typically determined by first dividing the workday or workweek into "planning periods" such as shifts, hours, quarter-hours, etc. Then a series of stationary queueing models, most often *M/M/s* type models, is constructed, one model for each planning period. Each of these models is independently solved for the minimum number of servers needed to meet the service target in that period. We call this method of setting staffing requirements the *stationary independent period by period* (SIPP) approach. The period-by-period staffing requirements so derived are then used to set actual workforce schedules. In some businesses managers do this heuristically while in others these staffing requirements become the right-hand sides of key constraints in a large optimization model that derives the actual staffing schedule.

Despite its very widespread use, there is reason to suspect that the SIPP approach does not always work well. The assumptions implicit in using a series of stationary queueing models to set staffing requirements are that (1) delays in consecutive planning periods are statistically independent of one another; (2) within each planning period the system achieves steady state; and (3) the arrival rate does not change during the planning period. The extent to which these assumptions are violated and result in inappropriate staffing levels obviously depends on the system parameters as well as on how the staffing levels themselves change from period to period. Our earlier research, on queueing systems in which demand rates are time-varying but staffing

Subject classifications: Service systems, staffing: Use of queueing models. Queueing systems, cyclic: accuracy of stationary models. Call centers.
Area of review: SERVICES & MILITARY

0030-364X/01/4904-0549 $05.00
1526-5463 electronic ISSN

Operations Research © 2001 INFORMS
Vol. 49, No. 4, July–August 2001, pp. 549–564

remains constant, showed that the accuracy of stationary approximation models is strongly dependent on the magnitude of the service rate and on the relative amplitude of the arrival rate (Green et al. 1991, Green and Kolesar 1995). Therefore, in thinking about systems with variable staffing, there was reason to doubt that a period-by-period stationary approach would always correctly estimate the number of servers needed to achieve desired performance. The results described in this paper confirm that these doubts were justified. In particular, in §4 we examine data from an actual financial services call center and show that implementing the SIPP staffing levels in this facility would result in delay probabilities that violate the target in a number of planning periods, sometimes quite dramatically.

While use of the SIPP approach is common in industry, there has been little published research exploring the conditions when using such stationary models of nonstationary environments is reasonable. (A few papers have tested this approach in specific contexts—see Green and Kolesar 1989, Kolesar et al. 1975, and Kwan et al. 1988.) The potential problems of using a stationary model for a system with time-varying arrivals have been recognized by Sze (1984) and by Thompson (1993). Sze (1984) explored a specific operator staffing application in which demands clustered around the hour and half-hour and used a smoothing function to deal with this "lumpy" demand pattern. Thompson (1993) proposed the use of a "modified" customer arrival rate to compensate for the dependency between consecutive planning intervals. This modified arrival rate is calculated algorithmically in each planning period to estimate the number of customers who actually place a demand during the current period. He found that this method performs better than the traditional SIPP type approach.

In this paper, we numerically explore several issues related to determining staffing requirements in cyclic service systems. First, we identify how the system parameters affect SIPP reliability. Next, we determine specific situations in which SIPP is safe to use and those in which it is not. We then suggest and test simple modifications of SIPP that can improve its performance. Finally, we explore the impact of the choice of method on total staffing requirements. Our results demonstrate that reliable staffing levels can be obtained for a broad range of practical situations with the use of just two of these SIPP modifications. Because SIPP has been integrated into commercially available management support software packages that are used "off-the-shelf," our findings have practical implications.

This work is a continuation of our research on understanding the effects of nonstationarity in queues and on how and when to use simple stationary models in managing nonstationary systems (Green et al. 1991; Green and Kolesar 1991, 1995, 1997). However, whereas our previous work only considered queues with constant server staffing, this paper focuses on service systems in which the number of servers is varied over time in an attempt to better meet the changing rate of arrivals. Because our goal is to provide general managerial insights, we analyze the following somewhat simple scenario: Customer demands are random and periodic over a 24-hour day, as is the case in many of the applications mentioned above. There are a fixed number of equal-length, nonoverlapping planning periods, the service rate is constant, and the target level of service performance remains fixed over the day. We assume that management's goal is to minimize the number of staff-hours required over the day while meeting the targeted service level during each planning period.

This is the set of assumptions used to determine staffing in several actual service contexts including many police patrol systems (Green and Kolesar 1989) where police officers are assigned to one of three nonoverlapping shifts. In other applications where the planning period is shorter, e.g., one hour, and work shifts can be overlapping, (Kolesar et al. 1975, Segal 1974, etc.) this approach would be the first step in constructing work schedules that meet or exceed the requirements in every period. More generally, our objective is to accurately generate appropriate service constraints that could be used in workforce planning and scheduling models.

We describe the model and our methodology in more detail in §1. In §2, we present findings on how SIPP reliability depends on the various system parameters, and we identify regions in which SIPP will work well and others where its performance is clearly unacceptable. Section 3 examines simple modifications to SIPP that improve its reliability, and also identifies best choices for various parameter domains based on performance and cost. We examine SIPP and its variants using empirical data in §4 and offer concluding remarks and directions for further research in §5.

1. MODEL AND METHODOLOGY

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

$$\lambda(t) = \lambda + A\sin(2\pi t/24), \tag{1}$$

where λ is the average arrival rate over the period and $A > 0$ is the amplitude. The other model parameters are μ, the service rate and $s(t)$, the number of servers on duty at time t. We set the period of the sine function at 24 hours because of the many practical applications in which a daily cycle is evident. Those interested in periods of different length can, with modest effort, scale our results as desired.

Let $p_n(t)$ be the periodic steady-state probability that n customers are in the system at time t. These functions are the foundation of our results and are obtained by numerically solving the following standard set of differential equations that describe the system; see Gross and Harris (1974):

$$\begin{aligned}
p_0'(t) &= -\lambda(t)p_0(t) + \mu p_1(t), \\
p_n'(t) &= \lambda(t)p_{n-1}(t) + (n+1)\mu p_{n+1}(t) \\
&\quad -(\lambda(t) + n\mu)p_n(t), \qquad 1 \leqslant n < s(t), \\
p_n'(t) &= \lambda(t)p_{n-1}(t) + s(t)\mu p_{n+1}(t) \\
&\quad -(\lambda(t) + s(t)\mu)p_n(t), \qquad n \geqslant s(t).
\end{aligned} \tag{2}$$

Details on our numerical analysis methods are given in Green et al. (1991). We assume that the system operates continuously over an infinite time horizon, and we consider its long-run behavior. In this paper we focus on the probability of delay as the main performance measure of interest. 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 epoch t and is given by

$$p_D(t) = 1 - \sum_{n=0}^{s(t)-1} p_n(t). \tag{3}$$

The principal output from our differential equation solver (simulator) is a vector of 288 estimates of $p_D(t)$ made at 5-minute intervals over the cycle.

The analytic sequence for each scenario is as follows:

(i) Fix the scenario's exogenous parameters: λ, the mean arrival rate; μ, the service rate; and $RA = A/\lambda$, the relative amplitude. Fix the managerial parameters: τ, the target probability of delay; and PP, the length of the planning period.

(ii) Divide the cycle into nonoverlapping intervals of length PP, starting at $t = 0$. For each planning interval compute the average arrival rate by integrating Equation (1) over the planning interval. Then use this average arrival rate, the service rate, and an iterative version of the Erlang delay equation (Cooper 1972, p. 100) to find the minimum staffing needed in the interval to achieve the target delay probability τ. This produces a vector of staffing levels $\{s(t), t = 1, 24/PP\}$.

(iii) Run the simulator with the exogenous parameters specified as in (i) and the $\{s(t)\}$ as determined in (ii). This produces the output vector $\{p_D(t)\}$ mentioned above.

(iv) Using the vector $\{p_D(t)\}$, compute various summary performance measures including the 24-hour average probability of delay, the instantaneous (5-minute) maximum probability of delay, the maximum of the half-hour average probabilities of delay, the number of half-hours in which the average probability of delay exceeds the target, and the number of half-hours in which the average probability of delay exceeds the target by at least 10%.

Some explanation of the summary choices suggested in (iv) is in order. First, we observe that there is no "right way" to evaluate SIPP performance relative to the target. Essentially, one wishes to compare the actual $p_D(t)$ curve to the target value. In a good solution the curves will be "close." But by what measure? If the performance target is taken literally and strictly, the most appropriate measure is arguably the number of planning periods in which the average of $p_D(t)$ doesn't exceed the target. But, there are several problems in using such a measure. First, to be able to make comparisons of systems with planning periods of different lengths, a consistent measure is needed. Second, when planning periods are long, e.g., 8 hours, although the average probability of delay over the period may be within the target, it is possible for $p_D(t)$ to exceed the target during a large fraction of the planning period. Moreover, our practical experience suggests that to most managers the target is not a strict constraint, it is typically a somewhat subjectively chosen "goal." For these reasons, although we computed all the summary statistics mentioned in (iv), as well the equivalent measures using hours and quarter-hours rather than half-hours, we focus primarily on the following measure: *the number of half-hours in which the target is exceeded by at least 10%*. We observe that this measure of reliability is an intermediate value for comparing systems with planning periods ranging from 1/4 hour to 2 hours in length. It is also a more conservative choice than one hour for measuring reliability in that it is a tougher metric of performance, and a half-hour appears to be a commonly used planning period length in many actual implementations of SIPP.

We analyze models of service systems that span a number of actual situations we have experienced personally or have encountered in the literature. However, our analysis has also been limited by the capabilities of our numerical analysis routine and our computing facilities. A scenario (model) is characterized by the following five parameters:

- The service rate, μ
- The average arrival rate, λ
- The relative amplitude, $RA = A/\lambda$
- The target probability of delay, τ
- The length of the planning period, PP

An important derived measure is $\rho = \lambda/\mu$, the average number of servers that are busy and an important measure of system "size." Our core set of models considered service rates starting at a low of $\mu = 2$, that is with average service times as long as 30 minutes, doubling up to 64 ($\mu = 2, 4, 8, 16, 32, 64$). We considered average customer arrival rates starting at a low of $\lambda = 4$ customers per hour and doubling up to 4096 ($\lambda = 4$, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096.) Not all of the 66 (μ, λ) combinations implied by the above were either computationally feasible (when $\lambda \gg \mu$ 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 36 core (μ, λ) combinations contained in Table 1. Note that this set of combinations covers a range of ρ from 2 to 64 increasing by doubles. While this set of combinations covers a very broad range and contains many practical cases, we do

Table 1. The core mu and lambda combinations analyzed.

	Average Arrival Rate, Lambda										
	Service Rate										
Mu	4	8	16	32	64	128	256	512	1024	2048	4096
2	X	X	X	X	X	X					
4		X	X	X	X	X	X				
8			X	X	X	X	X	X			
16				X	X	X	X	X	X		
32					X	X	X	X	X	X	
64						X	X	X	X	X	X

X = an included mu, lambda combination

not contend that it covers all regions of possible interest. It is also important to note that the scenarios are not spread uniformly through the experimental region. In our examination of the reliability of SIPP and its variants, we considered three relative amplitudes: $RA = 0.1, 0.5$, and 1.0; three probability delay targets: $\tau = 0.05, 0.10$, and 0.20; and four planning period lengths: $PP = 0.25, 0.05, 1.0$, and 2.0 hours. Thus, in evaluating SIPP we considered 1,296 scenarios. Additional scenarios were used to explore issues concerning the effect of the number and timing of planning periods and in the early stages of our work.

2. THE ACCURACY OF THE SIPP APPROACH

What are the factors that influence SIPP reliability, and what is the direction of their influence? When does SIPP specify "safe" staffing levels? These are the questions we address in this section.

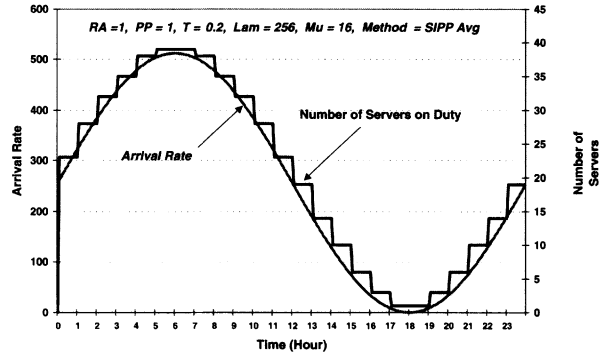
2.1. An Example Where SIPP Fails

Before describing the results of our analysis, we present some details for a hypothetical situation for which SIPP does not produce satisfactory staffing requirements. Consider an incoming telephone call center that is open 24 hours a day. The average length of calls is about 4 minutes, and the average call rate over the day is 250 calls per hour, with a peak of about 500 calls per hour in the early afternoon and a drop to almost zero in the middle of the night. The call center management uses one-hour planning periods and has set a service target of 20% probability of delay. This situation could be reasonably approximated by one of the models we studied. It has the following parameters: $\lambda = 256$, $\mu = 16$, $RA = 1$, $PP = 1$, $\tau = 0.2$. The standard SIPP method for this system suggests staffing levels (shown in Figure 1) that result in an actual 24-hour average probability of delay of 0.18—which meets the target on average. However, the instantaneous peak probability of delay is over 0.44, and the service target is exceeded in 16 of the 48 half-hour planning periods of the day. In 11 of these periods, the target is exceeded by more than 10%. Thus, the actual performance of this system when staffed according to SIPP will be considerably worse than desired. (See Figure 2 for a plot of the actual probability of delay curve.) It is important to note that this example is not a "worst case." There are, as we shall see, other parameter choices corresponding to practical situations in which SIPP performance is far worse. We will return to this example later in the paper and discuss alternatives to improve performance. We shall see that some SIPP variants perform well here and that this scenario is, in fact, a borderline case for which a change in any one of the system parameters could significantly improve or further degrade SIPP reliability.

2.2. The Factors Affecting SIPP Reliability

Based on our earlier work, we hypothesized that the standard SIPP approach would be particularly risky when the arrival rate changes substantially during the planning

Figure 1. Arrival rate and SIPP staffing in the sinusoidal model.



period. This happens when the period lengths are long (relative to the cycle length) or when the relative amplitude is high. We further hypothesized that particularly for short planning periods, the accuracy of the SIPP approach should also depend on the same factors that determine the accuracy of what we have called the *pointwise stationary approximation* (PSA) for nonstationary queues with a constant number of servers (Green and Kolesar 1991). The PSA models the behavior of the nonstationary queueing system at each point in time by a stationary model with the arrival rate at that epoch, and thus is related to using the SIPP approach. Our previous research showed that two critical factors determine the accuracy of the PSA: the service rate and the relative amplitude of the arrival process. The numerical results we obtained here support these hypotheses and, in addition, revealed another important factor—ρ, the size of the system.

Table 2 summarizes the results of the 1,296 simulations over the parameter domain described above. Note that the experimental design is completely balanced in relative amplitude, planning period length, service rate, delay target, and presented load (ρ) in that every parameter combination is included. For each scenario the table contains our main reliability measure—the count of the number of half-hours in which the average 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.

Figure 2. Probability of delay with SIPP staffing in the sinusoidal model.

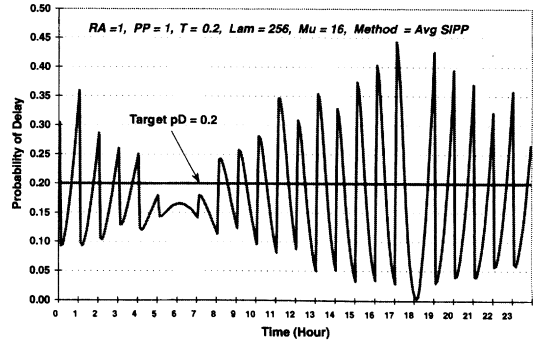


Table 2. SIPP reliability: the number of half-hours in which pD exceeds 110% of target.

			Rho, Mu																													
			2						4						8						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	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	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	0	0	0	0	0	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
			0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	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	0	0	0	0	0	0	0	0	0	0	0	4	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	0	0	0	0	0	0	0	0	0	0	0	3	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	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	0	0	0	1	1	0	0	0	0	8	2	1	0	0	0
			0.1	0	0	0	0	0	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	3	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	0	0	0	0	0	1	0	0	0	0	0	5	1	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	0	0	0	0	0	4	1	0	1	1	1	11	8	7	8	8	8
			0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	1	1	1	10	8	5	7	6	7
			0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	1	1	5	3	3	3	4	4
0.5	0.25		0.05	0	0	0	0	0	0	1	0	0	0	0	0	5	0	0	0	0	0	16	0	0	0	0	0	19	11	0	0	0	0	21	18	5	0	0	0
			0.1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	14	0	0	0	0	0	18	8	0	0	0	0	22	17	5	0	0	0
			0.2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	15	1	0	0	0	0	19	9	0	0	0	0	22	17	7	0	0	0
	0.5		0.05	0	0	0	0	0	0	2	0	0	0	0	0	6	2	0	0	0	0	15	3	0	0	0	0	19	14	1	0	0	0	21	18	9	1	0	0
			0.1	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	12	4	0	0	0	0	18	11	0	0	0	0	22	19	7	0	0	0
			0.2	0	0	0	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0	16	2	0	0	0	0	20	11	0	0	0	0	21	14	7	0	0	0
	1		0.05	1	0	0	0	0	0	2	0	0	0	0	0	7	2	1	2	2	2	14	6	5	5	4	4	18	8	8	7	11	10	19	10	10	14	16	16
			0.1	0	0	0	0	0	0	1	0	0	0	0	0	5	2	0	0	0	1	15	6	2	2	1	2	19	10	6	7	5	6	22	16	10	13	13	12
			0.2	0	0	0	0	0	0	1	0	0	0	0	0	7	2	0	0	1	1	8	2	0	0	0	0	18	7	5	4	7	8	22	14	8	11	15	16
	2		0.05	4	3	3	3	2	2	3	2	2	2	2	2	9	7	7	9	9	9	14	13	13	12	13	13	18	18	17	19	19	19	18	19	18	20	21	21
			0.1	0	0	0	0	0	0	4	3	3	3	2	2	9	7	6	6	6	8	12	11	12	11	12	12	16	17	15	15	18	19	21	20	19	20	21	22
			0.2	0	0	0	0	0	0	6	1	2	1	3	3	11	5	4	5	5	6	13	11	10	12	12	12	16	16	15	17	17	17	23	20	17	19	19	19
1	0.25		0.05	4	0	0	0	0	0	12	1	0	0	0	0	18	8	0	0	0	0	22	17	4	0	0	0	23	19	13	0	0	0	26	22	19	4	0	1
			0.1	2	0	0	0	0	0	10	0	0	0	0	0	18	4	0	0	0	0	22	17	4	0	0	0	24	21	8	0	0	0	27	22	17	1	0	0
			0.2	2	0	0	0	0	0	8	0	0	0	0	0	17	2	0	0	0	0	21	15	2	0	0	0	25	21	12	0	0	0	27	22	17	2	0	0
	0.5		0.05	6	2	0	0	0	0	11	3	0	0	0	0	18	7	1	0	0	0	23	15	6	1	1	2	24	21	15	2	0	3	26	22	20	14	14	17
			0.1	3	1	0	0	0	0	11	3	0	0	0	0	20	9	1	0	0	0	22	16	2	0	0	0	24	20	10	3	1	1	27	22	18	9	3	5
			0.2	2	0	0	0	0	0	8	1	0	0	0	0	17	8	1	0	0	0	19	14	1	0	0	0	24	20	10	0	0	0	28	23	19	13	0	2
	1		0.05	7	4	3	2	2	2	14	8	5	8	8	8	19	11	10	8	7	6	20	11	11	9	10	11	24	19	15	19	19	21	26	21	21	21	27	28
			0.1	4	2	0	0	0	0	12	4	1	1	2	1	16	7	4	5	5	6	20	13	10	11	14	12	23	18	13	18	20	20	27	21	23	21	21	24
			0.2	4	0	0	0	0	0	4	1	0	0	0	0	17	8	4	4	2	3	22	15	9	11	13	14	24	19	12	16	18	18	27	23	27	21	21	20
	2		0.05	9	6	4	5	5	4	14	12	12	15	15	14	16	15	13	15	14	13	23	22	19	21	25	26	24	21	23	25	25	25	30	27	31	28	29	29
			0.1	11	10	8	8	9	9	14	12	10	10	10	10	14	13	16	14	13	13	21	19	15	18	18	20	25	22	20	24	23	25	29	23	27	29	28	27
			0.2	5	2	2	2	2	2	12	11	11	11	10	12	16	15	13	13	13	15	21	17	17	17	18	17	26	21	17	20	19	19	30	26	26	26	28	27

Overall, we see in Table 2 that SIPP is reliable for 725 of the 1,296 scenarios (56%). SIPP is reliable in the upper left corner of the table; that is, for all models with low relative amplitude and low presented loads. SIPP is very unreliable in the lower right corner of the table, for models with high relative amplitude and high presented loads. Moreover, in most of these cases in the lower right corner SIPP is very unreliable; it violates the reliability standard in 20 or more of the 48 half-hours of the day. Scanning the entire table, we can conclude that over this experimental domain SIPP is often unreliable and that many practical scenarios fall into the unreliable regions. For an example of a region of practical concern examine the models with relative amplitude of 1.0, presented loads of 16 or more, and one-hour-long planning periods, for all of which SIPP is quite unsatisfactory.

By methodical inspection of Table 2, one can discern how a particular factor influences SIPP reliability when all other factors are fixed. Table 3 summarizes the data in Table 2 in a manner that facilitates an examination of the influence of each parameter individually while the other parameters vary—so-called *main effects*. It contains for each parameter the marginal counts and percentages of the number of reliable scenarios from Table 2. Table 4, also constructed from Table 2, shows two-way cross-tabulations of reliability counts by selected pairs of parameters. This table facilitates an examination of some important "interaction effects."

Now we discuss how each of the parameters influence SIPP reliability. To facilitate discussion of our findings, we define a "case" for any given parameter to be a set of scenarios in which that parameter varies over its range while all other parameters are held fixed. For example, Table 2 contains 432 "relative amplitude cases," one of which is the set $RA = 0.1, 0.5, 1.0$, while $PP = 0.5$, $\tau = 0.1$, $\mu = 4$ and $\rho = 16$. Examination of Tables 2 and 3 shows that:

1. SIPP reliability is always nonincreasing as relative amplitude increases—all other parameters being held constant. This is true for each of the 432 relative amplitude cases that are contained in Table 2. The power of this influence can be seen in that relative amplitude still has a dominant influence even when all other parameters vary. Specifically, 89% of the 432 models with relative amplitude of 0.1 are reliable, while only 28% of the 432 models with relative amplitude of 1.0 are reliable (Table 3). Thus, relative amplitude is a dominant influence on SIPP reliability.

2. SIPP reliability is almost always nonincreasing as the length of the planning period increases—all other parameters being held constant. This is true for all but 39 of the 324 planning period cases contained in Table 2; most of these are low service rate scenarios. The planning period length still has a dominant influence, even when all other parameters vary: 79% of the 324 models with 1/4-hour planning periods are reliable, while only 28% of the 324 models with 2-hour planning periods are reliable (Table 3).

3. SIPP reliability is almost always nondecreasing as the service rate increases—all other parameters being held constant. This is true for all but 60 of the 216 service rate cases

Table 3. Main effects: counts of unreliable models by parameter values: all 1,476 models.

	RA			Mu					Rho					Lambda							
	0.1	0.5	1	2	4	8	16	32	64	1	2	4	8	16	32	64	4	8	16	32	64	128	256	512	1024	2048	4096
Unreliable Models	30	164	277	83	75	71	80	82	80	21	33	53	65	75	99	125	9	17	34	46	55	74	64	62	49	40	21
Percent Unreliable	6.1	33.3	56.3	38.4	29.8	28.2	31.7	32.5	31.7	11.7	15.3	24.5	30.1	24.7	45.8	57.9	12.5	15.7	23.6	25.6	25.5	34.3	35.6	43.1	45.4	55.6	58.3
Total Models	492	492	492	216	252	252	252	252	252	180	216	216	216	216	216	216	72	108	144	180	216	216	180	144	108	72	36

	Target			Planning Pd		
	0.05	0.1	0.2	0.25	0.5	1	2
Unreliable Models	180	155	136	22	40	153	256
Percent Unreliable	36.6	31.5	27.6	6.0	10.8	41.5	69.4
Total Models	492	492	492	369	369	369	369

contained in Table 2, and most of these 60 cases are models with large relative amplitudes and long planning periods. Service rate is still a dominant influence, even when all other parameters vary: only 32% of the 216 models with service rate of 2 are reliable, while 66% of the 216 models with service rate of 64 are reliable (Table 3.)

4. SIPP reliability is almost always nonincreasing as the system size (ρ) increases—all other parameters being held constant. This is true for all but 13 of the 216 system size cases contained in Table 2. Most (8) of these 13 cases are models with 2-hour-long planning periods. Presented load is a strong influence, even when all other parameters vary: 81% of the 216 models with $\rho = 2$ are reliable, while only 31% of the 216 models with $\rho = 64$ are reliable (Table 3.)

5. SIPP reliability tends to decrease as the arrival rate increases, but the effect—most evident for large values—is not consistent or strong. This, too, is confirmed in Table 3.

6. There is a weak relationship between the probability delay target and SIPP reliability. SIPP is somewhat more reliable for less strict targets. (See Table 3). This tendency surprised us because we had hypothesized that SIPP would be more accurate for tougher targets, such as 0.05, where because little queueing occurs, behavior in successive planning periods would be more independent.

Table 4 presents SIPP reliability in a series of two-way tables in the key parameters. These illustrate the "first-order interaction" effects of the parameters on SIPP reliability. Some of these interactions are quite strong. In particular, high relative amplitude and long planning periods are a deadly combination—all the 108 such models that were run were unreliable. This is also true for high relative amplitude and low service rate. For high relative amplitude and large system size, and for long planning periods and large system size, we have similarly strong interaction effects—almost all such models are unreliable.

Before proceeding to interpret these findings we note that the directional conclusions listed above are not artifacts of the particular definition of reliability we used. We have carried out parallel analyses using counts of target exceedences (counts of intervals in which the actual probability of delay is larger than the target) in intervals ranging from 5 minutes to 2 hours in length. (We used both simple exceedances and exceedances of more than 110%.) The directional results are the same. Moreover, other measures such as 24-hour average delays, maximum hourly delays, and maximum instantaneous delays move in the same direction as our standard reliability measure. Thus, we believe that the above results are robust against a particular reliability measure.

2.3. Some Interpretation

While the theory available for the type of $M_t/M/S_t$ systems studied here does not permit us to confirm the above results analytically, theory did lead us to hypothesize a number of the above results in advance of the experiments and led us to our chosen experimental domains. The theory

Table 4. SIPP reliability: important interactions.

	Planning Period (Hours)				
RA	0.25	0.5	1	2	Total
0.1	2	3	10	31	46
0.5	23	29	67	96	215
1	44	63	95	108	310
Total	69	95	172	235	571

108 cases in each cell

	Mu						
RA	2	4	8	16	32	64	Total
0.1	17	7	4	6	6	6	46
0.5	59	45	32	26	26	27	215
1	72	65	54	42	37	40	310
Total	148	117	90	74	69	73	571

72 cases in each cell

	Rho						
RA	2	4	8	16	32	64	Total
0.1	0	0	0	0	17	29	46
0.5	7	24	38	42	49	55	215
1	35	42	51	57	59	66	310
Total	42	66	89	99	125	150	571

72 cases in each cell

Planning Period	Rho						
(Hours)	2	4	8	16	32	64	Total
0.25	3	5	9	13	15	24	69
0.5	5	8	13	18	21	30	95
1	10	17	31	32	40	42	172
2	24	36	36	36	49	54	235
Total	42	66	89	99	125	150	571

54 cases in each cell

also offers plausibility arguments. In Green et al. (1991) we showed that for fixed server queueing systems with sinusoidal Poisson input streams, the average probability of delay is monotone increasing in relative amplitude. Hence, we believe that the first two results are attributed to the fact that the variability in the arrival rate during a planning period increases as either the planning period gets longer or as the relative amplitude increases. A higher value of relative amplitude can also correspond to a more significant violation of the implicit SIPP assumption of independent planning periods. This is because the use of the average arrival rate can underestimate the actual workload in the beginning of the period (and hence at the end of the previous period) for periods during which the arrival rate is decreasing. In such cases, congestion may carry over from the previous planning period that cannot be adequately handled by the staffing determined by the SIPP model.

The service rate impacts SIPP reliability in two ways. First, it determines the speed with which customers are cleared from the system and hence the speed of convergence to steady state. Second, in service systems with con-

stant staffing, the service rate determines the magnitude of the lag between the time of the peak arrival rate and the time of peak congestion, and hence the degree to which the entire delay curve lags the PSA delay curve (which is in phase with the arrival rate curve). Although the concept of "lag" is not crisply defined for variable staffing systems such as studied here, it can be thought of as the amount of time during which the arrival rate at a given epoch will continue to impact system congestion. In constant staffing models, Whitt (1991) showed that as the service rate tends to infinity, the actual probability of delay at any time t, approaches the PSA for probability of delay at time t, and hence the lag approaches 0. Conversely, in Green and Kolesar (1995, 1997), we showed that for small values of μ such as $\mu \leqslant 2$, the lag is significant, e.g., more than a half-hour). Thus, we believe that when planning periods are short and service rates are low, the congestion originating from the arrivals in one period is likely to impact later periods and hence invalidate the implicit SIPP assumption of independence among periods. Moreover, the resulting congestion in planning periods following the peak arrival rate will be greater for higher relative amplitudes, and hence the SIPP approach will be more likely to underestimate the required staffing in these situations.

Higher average arrival rates exacerbate the effect of the relative amplitude. As more customers arrive during the planning periods near the peak of the cycle, they cause increasingly greater congestion than predicted by use of the average during that planning period. And because larger system size corresponds to higher arrival rates and/or lower service rates, it follows that system size will be negatively correlated to SIPP reliability.

2.4. Region of SIPP Reliability

So, in summary when is SIPP "safe" to use? The answer is not simple because, as Tables 2 through 4 show, SIPP accuracy depends upon the values of virtually all the system parameters and on interactions between them. Our analysis suggests that SIPP tends to be safe for systems with low relative amplitudes, short planning periods, high service rates, and small size. The converse is also true; SIPP tends to be unsafe when one or more of the following is true: the system has large relative amplitudes, long planning periods, low service rates, or is large. From Table 2, we can specify particular regions where SIPP is unsafe as well as regions where it is safe.

- SIPP is unsafe whenever $RA = 1$ and planning periods are 2 hours. All the 108 models with these parameter values are unreliable.
- SIPP is unsafe whenever $RA = 1$ and $\mu = 2$. All 72 cases for these parameter settings are unreliable.
- SIPP is almost always unsafe when $RA = 0.5$ and planning periods are 2 hours. The only exceptions occur when $\rho = 2$.
- When $RA = 0.1$, SIPP is safe for planning periods of 0.25 or 0.5 hour whenever $\mu \geqslant 4$; for 1-hour periods it is safe whenever $\mu \geqslant 16$.
- When $RA = 0.5$, SIPP is safe for planning periods of 0.25 or 0.5 hour whenever $\mu \geqslant 32$.

One can sharpen these insights for systems typical of a particular industry or environment by focussing on the appropriate parameter families. As an example we selected a set of parameter values that are broadly descriptive of many call center operations we have seen—specifically, relative amplitude of 1.0, target delay probability of 0.2, and 1/2-hour planning periods. Table 5 contains results for the 36 such models contained in our experimental domain (service rates and sizes of 2, 4, 8, 16, 32, and 64.) To provide a richer view of SIPP performance, the table contains four performance measures:

- The 24-hour average delay probability
- The maximum delay probability over the 288 five-minute intervals of the day
- The maximum delay probability over the 48 half-hour intervals of the day
- The number of half-hour periods for which the service target is exceeded by more than 10%—our standard measure of reliability.

The shaded region—basically, the upper right section of the table—indicates unreliable scenarios. So, *in this practically important environment*, SIPP is safe if the service rate is large relative to the size. A specific rule of thumb is SIPP is safe for call center models when $\rho/\mu = \lambda/\mu^2 < 1$.

3. ALTERNATIVES TO SIPP

The results in the previous section identified shortcomings of the standard SIPP approach. The problems are serious enough to invalidate the use of SIPP in many service systems. The next logical question is, "For systems for which SIPP is unreliable, are there simple alternatives that do better?" In this section, we explore the reliability of three alternative SIPP-based approaches for determining staffing requirements. All are easy and fast to implement.

3.1. SIPP Max

We believe that many of SIPP's reliability problems are because of the use of the planning period average λ to represent the arrival rate over the entire planning period. Because this often leads to understaffing, one potential improvement is to use the maximum value of λ for the planning period instead. We call this modification the SIPP Max method. We ran the same 1296 scenarios described in §2 using the SIPP Max method. Our numerical results, displayed in Table 6 for $RA = 0.5$ and $RA = 1$, reveal the following:

- SIPP Max is always as or more reliable than the standard SIPP (which we henceforth call SIPP Avg) in that it never produces more errors. While SIPP Avg is unreliable for 571 of the 1,296 scenarios run, SIPP Max is unreliable for 111 scenarios.
- SIPP Max is safe whenever $\mu \geqslant 8$.
- When $RA = 0.1$, SIPP Max is always safe.

Table 5. SIPP reliability for some call center models.

Mu		Rho 2	4	8	16	32	64
2	24-hr Avg pD	**0.12**	**0.14**	**0.16**	**0.17**	**0.20**	**0.25**
	Max pD (5 min)	**0.28**	**0.37**	**0.54**	**0.60**	**0.87**	**0.99**
	Max PD (1/2 hr)	**0.23**	**0.31**	**0.44**	**0.52**	**0.82**	**0.97**
	Periods Above 110% of Target	**2**	**8**	**17**	**19**	**24**	**28**
4	24-hr Avg pD	0.12	**0.14**	**0.16**	**0.17**	**0.18**	**0.21**
	Max pD (5 min)	0.23	**0.28**	**0.36**	**0.39**	**0.56**	**0.83**
	Max PD (1/2 hr)	0.20	**0.23**	**0.29**	**0.31**	**0.43**	**0.69**
	Periods Above 110% of Target	0	**1**	**8**	**14**	**20**	**23**
8	24-hr Avg pD	0.12	0.14	**0.16**	**0.16**	**0.18**	**0.19**
	Max pD (5 min)	0.21	0.25	**0.29**	**0.30**	**0.38**	**0.57**
	Max PD (1/2 hr)	0.18	0.20	**0.23**	**0.22**	**0.27**	**0.37**
	Periods Above 110% of Target	0	0	**1**	**1**	**10**	**19**
16	24-hr Avg pD	0.12	0.14	0.16	0.16	0.18	**0.19**
	Max pD (5 min)	0.21	0.24	0.27	0.27	0.33	**0.45**
	Max PD (1/2 hr)	0.18	0.19	0.20	0.19	0.22	**0.25**
	Periods Above 110% of Target	0	0	0	0	0	**13**
32	24-hr Avg pD	0.12	0.14	0.16	0.16	0.18	0.19
	Max pD (5 min)	0.22	0.24	0.27	0.28	0.33	0.43
	Max PD (1/2 hr)	0.18	0.18	0.20	0.18	0.20	0.22
	Periods Above 110% of Target	0	0	0	0	0	0
64	24-hr Avg pD	0.12	0.14	0.16	0.16	0.18	**0.19**
	Max pD (5 min)	0.23	0.25	0.28	0.30	0.36	**0.47**
	Max PD (1/2 hr)	0.18	0.18	0.20	0.19	0.20	**0.24**
	Periods Above 110% of Target	0	0	0	0	0	**2**

Models with RA = 1/2, Target = 0.2, Bold type = **Unreliable**

• When $RA = 0.5$, SIPP Max is safe for $\mu \geqslant 4$ when planning periods are half-hour or longer.

Another way of stating the above is that SIPP Max tends to be unreliable for low service rates, short planning periods, and high relative amplitudes. This is not surprising in light of the theory described in the previous section. For small service rates such as $\mu = 2$, the lag between the arrival rate curve and the delay curve is long enough so that the use of the maximum λ during the planning period will not necessarily capture the demand rate that is responsible for the congestion during that period. This is particularly true for shorter planning periods. This deficiency will clearly be worse for higher values of relative amplitude.

Because in every planning period SIPP Max uses the same or more staff as SIPP Avg for the same scenario, its $p_D(t)$ curve will always be lower than or equal to that for SIPP Avg. As a consequence, SIPP Max results in a smaller value for the maximum epoch probability of delay as well as for the 24-hour average probability of delay and the other measures we used for comparison. Of course, the increased reliability of SIPP Max comes with a cost: higher staffing requirements. For the cases we studied, the average difference in suggested staffing is 10.6% and the largest difference is 15.3%. However, in the cases for which both methods are reliable, the maximum difference is about 8%. To illustrate these differences, we revisit our opening example in which $RA = 1$, $\lambda = 256$, $\mu = 16$, $\tau = 0.2$, and planning periods are 1 hour. As noted previously, this is a case for which SIPP Avg performs badly. Figure 3 shows the actual $p_D(t)$ curve using the SIPP Max method. Note that using SIPP Max results in the target being met at all times during the cycle. The required staffing relative to SIPP Avg in this case is 532 servers vs. 496 servers—an increase of 7.25%.

3.2. SIPP Mix

Is there a compromise between SIPP Avg and SIPP Max, i.e., a method that has the same reliability as SIPP Max but with lower staffing requirements? Because of the lag between the arrival rate and the resulting delay, SIPP Avg is more likely to understaff when the arrival rate is decreasing (see, for example, Figure 2). Thus, using the planning

Figure 3. Probability of delay with SIPP Max staffing in the sinusoidal model.

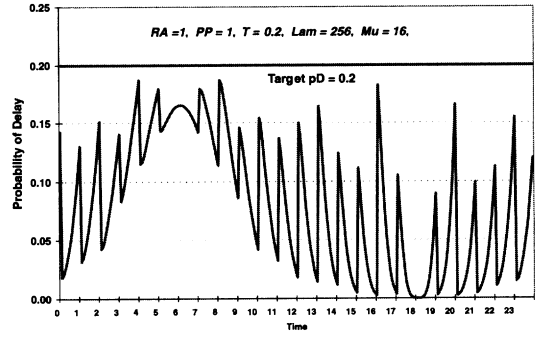


Table 6. SIPP max reliability: the number of half-hours in which pD exceeds 110% of target.

			Rho, Mu																															
			2						4						8						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	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64
0.5	0.25	0.05	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	16	0	0	0	0	0	17	2	0	0	0	0	21	13	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	13	0	0	0	0	0	18	2	0	0	0	0	20	12	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	18	5	0	0	0	0	20	13	0	0	0	0
	0.5	0.05	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	9	0	0	0	0	0	14	0	0	0	0	0	20	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	6	0	0	0	0	0	13	0	0	0	0	0	18	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	4	0	0	0	0	0	15	0	0	0	0	0	18	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	2	0	0	0	0	0	2	0	0	0	0	0	8	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	0	0	0	0	0	1	0	0	0	0	0	8	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	0	0	0	0	0	3	0	0	0	0	0	7	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	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	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	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	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	7	0	0	0	0	0	17	0	0	0	0	0	21	4	0	0	0	0	23	18	0	0	0	0	25	21	0	0	0	0
		0.1	2	0	0	0	0	0	6	0	0	0	0	0	17	0	0	0	0	0	20	4	0	0	0	0	22	16	0	0	0	0	25	20	0	0	0	0
		0.2	1	0	0	0	0	0	3	0	0	0	0	0	17	0	0	0	0	0	20	3	0	0	0	0	23	15	0	0	0	0	25	20	0	0	0	0
	0.5	0.05	1	0	0	0	0	0	4	0	0	0	0	0	12	0	0	0	0	0	17	0	0	0	0	0	21	1	0	0	0	0	22	9	0	0	0	0
		0.1	0	0	0	0	0	0	2	0	0	0	0	0	9	0	0	0	0	0	15	0	0	0	0	0	20	3	0	0	0	0	22	6	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	17	0	0	0	0	0	19	0	0	0	0	0	22	6	0	0	0	0
	1	0.05	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	8	0	0	0	0	0	9	0	0	0	0	0	12	2	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	4	0	0	0	0	0	9	1	0	0	0	0	12	2	0	0	0	0
		0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	9	0	0	0	0	0	11	0	0	0	0	0
	2	0.05	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	4	0	0	0	0	0	5	0	0	0	0	0	6	0	0	0	0	0
		0.1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	5	1	0	0	0	0	5	1	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	0	0	0	0	3	0	0	0	0	0	5	0	0	0	0	0

period maximum arrival rate should be most cost effective for those planning periods in which the arrival rate is either strictly decreasing or reaches its maximum value. Therefore, we explored what we call SIPP Mix, which uses the average planning period arrival rate for periods in which the arrival rate is strictly increasing, and the maximum planning period arrival rate otherwise. Our results indicate that although SIPP Mix is a better choice than SIPP Max in some situations, overall it is not as reliable. Specifically:

- When $RA \leqslant 0.5$, SIPP Mix is almost always as reliable as SIPP Max when planning periods are short, e.g., 0.25 or 0.5 hour.
- SIPP Mix is much less reliable than SIPP Max for longer planning periods, particularly as λ and μ increase.
- When $RA = 1$, even for shorter periods, SIPP Mix may be unreliable for large values of both λ and μ.

In Green et al. (1991), we showed that probability of delay is monotone increasing in event frequency (as measured by both λ and μ) as well as relative amplitude. Thus, the above results can be interpreted as due to the increasing inadequacy of the use of the average arrival rate to estimate actual congestion during any planning period as event frequency increases. This inadequacy is compounded as the variability of the arrival rate during the period increases, i.e., for longer planning periods or higher relative amplitude.

3.3. A Lagged SIPP Approach

Our results show that SIPP Max is the most reliable of the three SIPP methods we explored, i.e., there are no scenarios for which SIPP Max is not reliable and one of the others is reliable. Therefore, from Table 6, which gives the results for SIPP Max, we see that none of the three SIPP methods works for some cases when $\mu = 2$ or 4. This is likely the result of the previously described lag between the arrival rate curve and the probability of delay curve. Borrowing from our work on estimating peak congestion in service systems with time-varying demands and a constant number of servers (Green and Kolesar 1997), we explored a fourth method for estimating staffing requirements: Lag SIPP. The idea of this method is to estimate L, the actual lag that would exist if the number of servers were constant, by an approximation based on a nonstationary infinite server model. Then, instead of basing the staffing during a planning period on the average (or maximum) arrival rate during the planning period, we use an average (or maximum) arrival rate calculated from shifting (advancing) the $\lambda(t)$ curve by L units. So, for example, if $L = 0.5$ 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 $[t_0-0.5, t_0+0.5]$.

For the infinite server model with sinusoidal arrival rate and exponential service, Eick et al. (1993) showed that the lag (in server occupancy) is solely a function of the service rate and is given by

$$L = (\cot^{-1}(\mu/\gamma))/\gamma, \tag{4}$$

where $\gamma = 2\pi/24$, in our case of a cycle time of 24 hours. In Green and Kolesar (1998) we proposed an approximation to (4) of $L \approx 1/\mu$ for infinite server systems and showed that the error using this approximation is less than 1% for values of $\mu \geqslant 2$. Using this infinite server lag to predict delays in a finite server model introduces another source of error. Specifically, for a given service rate, the lag predicted by the infinite server model (which has no delays) underestimates the actual lag in the finite server system and this underestimate increases as the peak probably of delay increases. Green and Kolesar (1997) showed that a lagged PSA approach for predicting peak probabilities of delay is quite accurate for small values of μ and low delay probabilities, when simpler methods are very inaccurate. More importantly, using the lagged arrival rate almost always results in the correct identification of the number of servers needed to meet a specified target peak probability of delay in these cases.

We used the "Lag" approach combined with each of the three methods described above for calculating the arrival rate for each planning period, resulting in what we call the *Lag Avg, Lag Max*, and *Lag Mix* methods. In virtually all cases, the Lag approach results in fewer errors than its nonlagged counterpart. In particular, the appropriate choice of Lag Avg or Lag Max produces no or few errors when μ is relatively small, hence increasing the domain for which the SIPP method is reliable. Specifically, our numerical results show:

- Lag Avg is always reliable when relative amplitude is low, i.e., 0.1 or 0.5, and planning periods are short, i.e., 0.25 or 0.5 hour.
- Lag Max is reliable when $RA = 0.1$ and 0.5 regardless of the length of the planning period. When $RA = 1$, Lag Max produces no errors whenever $\mu \geqslant 8$.
- Though Lag Max is not perfect for $\mu = 2$ and 4, it never results in more than three half-hour periods for which the target is exceeded by more than 10%.

Table 7 shows the numerical results using Lag Max for $RA = 1$. It is important to note that while use of the Lag approach doesn't close all the gaps we observed when we examined the three nonlagged SIPP methods, there are only 15 of the 1,296 scenarios for which Lag Max is not reliable. Furthermore, 9 of these occur when $\mu = 2$, i.e., the average service time is 30 minutes and the planning period is 15 or 30 minutes. This combination seems to be an unlikely one in real applications.

What is the impact of using the Lag approach on staffing costs? Use of Lag Avg usually produces better results than SIPP Avg while using the same number of staff-hours. (On average, Lag Avg *saves* 0.2% in staff-hours and never results in more than a 2% increase over SIPP Avg.) Similarly, Lag Max reduces the number of errors relative to SIPP Max. For $\mu = 2$ and 4 where Lag Max is most important, the number of staff-hours is sometimes the same, sometimes greater and in a couple of cases, less than when using SIPP Max. When it is greater, the difference is less than 5%.

Table 7. Lag Max reliability: the number of half-hours in which pD exceeds 110% of target, RA = 1.

		Rho, Mu																	
		16						32						64					
Plan Pd	Target	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64
0.25	0.05	2	0	0	0	0	0	2	1	0	0	0	0	3	0	0	0	0	0
	0.1	0	0	0	0	0	0	3	0	0	0	0	0	3	1	0	0	0	0
	0.2	0	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0
0.5	0.05	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
	0.1	0	0	0	0	0	0	0	0	0	0	0	0	2	1	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	1	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	2	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

4. AN EXAMPLE USING EMPIRICAL DEMAND DATA

In this section, we examine how well SIPP and its variants perform on a model derived from empirical data. Figure 4 shows a telephone call arrival curve derived from actual data provided by a financial services company's incoming telephone call center. The figure shows the following broad pattern. Call volumes are very low at midnight and stay low in the early morning hours, dropping to about 30 calls per hour at about 1 a.m. They then rise smoothly until about 8:30 a.m., when they peak at about 2,100 calls per hour. Call volumes then drop slightly and remain at about 1,900 calls per hour until about 1 p.m., when they peak again at about 2,100 calls per hour. They drop off thereafter, smoothly falling back to the low at midnight. This pattern is at once similar enough to the pattern seen in Figure 1 to reinforce our work with the sinusoidal model, and different enough for us to want to test how SIPP and its variants work in this particular call center scenario.

Given our results of §2, we predicted that SIPP Avg would not be reliable for this application, given a RA close to 1, long planning periods, and large system size. In order to determine how well the sinusoidal model provides insights on the performance of SIPP and its variants for this case, we analyzed this system by running our differential equation solver using the empirical arrival rate function as well as a sinusoidal approximation to this function. For the sinusoidal model we used $\lambda = 948$, which is the 24-hour average of the empirical call rate, and $RA = 1$. The other parameters values used in both cases reflect the way that the call center actually operated. The historical service rate was about 10 calls per hour, 1-hour planning periods were used and the service target was about 0.2.

The staffing levels suggested by SIPP Avg for the empirical demand curve are shown in Figure 4—a total of 2,511 staff-hours are deployed over the 24-hour cycle. Figure 5 is the resulting probability of delay curve. It illustrates that the SIPP staffing is clearly very inadequate. This is confirmed by the results in Table 8, which gives our key performance measures for the empirical model as well as those predicted by the sinusoidal model using SIPP Avg and the five variants. Focusing on the SIPP Avg column, we see that for the empirical demand even the 24-hour average delay probability is 28% and that the delay probability target is exceeded by more than 10% in 26 half-hours. The sinusoidal model for this case predicted that SIPP Avg would result in a 24-hour average delay probability of 24%

Figure 4. A financial services call center: incoming call volume by half hour.

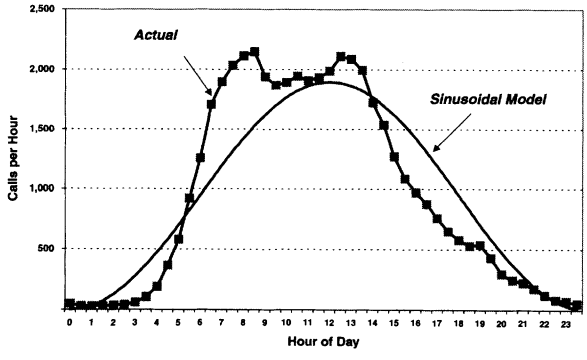


Figure 5. A financial services call center: Probability of delay with SIPP Avg staffing.

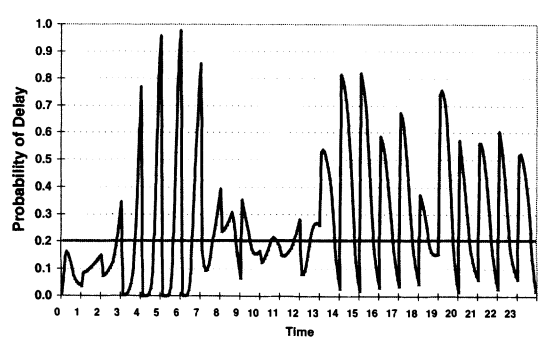


Table 8. Empirical financial services call center model results.

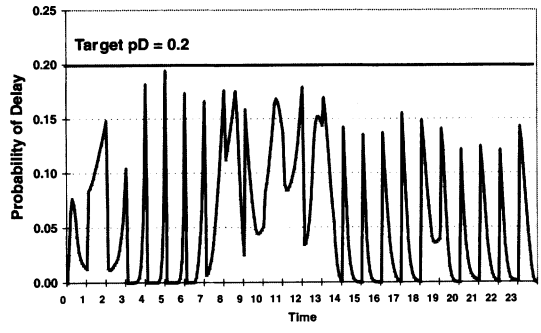
Measure	Method					
	L_Avg	L_Max	L_Mix	S_Avg	S_Max	S_Mix
Empirical Arrival Rate with Mean of 948, Mu = 10, Target = 0.2, PP = 1						
24-hr Avg pD	0.271	0.075	0.153	0.284	0.082	0.148
Max pD (5 min)	0.997	0.194	0.997	0.977	0.263	0.977
Max pD (1/2 hr)	0.637	0.158	0.637	0.629	0.158	0.469
Half-Hours Above Target	24	0	7	31	0	7
Half-Hours Above 110% Target	19	0	6	26	0	6
Staffing Man-Hours	2514	2739	2639	2511	2731	2629
Sinusoidal Arrival Rate with Mean of 948, RA = 1, Mu = 10, Target = 0.2, PP = 1						
24-hr Avg pD	0.237	0.084	0.164	0.239	0.087	0.147
Max pD (5 min)	0.823	0.208	0.823	0.824	0.231	0.709
Max pD (1/2 hr)	0.483	0.186	0.441	0.618	0.186	0.317
Half-Hours Above Target	23	0	10	31	0	11
Half-Hours Above 110% Target	22	0	10	25	0	9
Staffing Man-Hours	2519	2707	2615	2520	2706	2613

with 25 half-hours exceeding the target by more than 10%. These results confirm that the insights developed from the sinusoidal model in §3 are valid for actual demand curves.

Table 8 also shows that the sinusoidal model correctly identifies that both the SIPP Max and Lag Max methods are reliable for the empirical case, with SIPP Max being slightly more efficient. While the sinusoidal model indicates a savings of only 1 staff-hour using SIPP Max and the empirical results show a savings of 8 staff-hours, in both cases the difference in using the two methods could be considered negligible at well under 0.5%. As we state in our concluding section, we suggest that Lag Max be used for these parameter values. Figure 6 shows the $p_D(t)$ curve for the Lag Max solution for the empirical demand. From Table 8, we see that the actual total increase in staff-hours using Lag Max instead of SIPP Avg is about 9%, whereas the sinusoidal models predicts an increase of about 7.5%.

We should remark in closing this section that we have seen a number of empirical demand curves that had similar characteristics to the curve in Figure 4. These curves display: (1) a broad, smooth arrival pattern with a single dominant peak not too dissimilar to a sinusoid with relative amplitude of 1; and (2) a subpattern within the work day that shows two subpeaks with a plateau in between. We take the results of this section as indicating that the findings of §§2 and 3 should apply to such scenarios.

Figure 6. A financial services call center: Probability of delay with Lag Max staffing.



5. SUMMARY, CONCLUSIONS AND FURTHER RESEARCH

The findings in this paper have important implications for the design and management of many types of service systems. First, our results on the reliability of the standard SIPP approach show that managers should be cautious when using this method for determining staffing requirements. In many cases the standard SIPP method will suggest staffing levels that result in actual delays that exceed targeted levels. 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 in an LP-based scheduling model, the model's schedule frequently adds staff (slack) in some periods because of other constraints. However, it is far from certain that such slack would be added where needed most to compensate for SIPP's shortcomings.

Second, our proposed modifications of SIPP provide practitioners with good alternatives which are simple to implement in those cases in which SIPP is unreliable. Table 9 shows the "best method" to use among the six we examined based on reliability, efficiency, and simplicity of the methodology. This table was produced by applying the following rule to each scenario: If all six methods are unreliable, the outcome is 0; else, select the winner from all the methods that have 0 errors (as defined previously) that uses the least staff-hours. In case of ties, select the winner according to the following priority order: SIPP Avg, SIPP

Table 9. Winning SIPP method.

			Rho, Mu																													
			2						4						8						16						32						64					
RA	PP	Targ	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64	2	4	8	16	32	64
0.1	0.25	0.05	4	4	7	7	7	7	4	4	4	4	4	4	4	4	4	7	4	4	4	4	7	7	7	7	4	4	4	4	4	4	7	4	7	4	7	7
		0.1	4	4	4	4	4	4	4	4	7	7	7	4	4	4	4	4	4	4	4	4	7	7	7	7	4	4	7	7	7	7	7	4	4	4	4	4
		0.2	4	4	7	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	7	4	4	4	4	4	4	4	4	4	4	4	7	7	4	4
	0.5	0.05	4	7	4	4	4	4	4	7	7	7	7	4	4	4	4	4	4	4	4	7	4	4	4	4	4	4	4	4	7	7	7	4	4	4	4	4
		0.1	4	4	4	4	4	4	4	7	4	4	4	4	4	7	7	7	7	7	4	7	4	4	4	4	4	4	4	4	4	4	7	7	7	4	4	4
		0.2	4	7	4	4	4	4	4	7	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	7	7	7	7	7	7	4	4	4	4	4
	1	0.05	7	4	4	4	4	4	4	4	4	4	4	4	7	7	7	7	7	4	4	4	4	4	4	4	7	7	4	4	4	4	7	7	7	4	4	4
		0.1	4	4	4	4	4	4	7	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	7	7	4	4	4	4	7	7	7	7	7	7
		0.2	7	4	4	4	4	4	7	4	4	4	4	4	4	7	7	7	4	4	4	4	4	4	4	4	7	4	4	7	4	4	7	7	4	4	4	4
	2	0.05	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	7	7	7	4	4	7	6	4	5	5	5	6	5	5	5	5	5
		0.1	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	7	4	4	4	4	7	4	4	6	6	5	6	5	5	5	5	5
		0.2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	7	7	4	4	4	4	7	4	4	6	6	6	6	6	6	5	5	5
0.5	0.25	0.05	4	4	4	4	4	4	7	4	4	4	7	4	7	4	7	4	4	4	7	4	4	7	7	4	7	7	7	7	4	7	7	7	7	4	4	4
		0.1	4	4	4	4	4	4	4	4	4	4	4	4	7	4	7	7	7	7	7	4	7	7	4	4	7	7	4	7	7	7	7	7	7	7	7	7
		0.2	4	4	7	4	4	7	4	4	4	4	4	4	7	4	4	4	4	4	7	7	4	4	4	4	7	7	7	7	7	4	7	7	7	4	4	4
	0.5	0.05	4	7	7	4	4	4	7	7	7	7	7	7	7	7	4	4	4	4	7	7	7	7	7	4	7	7	7	4	4	4	8	7	7	7	4	4
		0.1	4	7	7	7	4	4	4	7	7	4	4	4	7	4	4	4	4	4	7	7	7	7	7	7	7	7	4	4	7	4	7	7	7	4	7	7
		0.2	4	4	4	4	4	4	7	4	4	4	4	4	7	4	4	7	4	4	7	7	7	7	4	7	7	7	4	4	4	4	7	7	7	7	4	7
	1	0.05	7	4	4	4	4	4	7	4	4	4	4	4	6	6	7	5	5	5	8	6	6	5	5	5	8	6	8	8	8	5	8	6	8	8	8	5
		0.1	7	4	4	4	4	4	7	4	4	4	4	4	7	7	4	4	4	5	6	6	7	6	6	5	8	6	6	5	8	8	8	6	6	8	8	5
		0.2	7	7	7	7	7	7	7	7	7	7	7	7	7	7	4	4	5	5	7	7	7	4	4	4	8	6	7	6	5	5	8	6	6	5	5	8
	2	0.05	6	5	5	5	5	5	6	8	5	5	5	5	5	5	5	5	5	5	6	5	5	5	5	5	8	8	8	8	8	5	8	5	5	8	5	5
		0.1	4	4	4	4	4	4	6	6	6	6	5	5	6	5	5	5	5	5	6	5	5	5	5	5	6	5	8	8	8	8	8	5	8	5	8	5
		0.2	4	4	4	4	4	4	6	6	9	5	5	5	6	5	5	5	5	5	6	5	5	5	5	5	6	8	5	5	5	5	5	5	5	5	5	5
1	0.25	0.05	7	4	7	7	7	7	7	7	7	7	7	7	8	7	4	4	7	4		7	7	4	4	7	0	0	7	7	4	7	0	8	6	6	4	5
		0.1	7	4	7	7	7	4	7	4	4	4	7	4	7	7	7	4	7	7	8	7	7	4	4	4	0	7	7	7	7	7	0	0	6	6	4	4
		0.2	7	4	4	4	4	4	7	4	4	4	4	4	7	7	7	7	4	4	7	7	7	7	7	7	0	7	7	4	4	4	0	7	7	7	4	4
	0.5	0.05	7	7	4	4	4	4	7	7	4	7	4	4	8	7	7	4	4	4	8	7	7	6	5	5	0	0	6	6	4	8	8	8	6	6	5	5
		0.1	7	7	4	4	4	4	7	7	4	7	7	7	7	7	7	4	4	4	8	7	7	7	7	4	8	8	7	6	5	5	0	0	6	6	8	8
		0.2	7	7	7	7	4	4	7	7	7	4	4	4	7	7	7	4	4	4	7	7	7	7	7	7	8	7	7	7	7	4	8	8	6	6	4	8
	1	0.05	7	6	5	5	5	5	6	6	8	5	5	5	8	6	8	8	8	5	8	6	5	5	5	5	0	6	5	8	8	5	8	8	5	5	5	5
		0.1	6	7	4	4	4	4	6	6	6	6	5	5	8	6	5	5	5	5	8	6	8	8	8	8	8	8	5	5	5	5	0	8	8	8	8	8
		0.2	7	4	4	4	4	4	6	6	4	4	4	4	6	6	6	5	5	5	8	6	6	8	5	5	8	6	5	8	5	5	8	6	8	5	8	8
	2	0.05	6	8	5	5	5	5	6	5	8	8	5	5	8	5	5	5	5	5	8	5	5	5	5	5	8	5	5	8	5	5	8	8	5	5	8	8
		0.1	5	5	5	5	5	5	5	5	5	5	5	5	8	5	5	5	5	8	8	5	8	8	5	5	8	8	5	5	5	5	8	8	8	8	8	8
		0.2	6	6	5	5	5	5	6	5	5	5	5	5	6	5	5	5	5	5	8	5	8	8	8	5	8	5	5	5	5	5	8	5	5	5	5	5

Codes of Winning SIPP Variant: 0 = No Method Feasible, 4 = SIPP Avg, 5 = SIPP Max, 6 = SIPP Mix, 7 = Lag Avg, 8 = Lag Max, 9 = Lag Mix

Max, Lag Avg, Lag Max, SIPP Mix, and Lag Mix. As illustrated by this table, there are no rigid, simple rules that describe which method is best across all potential choices of parameters. However, our findings suggest the following guidelines for 24-hour service systems:

(1) SIPP Avg can be reliably used when relative amplitude is low, i.e., $RA \leqslant 0.5$; planning periods are short, 0.25 or 0.5 hours; and the service rate is high, i.e., $\mu \geqslant 32$.

(2) Lag Avg is reliable and efficient when relative amplitude is low and planning periods are short.

(3) For large relative amplitude or long planning periods, SIPP Max or Lag Max will assure reliability. SIPP Max can be reliably used for $\mu \geqslant 8$. In some of these cases, SIPP Mix is equally accurate and will save staff-hours. Lag Max is safer when $\mu < 8$.

As a simple guideline for practitioners, we recommend that Lag Avg be used for low values of RA and short planning periods, and Lag Max be used for all other situations. Of course, SIPP Max or Lag Max typically will use more staff-hours, and a manager may want to consider the trade-off between higher labor costs and what may be 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 trade-offs 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 involving a 24-hour cyclic system and, with appropriate scaling, any continuously operating system with a different cycle length.

Though the findings described in this paper are based on sinusoidal arrival rates, we believe that the directional results are quite general. Our results with the empirical model described in §4 reinforce this judgment. These results indicate that the regions of reliability identified for each method using a sinusoidal arrival rate are fairly robust with respect to other 24-hour cyclic systems.

We should note that our findings are based on probability of delay performance targets. Of course, there are many other possible performance measures including expected delay and the probability that the delay exceeds a specified duration. We worked with probability of delay for several reasons. First, probability of delay metrics are used in the majority of actual implementations with which we are familiar, as well as in the literature on these problems. Because computation of probability delay was technically feasible given our approach of numerical solution of the system's differential equations. Finally, the use of a single parameter performance metric simplifies the analysis and clarifies the findings. Because the other possible performance measures are closely correlated to probability of delay, we believe that the directional findings cited here will hold for these measures as well. We also note that our findings on the reliability of our proposed alternatives to SIPP do not depend on the magnitude of the performance target.

While this research greatly clarifies a number of the issues involved in the staffing of cyclic queueing systems and makes practical proposals for improved practice, it is not the last word. A number of important questions are still on the table for future research. Among these we include the following:

1. Setting optimal staffing levels. The family of SIPP-based methods studied here attempts to produce minimal feasible staffing levels. They do so by segmenting the problem into planning periods. Clearly, this is not optimal in general. What is not at all clear is how to actually determine an optimal cover without resorting to an extensive trial-and-error approach. This is the problem addressed by Thompson (1997) for one set of situations.

2. The SIPP-based models used here and, more importantly, in industry, assume that the system parameters are known. Empirical data available to us show that this is not always a reasonable assumption, particularly for the customer arrival rates. The deviations from forecast in every data set we have examined have exceeded those allowed by the Poisson process model. Thus, it is of interest to build staffing models that explicitly incorporate forecast error.

3. The family of models studied here is assumed to operate continuously over time. While many real-world service systems do so, there are also many that open up, operate over a segment of the day, and then shut down—until they start up again on the next day. The behavior of such limited operating horizon systems is probabilistically different from those studied in this paper. Because managers of some of these systems also employ SIPP methods, it is important to explore SIPP's reliability in this setting. We are well into a research agenda on these models. Our initial findings on SIPP reliability in these limited operating horizon situations confirm the directional results we obtained in this paper's study of continuous systems. However, the ranges of SIPP reliability appear to be smaller, and additional theory is needed to guide the choice of good alternatives when SIPP fails.

REFERENCES

Agnihothri, S. R., P. F. Taylor. 1991. Staffing a centralized appointment scheduling department in Lourdes Hospital. *Interfaces* **21** 1–11.

Andrews, B. H., H. L. Parsons. 1989. L. L. Bean chooses a telephone agent scheduling system. *Interfaces* **19** (6) 1–9.

——, ——. 1993. Establishing telephone-agent staffing levels through economic optimization. *Interfaces* **23** 14–20.

Brewton, J. P. 1989. Teller staffing models. *Financial Manager's Statement*. July–August 22–24.

Brigandi, A. J., D. R. Dargon, M. J. Sheehan, T. Spencer III. 1994. AT&T's call processing simulator (CAPS) operational design for inbound call centers. *Interfaces* **24** 6–28.

Brusco, M. J., L. W. Jacobs, R. J. Bongiorno, D. V. Lyons, B. Tang. 1995. Improving personnel scheduling at airline stations. *Oper. Res.* **43** 741–751.

Cooper, R. B. 1972. *Introduction to Queueing Theory*. Macmillan Co., New York.

Edie, L. C. 1954. Traffic delays at toll booths. *Oper. Res.* **2** 107–138.

Eick, S. G., W. A. Massey, W. Whitt. 1993 $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* **39** 241–252.

Gopalakrishnan, M., S. Gopalakrishnan, D. M. Miller. 1993. A decision support system for scheduling personnel in a newspaper publishing environment. *Interfaces* **23** 104–115.

Green, L., P. Kolesar. 1984. The feasibility of one-officer patrol in New York City. *Management Sci.* **20** 964–981.

——, ——. 1989. Testing the validity of a queuing model of police patrol. *Management Sci.* **35** 127–148.

——,——. 1991 The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* **37** 84–97.

——, ——. 1995. On the accuracy of the simple peak hour approximation for Markovian queues. *Management Sci.* **41** 1353–1370.

——, ——. 1997. The lagged PSA for estimating peak congestion in Markovian queues with periodic arrival rates. *Management Sci.*, **43** 80–87.

——, ——. 1998. A note on approximating peak congestion in $M/G/\infty$ queues with sinusoidal arrivals. *Management Sci.* **44**.

——, ——, A. Svoronos. 1991. Some effects of nonstationarity on multiserver Markovian queuing systems. *Oper. Res.* **39** 502–511.

Gross, D., C. M. Harris. 1985. *Fundamentals of Queuing Theory*, 2nd ed. John Wiley & Sons, New York.

Holloran, T. J., J. E. Byrne. 1986. United Airlines station manpower planning system. *Interfaces* **16** 39–50.

Kwan, S. K., M. M. Davis, A. G. Greenwood. 1988. A simulation model for determining variable worker requirements in a service operation with time-dependent customer demand. *Queueing Systems* **3** 265–276.

Kolesar, P. J., K. L. Rider, T. B. Craybill, W. W. Walker. 1975. A queuing-linear programming approach to scheduling police patrol cars. *Oper. Res.* **23** 1045–1062.

Larson, R. C. 1972. *Urban Police Patrol Analysis*, MIT Press, Cambridge, MA.

Segal, M. 1974. The operator scheduling problem: a network flow approach. *Oper. Res.* **22** 808–823.

Stern, H. I., M. Hersh. 1980. Scheduling aircraft cleaning crews. *Transport. Sci.* **14** 277–291.

Sze, D. Y. 1984. A queueing model for telephone operator staffing. *Oper. Res.* **32** 229–249.

Taylor, P. E., S. J. Huxley. 1989. A break from tradition for the San Francisco Police: patrol officer scheduling using an optimization-based decision support system. *Interfaces* **19** 4–24.

Thompson, G. M. 1993. Accounting for the multi-period impact of service when determining employee requirements for labor scheduling. *J. Oper. Management* **11** 269–287.

——. 1997. Labor staffing and scheduling models for controlling service levels. *Naval Res. Logist.* **44** 719–740.

Whitt, W. 1991 The pointwise stationary approximation for $M_t/M_t/S$ queues is asymptotically correct as the rates increase. *Management Sci.* **37** 307–314.