2007

Managing Trade-Offs in Call Center Agent Scheduling: Methodology and Case Study

Vijay Mehrotra
University of San Francisco, vmehrotra@usfca.edu

Managing Trade-Offs in Call Center Agent Scheduling: Methodology and Case Study

Robert Saltzman
Vijay Mehrotra
Decision Sciences Department
San Francisco State University
1600 Holloway Avenue, San Francisco, CA 94132
saltzman@sfsu.edu
vjm@sfsu.edu

Keywords: Call Center, Scheduling, Optimization

Abstract

This paper develops a flexible and tractable scheduling methodology that produces near-optimal call center agent schedules while taking into account the costs associated with customer waiting time, customer abandonment, and call center agents. Our methodology combines integer programming (to find a desirable staffing plan for a given total number of agents) and simulation modeling (to evaluate the weekly costs of a given staffing plan). We describe the advantages of this approach over the traditional scheduling method, and test both methods by building schedules based on actual demand and shift data from an actual call center operated by Expedia.com under a variety of cost scenarios. The new scheduling approach not only out-performs the traditional staffing approach in all scenarios examined, it reduces total weekly costs of the call center's existing agent schedule by 8-25%, depending on the scenario.

1. INTRODUCTION

This paper introduces a new approach to creating near-optimal weekly agent schedules for call centers developed as a result of our work with an inbound call center in Tacoma, WA supporting Expedia.com, a web-based travel service that primarily serves consumers and small businesses. During a typical day, the agents in this call center answer several thousand calls from customers who, for example, are considering an on-line purchase but need additional information, have questions about tickets already purchased on-line, or have encountered some difficulty with the on-line booking process.

Because the company's "products" (travel bookings) are viewed by many consumers as commodities, call center management is simultaneously under pressure to minimize labor costs (because of low margins) while minimizing customer abandonment (since abandoned calls reduce revenue and profit) and waiting times (which, in addition to driving customer abandonment, also increase customer dissatisfaction and future defections to competitors). In addition, although the company historically has had a culture in which all agents worked full-time shifts, management has also expressed interest in exploring limited use of part-time agents, either as in-house employees or through an outsourced service provider.

The major contributions of this paper are to (1) develop a flexible and tractable scheduling methodology that explicitly accounts for the costs of customer waiting time, customer abandonment, and call center agents; (2) demonstrate how this methodology differs from traditional agent scheduling techniques; and (3) evaluate the performance of this method based on actual operational data from Expedia's Tacoma call center.

While our initial motivation for developing this scheduling methodology was derived from one company's operations, the conditions under which this call center operates are increasingly common. Over the past 20 years, rapid changes in technology and regulation have resulted in many industries (including financial services, travel, and consumer products) becoming far more competitive, leading to increased commoditization, decreased profit margins, and more demanding customers.

The remainder of the paper is organized as follows. In Section 2, we provide a survey of the literature associated with personnel scheduling in general and call centers in particular. Section 3 describes the agent scheduling problem in detail and the traditional approach to solving it, while Section 4 presents our new staffing approach. Section 5 describes the implementation of the scheduling methodology and the operational data associated with the Tacoma call center that is used for our numerical examples. In Section 6, we utilize this call center's data to compare the performance of the traditional scheduling model with the new methodology on a wide range of cost scenarios. Section 7 summarizes our work and indicates avenues for future research in this area.

 ISBN # 1-56555-316-0

2. OVERVIEW AND LITERATURE SURVEY

Research into personnel scheduling problems dates back to [Dantzig 1954], and there have been over 700 papers published in this general area. This body of research is surveyed and cataloged by [Ernst et al. 2004]. In many labor-intensive environments, the essential personnel scheduling challenge is to cost-effectively match the demand for personnel with supply of human resources available, given various restrictions on schedules [Hur et al. 2004].

Because labor comprises the majority of operating costs for inbound call centers [Mehrotra 1997], personnel scheduling is a particularly critical issue for call center managers. A survey of the research literature on call centers is provided in [Gans et al. 2003], while major issues in managing call centers are presented in [Brigandi et al. 1994, Cleveland and Mayben 1997, and Mehrotra 1997]. Industry-specific call centers analyses include an insurance company's inbound call center [Green et al. 2003], a retail phone-order business [Andrews and Parsons 1989], a government agency's consumer information system [Harris et al. 1987], and a software company's technical support call center [Saltzman and Mehrotra 2001].

The problem of call center agent scheduling is typically comprised of three components: (1) forecasting workload; (2) translating workload into agent targets; and (3) scheduling agents based on agent shifts and targets. The first step includes forecasting both the distribution of the call arrival pattern and the distribution of the service times associated with these calls. Within the call center literature, the standard forecasting approach is to treat call arrivals over the course of a day or week as a Nonhomogeneous Poisson Process (NHPP) with constant arrival rates over specific time intervals of 15-, 30- or 60-minutes that are independent of each other. In turn, these per-period arrival rates are typically predicted using historical time series data and exponential smoothing [Andrews and Cunningham 1995, Brown et al. 2005].

The second step is to translate the forecasted arrival rates into a demand for agents per period, which depends not only on the workload forecast but also on the customer waiting time targets. Grassman [1988] discusses many of the practical issues generally associated with this type of translation, while Green et al. [2001, 2003] describe the standard call center forecast translation process, which they refer to as the Stationary, Independent, Period by Period method (SIPP) method. The SIPP method treats every period as an independent stationary M/M/S queuing system, and the target number of servers for each period is set to be the minimum number for which the acceptable waiting time distribution will be achieved in steady state for the given workload forecast. Green et al. [2001] propose improvements to the way in which these agent targets are determined for call centers with cyclic demand, while Green et al. [2003] suggest similar modifications for call centers with limited daily operating hours. Ingolfsson et al. [2005] suggests estimating the target number of agents per period by using the randomization method described in Grassman [1977] to explicitly model the dependence between time periods. Atlason et al. [2002] address the dependence of the time intervals by combining the service level evaluation with the schedule optimization, using cutting planes and simulation.

Finally, the target number of agents per period becomes input for agent scheduling optimization models that seek to minimize overall staffing costs. The traditional scheduling approach is to formulate and solve an integer program that determines the optimal number of agents needed for each shift, and is described in detail later in this paper. The target number of agents for each period is treated as a strict lower bound; a notable exception is Thompson [1993], who relaxes this lower bound assumption while including penalties on employee shortages.

It is important to note that the chosen waiting time target – and the enforcement of that choice as a binding constraint in each period in the scheduling optimization – has a significant impact on traditional scheduling models and costs. Specifically, the requirement that each interval be staffed with enough agents to achieve the target waiting time is a more stringent requirement than having the overall (daily or weekly) waiting time distribution achieve this same target. The new scheduling methodology described here seeks to address this issue, as do several recent papers [Atlason et al. 2002, Ingolfsson et al. 2005].

In addition to the distribution of customer waiting times, more than 40% of call centers [Garnett et al. 2002] also look at the *abandonment rate*, or percentage of callers who hang up before talking to an agent, as a key performance measure. Companies view abandonment as highly undesirable since they lose current revenue from abandoning customers and may lose future revenue as well. Thus, the scheduling methodology described in this paper also explicitly includes abandonment costs in its objective function.

Since the traditional methodology first determines a target number of agents per period (based on waiting time targets) and then using those targets as constraints in a cost minimization, the traditional scheduling optimization model does not have the opportunity to make economic trade-offs between cost and service quality as advocated in [Andrews and Parsons 1993] and [Hillier and Lieberman 1986]. While the choice of the waiting time target can be quite arbitrary, it can have a major affect on

the target number of agents and overall staffing costs. By contrast, the new scheduling methodology described here explicitly includes the relative costs of labor, waiting, and abandonment in its objective function in order to address the associated tradeoffs between costs and service quality.

3. PROBLEM DEFINITION, NOTATION, AND THE TRADITIONAL SCHEDULING APPROACH

Given a planning horizon of m intervals (typically 15 or 30 minutes each), the call center manager's core scheduling problem is to determine how many agents to assign to each of n available shifts. For each period $i = 1, 2, \ldots, m$, and each shift j = 1, 2, …, n, we define the binary parameter A_{ij} to be 1 if an agent working shift i is available to answer calls during period j; A_{ij} is 0 if the agent is not available to answer calls during this period (which may mean the agent has not yet arrived for work; is committed to some non-phone activity such as a break, lunch, or training; or has completed work for the day). Without loss of generality, we assume that there are a total of f full-time shifts (where $1 \le f \le n$); shifts indexed by $f+1, f+2, \ldots, n$, are part-time shifts.

The n x m activity matrix $A = \{A_{ij}\}$ contains a key set of input parameters for the agent scheduling problem. For example, the A matrix for the call center that is analyzed later in the paper is shown in Figure 1. The 49 shifts represented by the columns of A include 26 shifts that had previously been used by full-time agents and 23 proposed new 4-hour shifts for part-time agents (as discussed in Section 6).

During any period i, we define c_i as number of available agents to handle inbound calls during that period. We also define x_j as the number of agents scheduled to work shift on shift j. We refer to the m-vector $c = \{c_i\}$ as the call center's *capacity* and the n-vector $x = \{x_j\}$ as the *agent schedule.* For a given agent schedule x, the activity matrix A determines the overall capacity through the equation $Ax = c$. Also, for each shift j, we define h_j as the total number of paid periods for shift j. Once a cost-effective schedule has been found, the manager assigns individual agents to the scheduled shifts.

The algorithm most often used to determine call center staffing levels, referred to here as the "traditional approach," takes a service level view of the problem. Given each period's expected volume of calls and mean handling time, the SIPP method assumes that system behavior in every period i is well-approximated by an M/M/S queuing system and is independent of the preceding period's behavior. It calculates the staffing targets s_i separately for each period i with queuing formulas found in [Hillier and Lieberman 1986], for example. Using these estimated staffing requirements, the following manpower scheduling problem [Gans et al. 2003] is then solved to determine how many agents x_j should be assigned to each shift j.

$$\text{Minimize} \sum_{j=1}^{n} h_j w_j x_j \tag{1}$$

$$\text{subject to: } \sum_{j=1}^{n} A_{ij} x_j \ge s_i, \text{ for all } i = 1, \ldots, m \tag{2}$$

$$\sum_{j=f+1}^{n} x_j \le PT_{\max} \tag{3}$$

$$x_j \ge 0 \text{ and integer, for all } j = 1, \ldots, n \tag{4}$$

where w_j in the objective function (1) is the wage rate per period of an agent who works shift j. Constraint set (2) forces agent capacity to meet or exceed the target staffing level in every period. Anticipating the possibility of analyzing the impact of part-time agents, constraint (3) keeps the total number of part-time agents from exceeding a desired maximum PT_{max}. Finally, non-negativity and integrality of the staffing decisions are specified in constraint set (4).

Because the service level target must be met in every period, the traditional approach tends to generate larger x_j values than are really needed. Consequently (as shown in Section 6), when the costs of waiting time, customer abandonment, and staffing are all taken into account, the traditional approach consistently produces schedules with relatively high total costs. This overstaffing is also more acute when there are either a small number of part-time shifts or none at all.

4. NEW SCHEDULING APPROACH: WEEKLY STAFFING PLAN ROUTINE

The new scheduling approach presented here, called the Weekly Staffing Plan Routine (WSPR), seeks to minimize the call center's total staffing and waiting costs per week. Service costs are based on the number of periods associated with each shift, the per period wage rate for each shift, and the number of agents scheduled to work each shift. Expected total waiting costs are the sum of expected abandonment costs, (which are the product of the expected number of abandoned calls per week N_{Ab} and the associated cost of an abandoned call C_{Ab}) and expected caller waiting costs (which depend on the expected number of calls on hold L_q and the associated cost per period of keeping a caller on hold C_{Lq}). Thus, for a given schedule x, the expected total costs (ETC) per week are:

$$\text{ETC} = \sum_{j=1}^{n} h_j w_j x_j + N_{Ab} C_{Ab} + m L_q C_{L_q} \tag{5}$$

 ISBN # 1-56555-316-0

			Shift *j*																					
			X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}		X_{47}	X_{48}	X_{49}	Σx
Agents scheduled			0	11	11	0	0	0	0	0	0	0	4	8	0	0	15	0	3	0		0	0	1	64
Time Period	Day	*i*	A^1	A^2	A^3	A^4	A^5	A^6	A^7	A^8	A^9	A^{10}	A^{11}	A^{12}	A^{13}	A^{14}	A^{15}	A^{16}	A^{17}	A^{18}		A^{47}	A^{48}	A^{49}	Cap C_i
6:00	Mon	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	11
6:15	Mon	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	11
6:30	Mon	3	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	22
6:45	Mon	4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	22
7:00	Mon	5	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	32
7:15	Mon	6	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	32
7:30	Mon	7	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0		0	0	0	41
7:45	Mon	8	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0		0	0	0	41
8:00	Mon	9	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0		0	0	0	41
8:15	Mon	10	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0		0	0	0	41
8:30	Mon	11	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0		0	0	0	42
8:45	Mon	12	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0		0	0	0	42
9:00	Mon	13	1	1	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0		0	0	0	43
9:15	Mon	14	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0		0	0	0	53
9:30	Mon	15	1	1	1	1	0	1	0	1	1	1	1	1	0	0	0	0	0	0		0	0	0	44
9:45	Mon	16	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0		0	0	0	53
10:00	Mon	17	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0		0	0	0	53
10:15	Mon	18	0	1	1	0	1	1	1	1	0	1	1	1	1	1	0	0	0	0		0	0	0	53
10:30	Mon	19	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0		0	0	0	57
10:45	Mon	20	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0		0	0	0	42
11:00	Mon	21	1	0	1	1	1	0	1	0	1	1	1	0	1	1	1	1	0	0		0	0	0	39
11:15	Mon	22	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	0	0		0	0	0	58
11:30	Mon	23	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1		0	0	0	52
11:45	Mon	24	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1		0	0	0	52
12:00	Mon	25	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1		0	0	0	55
12:15	Mon	26	1	1	1	0	1	1	1	1	0	0	1	1	0	1	1	1	1	1		0	0	0	55
12:30	Mon	27	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1		0	0	0	71
12:45	Mon	28	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1		0	0	0	63
13:00	Mon	29	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1		0	0	0	70
13:15	Mon	30	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1		0	0	0	63
13:30	Mon	31	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0		0	0	0	82
13:45	Mon	32	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1		0	0	0	79
14:00	Mon	33	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	1	1		0	0	0	79
14:15	Mon	34	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1		0	0	0	67
14:30	Mon	35	0	0	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	1		0	0	0	59
14:45	Mon	36	0	0	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1		0	0	0	56
20:15	Fri	298	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	1	1	31
20:30	Fri	299	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	1	28
20:45	Fri	300	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	1	28

Figure 1. Agent Schedule and Capacities by 15-Minute Period for the Tacoma Call Center.

Call center managers influence staffing costs by controlling the total number of agents employed, and waiting costs by how they allocate these agents to the various shifts. Some schedules decrease waiting costs by reducing caller abandonment and holding time, but increase staffing costs; others decrease staffing costs at the expense of higher waiting costs. However, judiciously chosen schedules can simultaneously reduce both staffing and waiting costs if the corresponding capacities are well-matched to agent requirements. The WSPR approach

allows explicit trade-offs to be made among service and waiting costs by employing a combination of optimization and simulation modeling.

Figure 2 gives an overview of WSPR, which addresses the weekly agent scheduling problem with a combination of optimization and simulation. For a sequence of total agents hours T worked per week specified by the manager, from T_{min} to T_{max}, WSPR first solves an optimization problem that identifies a schedule $x^*(T)$ that best matches agent capacities to agent requirements, and then evaluates its total cost by running a discrete event dynamic simulation model of the call center. The simulation model's central role is to evaluate call center performance resulting from a given staffing plan by accurately estimating the average number of callers waiting on hold and abandoned calls per week. These averages are based on multiple independent replications of the model, each of which simulates the entire planning horizon consisting of periods 1, 2, …, m. With mean performance measures and user-supplied costs for agents, abandoned calls, and calls waiting on hold, the average total cost per week of a given staffing plan can then be calculated via equation (5).

---

0. Read Key Input Data:
Shift matrix A (m = 300 time periods by n = 49 shifts)
Agent requirements s_i, pre-computed for i = 1, 2, …, m
Mean call volume v_i, for i = 1, 2, …, m (#)
Mean call length l_i, for i = 1, 2, …, m (#)
Mean extension out time e_i, for i = 1, 2, …, m (#)
(#) Explained more fully in Section 5

1. For total agent hours T from T_{min} to T_{max}
Find best requirements-matching staffing plan $x^*(T)$
Find mean total cost/week Z of $x^*(T)$ via simulation
Next T

2. Select $x^*(T)$ with lowest Z as the *near-optimal staffing plan* (x^b, Z^b)

Figure 2. Weekly Staffing Plan Routine (WSPR) Overview

For a particular total number of agent hours worked per week T, WSPR first finds the schedule $x^*(T)$ that best matches agent capacities with agent targets in all periods. More specifically, $x^*(T)$ is found by solving the integer program (6)-(10), where PT_{max} is the maximum number of part-time agent hours available, and h_j represents the total number of paid *hours* per week included in shift j, e.g., in the example presented in the next section, h_j = 40 hours for j = 1, 2, …, f and h_j = 20 hours for $j = f+1, f+2, \ldots, n$.

$$\text{Minimize } \sum_{i=1}^{m} | c_i - s_i | = Z \qquad (6)$$

$$\text{subject to: } \sum_{j=1}^{n} A_{ij} x_j = c_i, \text{ for all } i = 1, \ldots, m \qquad (7)$$

$$\sum_{j=1}^{n} h_j x_j \le T \qquad (8)$$

$$\sum_{j=f+1}^{n} x_j \le PT_{\max} \qquad (9)$$

$$x_j \ge 0 \text{ and integer, for all } j = 1, \ldots, n \qquad (10)$$

The objective function (6) minimizes the sum of the absolute differences between agent targets and scheduled agent capacities in all periods. Constraint set (7) converts the schedule x into capacities by period, while constraint (8) requires the total number of agent hours allocated to the shifts to be no more than the specified amount T. Constraints (9) and (10) are the same as constraints (3) and (4) of the traditional scheduling approach.

WSPR takes the resulting staffing plan $x^*(T)$ and finds its average total weekly cost by running the corresponding agent capacities through a simulation model. The lowest total cost plan among those found during the T loop is designated as (x^b, Z^b), where Z^b is the value of the objective function (5) for the best schedule x^b.

While x^b is a high-quality schedule, it is not the global optimum. Computational experience reported in Saltzman [2005] with heuristic search procedures that start at (x^b, Z^b), e.g., tabu search, demonstrated the existence of lower cost schedules. These heuristics, however, are fairly time consuming and tend to produce only very modest reductions in total costs; as such, they were not incorporated into WSPR.

Finally, we note that WSPR is a flexible scheduling methodology. The critical characteristic of its objective function (5) is that it includes costs for staffing, waiting, and abandonment – but because it is evaluated based on simulation output, it is not restricted to any particular functional form. Furthermore, WSPR's embedded integer program (6)-(10) allows for any arbitrary methodology to determine the agent targets s_i and any activity matrix A to be incorporated. The right hand side of constraint (8) can denote either a restriction on total agent hours or on total budget (with the coefficients h_j being replaced by the product $h_j w_j$). The right hand side of constraint (9) can be set arbitrarily small or large depending on the availability of part-time personnel.

5. IMPLEMENTATION AND OPERATIONAL DATA FOR NUMERICAL EXAMPLE

To evaluate the effectiveness of the WSPR methodology, we solved the agent scheduling problem using both the traditional approach and WSPR with data from Expedia's inbound call center in Tacoma, WA. For both

methods, the agent requirements s_i were determined using the traditional SIPP method. Both the traditional method's optimization problem (1)-(4) and WSPR's optimization model (6)-(10) were solved using OPL Studio [ILOG 2000], with the WSPR model first being converted into an integer linear program, as in [Saltzman 2005]

WSPR's core simulation model, which evaluates the cost of a given staffing plan, was built in Arena, a well-known simulation package [Kelton et al. 2004]. Its structure is similar to that of [Saltzman 2005] with two main differences: first, the current model simulates five different workdays per week, and second, it reads in all time-varying data from Excel files at the beginning of each run. The model's crucial role is to accurately estimate the time-averaged number of callers waiting on hold (L_q) and the number of abandoned calls per week (N_{Ab}), and send this information back to the Visual Basic code run from Excel that calculates the current staff plan's total weekly cost.

5.1 WSPR's Core Simulation Model Data

The Arena model reads in four main data sets which vary by period i = 1, 2, …, m: (1) agent capacities c_i, (2) mean call volumes v_i, which are used as the call arrival forecasts, (3) mean call length l_i, and (4) mean call extension out times e_i. The v_i, l_i and e_i data come from the call center's automatic call distributor (ACD) or phone switch during May 2002. Unfortunately, only summary reports giving means by time-of-day were made available to us, i.e., we had no data for how long individual callers spent with agents. This is a typical limitation for call center personnel scheduling systems due to the level of data detail that is stored in ACD databases.

The mean volume of calls v_i in each half-hour period varied considerably throughout the day, ranging from a low of just 1 call in the first half-hour of each day, to a high of 116 calls on Wednesday from 12:00-12:30 PM. It also varied by day of the week, with Tuesday and Wednesday having the largest share of calls (22.3% each) and Friday having the smallest percentage (17.6%).

As is common practice when simulating the arrival of independent callers, the arrival pattern was assumed to be a Nonhomogeneous Poisson Process with constant arrival rates over each 15-minute period. ACD data showed that the mean call length l_i and mean extension out time e_i also varied by time-of-day and day-of-week. During one third of the calls, the agent spent extra time on an extension out call, that is, a call made to an internal four-digit extension such as that of a supervisor or help desk. Both call length and extension out time were assumed to be exponentially distributed, although the simulation model can easily accommodate many other distributional forms if data about individual call durations is available for fitting.

5.2 Modeling Abandonment and Determining the Mean Abandonment Time Parameter

Abandonment behavior is a function of a caller's patience level (abandonment time) and the amount of waiting experienced. Thus, *caller abandonment time* must be represented in a simulation model by an input distribution. However, determining an appropriate distribution to use is difficult because observed abandonment time is known only for a small fraction of customers who actually abandon the system. Most observations come from those who abandon the phone queue relatively quickly. Thus, the modeler has biased information about the mean abandonment time and can only guess at the distribution of abandonment time. Consequently, caller abandonment time was modeled simply as an exponential distribution of the time spent on hold, as in [Garnett et al. 2002], and we estimated the mean time to abandonment from the (censored) historical data about customer abandonment.

Using the actual agent schedules followed by the call center during the period for which the call history was provided, the distribution's mean was systematically altered in experiments until model output averaged across 10 weekly replications closely resembled that of the real system for several key performance measures (see Table 1). In particular, a mean of 17.5 minutes led to model behavior that matched up well with observed call center behavior.

Table 1. Validating the Simulation Model

Perform. Measure	Observed Mean	Model Mean	95% CI HW for Model Mean	Diff. in Means
Offered calls/week	11,296.5	11,284.1	60.7	-0.1%
Abandoned calls/week	509.0	505.9	39.8	-0.6%
% calls answered	95.5	95.7	0.35	0.2%
ASA (seconds)	45.5	45.0	3.6	-1.2%
Ave. no. on hold	N/A	1.97	0.15	N/A

5.3 Simulation Model Validation

Given these input parameter values, we validated the accuracy of our simulation model by using the actual agent schedules followed by the call center and comparing the simulation results to the actual operational results for several key performance measures: offered calls per week, abandoned calls per week, percentage of calls answered, average waiting time (referred to as Average Speed of Answer, or "ASA"), and mean queue length L_q. For each of the first four performance measures, the model's mean output is quite close in both absolute and relative terms to

the observed mean from 4 weeks in May 2002. The fourth column gives the half-width of the 95% confidence interval (CI) for the model mean. Since the CI contains the observed mean in each case, the average behavior of the simulation model does not appear to be significantly different from that of the actual call center.

6. EXPERIMENTAL TESTING

Both WSPR and the traditional approach were applied to the Tacoma call center data under 10 cost scenarios chosen to reflect a diverse set of attitudes about the relative importance of waiting and service costs (see Table 2). In particular, when managers set staffing levels using a small probability of delay, e.g., P_d = 5%, they are implicitly valuing holding and abandonment costs to be quite high relative to service costs. In this case, w_j = \$15/hour per agent for all shifts, C_{Ab} = \$25 in lost revenue per abandoned call and C_{Lq} = \$20 per waiting call per hour (Scenario 1). On the other hand, when managers perceive the costs of providing service to be much higher than those of holding and abandonment, e.g., w_j = \$30, C_{Ab} = \$10, and C_{Lq},= \$10, they are likely to set staffing levels based on a high probability of delay, such as P_d = 50% (Scenario 7). This setting might be more typical of a technical support call center in the software industry whose agents are more technically skilled than those in the travel industry, but whose customers have little choice as to where to get their software-specific questions answered. The complete set of cost scenarios and corresponding P_d values used to set staffing levels by period are shown in Table 2.

Table 2. Cost/P_d Scenarios Tested

Costs (\$)	Probability of Delay P_d					MRSL
w_j:C_{Ab}:C_{Lq}	.05	.10	.25	.50	.75	
15:25:20	1	2				
20:15:15		3	4			
25:15:10			5	6		
30:10:10				7	8	
30:5:5					9	10

The last column, MRSL, refers to the "minimum reasonable staffing level" [Thompson, 1995] in which the staffing requirement s_i in period i is set to $\lceil \lambda_i/\mu_i \rceil$. In this setting (Scenario 10), staffing costs are viewed to be much greater than waiting costs, resulting in relatively poor service.

Numerical results from all scenarios are shown in Table 3. Base case total costs in the right-most column are derived from the performance measures of the validated base case model (shown in the first row) and the costs parameters of a particular scenario. For example, Scenario 1 has average labor costs per week of \$39,600 (66 full-time agents paid \$15/hour for 40 hours/week), average waiting costs of (75 hours/wk.)(\$20/call/hour)(1.97 waiting calls) = \$2,955/week, and average abandonment costs of (506 abandoned calls/week)(\$25/call) = \$12,650/week, for an expected total weekly cost of \$55,205.

The bottom half of Table 3 indicates that the traditional approach's staffing plan incurs almost no waiting costs in any scenario because it considerably over-staffs the call center, employing many more agents than in the base case. The traditional approach always staffs at or above the required number of agents in every period, whereas WSPR's staffing stays close to the requirements, going slightly above or below as needed to minimize total costs. We can see that the traditional approach *increases* total expected weekly costs over the base case by 13.1-32.9%. WSPR, on the other hand, makes an explicit tradeoff between waiting and service costs and finds better staffing plans in every scenario tested. In Scenario 1, for example, WSPR slightly increases base case labor costs but offsets this increase with reductions in waiting costs. Overall, WSPR *reduces* base case total expected weekly costs by 7.5-24.6%, depending on the scenario, and allows call center managers to tailor the staffing plan to fit their perceived costs of waiting and service.

Experiments were also conducted to gauge the impact of part-time labor, e.g., agents who work 4-hour shifts, on total costs. However, due to space limitations, the results of these experiments are not reported here.

7. CONCLUSION AND FUTURE WORK

This article has addressed the issue of agent scheduling, a critical operational issue within call centers. After presenting the traditional methodology for agent scheduling, which assumes that the target probability of delay must be achieved in *every* period of the given planning horizon, we introduced a new scheduling methodology that includes a more robust objective function while relaxing this (often arbitrarily imposed) restriction on per-period delay probabilities. In addition, we compared staffing plans generated by the traditional service level approach to those based on this new approach, which explicitly considers the costs of labor, customer abandonment and waiting time in order to minimize the total overall costs. The new staffing approach substantially out-performed the traditional staffing approach in all scenarios; more importantly, it reduced total weekly costs of the Tacoma call center's existing (base case) staffing plan by approximately 8-25%.

There are several other areas of research to pursue in relation to this new scheduling approach, both in terms of modeling call center operations and of analyzing the performance of the scheduling methodology under different operating conditions. One compelling extension is to call centers with skill-based routing. While the methodology presented here is for call centers where the arrivals form a *single* queue (and each agent can handle each call in that queue), there has been a proliferation of multi-queue call centers in which different agents can handle calls from one

Table 3. Comparison of Results from WSPR and Traditional Approach for Ten Scenarios

Scen.	Staffing Approach	Full-Time Agents	ASA (sec.)	N_{Ab} (calls)	L_q (calls)	Costs per Week ($)			Expected Total Costs per Week ($)	Change from Base Case	Base Case Total Costs ($)
						Labor	Aband.	Waiting			
Base		66	45.0	506.0	1.97						
1	WSPR	70	18.5	198.3	0.80	42,000	4,958	1,193	48,151	-12.8%	55,205
2	WSPR	70	17.6	190.8	0.76	42,000	4,770	1,139	47,909	-13.2%	55,205
3	WSPR	64	31.7	346.7	1.36	51,200	5,201	1,528	57,928	-7.5%	62,606
4	WSPR	59	48.2	531.1	2.06	47,200	7,967	2,323	57,489	-8.2%	62,606
5	WSPR	59	48.2	531.1	2.06	59,000	7,967	1,548	68,515	-8.7%	75,068
6	WSPR	58	52.8	580.0	2.26	58,000	8,700	1,698	68,398	-8.9%	75,068
7	WSPR	49	112.8	1229.0	4.81	58,800	12,290	3,609	74,699	-12.9%	85,738
8	WSPR	49	111.7	1216.9	4.77	58,800	12,169	3,574	74,543	-13.1%	85,738
9	WSPR	35	285.3	3120.6	12.19	42,000	15,603	4,573	62,176	-24.6%	82,469
10	WSPR	34	303.4	3313.8	12.96	40,800	16,569	4,859	62,228	-24.5%	82,469
1	Traditional	110	0.04	0.5	0.00	66,000	13	3	66,015	19.6%	55,205
2	Traditional	104	0.14	1.4	0.01	62,400	35	9	62,444	13.1%	55,205
3	Traditional	104	0.14	1.4	0.01	83,200	21	7	83,228	32.9%	62,606
4	Traditional	96	0.59	6.7	0.03	76,800	101	29	76,929	22.9%	62,606
5	Traditional	96	0.59	6.7	0.03	96,000	101	19	96,120	28.0%	75,068
6	Traditional	89	1.90	20.8	0.08	89,000	312	63	89,375	19.1%	75,068
7	Traditional	89	1.90	20.8	0.08	106,800	208	63	107,071	24.9%	85,738
8	Traditional	85	3.22	35.0	0.14	102,000	350	107	102,457	19.5%	85,738
9	Traditional	85	3.22	35.0	0.14	102,000	175	53	102,228	24.0%	82,469
10	Traditional	80	6.04	70.2	0.27	96,000	351	102	96,453	17.0%	82,469

or more queues. This has led to significant recent research [Whitt 2006; Harrison and Zeevi 2005]. An excellent survey of this literature is provided by Koole and Pot [2006]. In this context, one major and very interesting research problem would be to extend WSPR and the integer program (6)-(10) to include multiple agent types, using the approximations for the target number of agents with different skill sets from Wallace and Whitt [2005] for the s_i in (6).

The WSPR method also invites additional empirical research. In our study, all agents were assumed to cost the call center the same amount. In light of the fact that many call center activities are now being outsourced to low cost agents overseas, an interesting question would be to examine the impact of using different costs for different types of agents (in-house vs. outsourced, full-time vs. part time). A second area to explore would be the effect of a nonlinear waiting cost function compared to the current linear approach. Another avenue to investigate is the impact on operating costs of designing additional agent shifts to include in the *A* matrix. Finally, it would be informative to obtain data sets and cost estimates from other call centers to see if the results reported here prove to be consistent across a variety of call center settings.

References

Andrews, B. H. and S. M. Cunningham. 1995. "L.L. Bean Improves Call-Center Forecasting, *Interfaces*, 25 (6), 1-13.

Andrews, B. H. and H. L. Parsons. 1989. "L. L. Bean Chooses a Telephone Agent Scheduling System," *Interfaces*, 19 (6), 1-9.

Andrews, B. H. and H. L. Parsons. 1993. "Establishing Telephone-Agent Staffing Levels Through Economic Optimization," *Interfaces*, 23 (2), 15-20.

Atlason, J., M. A. Epelman and S. G. Henderson. 2002. "Call Center Staffing with Simulation and Cutting Plane Methods," *Annals of Operations Research*, 127, 333-358.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective," *Journal of the American Statistical Association*, 100, 36-50.

Brigandi, A., D. Dargon, M. Sheehan and T. Spencer. 1994. "AT&T's Call Processing Simulator Operational Design for Inbound Call Centers," *Interfaces*, 24 (1), 6-28.

Cleveland, B. and J. Mayben. 1997. *Call Center Management on Fast Forward*, Call Center Press, Annapolis, MD.

Dantzig, G. B. 1954. "A Comment on Edie's 'Traffic Delays at Toll Booths,'" *Operations Research*, 2 (3), 339-341.

Ernst, A. T., H. Jiang, M. Krishnamoorthy, B. Owens and D. Sier. 2004. "An Annotated Bibliography of Personnel Scheduling and Rostering," *Annals of Operations Research*, 127 (1-4), 21-144.

Gans, N., G. Koole, and A. Mandelbaum. 2003. "Telephone Call Centers: Tutorial, Review and Research Prospects," *Manufacturing and Service Operations Management* 5 (2), 79-141.

Garnett, O., A. Mandelbaum and M. Reiman. 2002. "Designing a Call Center with Impatient Customers," *Manufacturing and Service Operations Management*, 4 (3), 208-227.

Grassman, W. K. 1977. "Transient Solutions to Markovian Queueing Systems," *Computers and Operations Research*, 4, 47-53.

Grassman, W. K. 1988. "Finding the Right Number of Servers in Real-World Queueing Systems," *Interfaces*, 18 (2), 94-104.

Green, L. V., P. J. Kolesar and J. Soares. 2001. "Improving the SIPP Approach for Staffing Service Systems that have Cyclic Demands," *Operations Research*, 49 (4), 549-565.

Green, L. V., P. J. Kolesar and J. Soares. 2003. "An Improved Heuristic for Staffing Telephone Call Centers with Limited Operating Hours," *Production and Operations Management*, 12 (1), 46-61.

Harris, C. M., K. L. Hoffman and P. B. Saunders. 1987. "Modeling the IRS Taxpayer Information System," *Operations Research*, 35 (4), 504-523.

Harrison, J. M. and A. Zeevi. 2005. "A Method for Staffing Large Call Centers Based on Stochastic Fluid Models," *Manufacturing and Service Operations Management*, 7 (1), 20-36.

Hillier, F. S. and G. J. Lieberman. 1986. *Introduction to Operations Research, 4th ed.*, Holden-Day, Oakland, CA.

Hur, D., V. A. Mabert, and K. M. Bretthauer. 2004. "Real-Time Work Schedule Adjustment Decisions: An Investigation and Evaluation," *Production and Operations Management*, 13 (4), 322-339.

ILOG. 2000. *ILOG OPL Studio 3.0 User's Manual*. ILOG, Gentilly, France.

Ingolffson, A., E. Akhmetshina, S. Budge, Y. Li and X. Wu. 2005. "A Survey and Experimental Comparison of Service Level Approximation Methods for Non-Stationary M/M/s Queueing Systems," Working Paper, University of Alberta.

Kelton, W. D., R. P. Sadowski and D. Sturrock. 2004. *Simulation with Arena, 3rd ed.*, McGraw-Hill, Boston.

Mandelbaum, A. 2000. *4CallCenters: Personal Optimization Tools for Call Centers (v2.01)*, http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm

Koole, G. and A. Pot. 2006. "An Overview of Routing and Staffing in Multi-Skill Contact Centers," Working Paper, Vrije University.

Mehrotra, V. 1997. "Ringing Up Big Business," *OR/MS Today*, 24 (4), 18-24.

Saltzman, R. M. 2005. "A Hybrid Approach to Minimize the Cost of Staffing a Call Center," *International Journal of Operations and Quantitative Management*, 11 (1), 1-14.

Saltzman, R. M. and V. Mehrotra. 2001. "A Call Center Uses Simulation to Drive Strategic Change," *Interfaces*, 31 (3), 87-101.

Thompson, G. M. 1993. "Representing Employee Requirements in Labour Tour Scheduling," *OMEGA International Journal of Management Science*, 21 (6), 657-671.

Thompson, G. M. 1995. "Labor scheduling using NPV estimates of the marginal benefit of additional labor capacity." *Journal of Operations Management*, 13 (1), 67-86.

Wallace, R. B. and W. Whitt. 2005. "A Staffing Algorithm for Call Centers With Skill Based Routing," *Manufacturing and Service Operations Management*, 7 (4), 276-294.

Whitt, W. 2006. "A Multi-Class Fluid Flow Model for a Contact Center With Skill-Based Routing," *International Journal of Electronics and Communication*, 60 (2), 95-102.

Biographies

Robert Saltzman teaches courses in spreadsheet-based decision modeling, operations management, simulation, and business statistics in the College of Business at SF State. He earned his PhD in OR from Stanford University, and has research interests in applied optimization and animated simulation modeling.

Vijay Mehrotra is an Assistant Professor in the College of Business at SF State. Prior to joining the faculty in 2003, he held the positions of Vice President at Blue Pumpkin Software and CEO at Onward, Inc. Vijay earned his M.S. and Ph.D. in OR from Stanford University. He writes a regular column in *OR/MS Today* entitled "Was It Something I Said?" Vijay's research interests include applications of stochastic processes and optimization, queuing networks, and the adoption of models and information technology.