PRODUCTION AND OPERATIONS MANAGEMENT

POMS

Vol. 16, No. 6, November-December 2007, pp. 665–688
ISSN 1059-1478 | 07 | 1606 | 665$1.25

DOI 10.3401/poms.

The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research

Zeynep Aksin • Mor Armony • Vijay Mehrotra
College of Administrative Sciences and Economics, Koc University, Rumeli Feneri Yolu, 34450 Sariyer-Istanbul, Turkey
Leonard N. Stern School of Business, New York University, West 4th Street, KMC 8–62, New York, New York 10012, USA
Department of Decision Sciences, College of Business, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132-1722, USA
zaksin@ku.edu.tr • marmony@stern.nyu.edu • vjm@sfsu.edu

Call centers are an increasingly important part of today's business world, employing millions of agents across the globe and serving as a primary customer-facing channel for firms in many different industries. Call centers have been a fertile area for operations management researchers in several domains, including forecasting, capacity planning, queueing, and personnel scheduling. In addition, as telecommunications and information technology have advanced over the past several years, the operational challenges faced by call center managers have become more complicated. Issues associated with human resources management, sales, and marketing have also become increasingly relevant to call center operations and associated academic research.

In this paper, we provide a survey of the recent literature on call center operations management. Along with traditional research areas, we pay special attention to new management challenges that have been caused by emerging technologies, to behavioral issues associated with both call center agents and customers, and to the interface between call center operations and sales and marketing. We identify a handful of broad themes for future investigation while also pointing out several very specific research opportunities.

Key words: call centers; staffing; skill-based routing; personnel scheduling; outsourcing
Submissions and Acceptance: Submissions and Acceptance: Received April 2007; revision received October 2007; accepted October 2007.

1. Introduction

Virtually all businesses are interested in providing information and assistance to existing and prospective customers. In recent years, the decreased costs of telecommunications and information technology have made it increasingly economical to consolidate such information delivery functions, which led to the emergence of groups that specialize in handling customer phone calls. For the vast majority of these groups, their primary function is to receive telephone calls that have been initiated by customers. Such operations, known as "inbound" call centers, are the primary topic of this paper.

Inbound call centers are very labor-intensive operations, with the cost of staff members who handle phone calls (also known as "agents") typically comprising 60–80% of the overall operating budget. Inbound call centers may be physically housed across several different locations, time zones, or countries.

Inbound call centers make up a large and growing part of the global economy. Although reliable industry statistics are notoriously hard to come by, the Incoming Call Management Institute (ICMI), a highly reputable industry association, regularly tracks published industry statistics from several sources (www.incoming.com/statistics/demographics.aspx). By 2008, various studies cited by ICMI predict the following:

- The United States will have over 47,000 call centers and 2.7 million agents.

- Europe, the Middle East, and Africa together will have 45,000 call centers and 2.1 million agents.
- Canada and Latin America will have an estimated 305,500 and 730,000 agents, respectively.

Meanwhile, the demand for call center agents in India has grown so fast that the labor supply has been unable to keep up with it: by 2009, the demand for agents in India is projected to be over 1 million, and more than 20% of those positions will be unfilled because of a shortage of available skilled labor.

When a customer calls an inbound call center, various call handling and routing technologies will attempt to route the call to an available agent. However, there is often no agent available to immediately answer the phone call, in which case the customer is typically put on hold and placed in a queue. The customer, in turn, may abandon the queue by hanging up, either immediately after being placed on hold or after waiting for some amount of time without receiving service. Once connected to an agent, a customer will speak with that agent for some random time, after which either the call will be completed or the customer will be "handed off" to another agent or queue for further assistance. The quality of the service is typically viewed as a function of both how long the customer must wait to receive service and the value that the customer attributes to the information and service that is received.

Call center managers are increasingly expected to deliver both low operating costs and high service quality. To meet these potentially conflicting objectives, call center managers are challenged with deploying the right number of staff members with the right skills to the right schedules in order to meet an uncertain, time-varying demand for service. Traditionally, meeting this challenge has required call center managers to wrestle with classical operations management decisions about forecasting traffic, acquiring capacity, deploying resources, and managing service delivery.

In recent years, the call center landscape has been altered by a wide variety of managerial and technological advances. Reduced information technology and telecommunications costs—the same forces that contributed significantly to the growth of the call center industry—have also led to rapid disaggregation of information-intensive activities (Apte and Mason 1995). For call centers, this translated into increased contracting of call center services to third parties (commonly referred to as "outsourcing") and the dispersion of service delivery to locations across the globe ("offshoring"). In addition, advances in telecommunications technologies enabled richer call center workflow, including increasingly intelligent routing of calls across agents and physical sites, automated interaction with customers while on hold, and call messaging that results in automatic callbacks to customers once an agent is available.

Also, as call centers now serve as the "public face" for many firms, there is increasing executive consideration of their vital role in customer acquisition and retention. Similarly, the managerial awareness of call centers' potential to generate significant incremental revenue by augmenting service encounters with potential sales opportunities has also been growing rapidly: for example, a recent McKinsey study revealed that credit card companies generate up to 25% of new revenue from inbound call centers (Eichfeld, Morse, and Scott 2006). However, for call center managers, there is significant additional complexity associated with managing this dual service-and-sales role without compromising response times, service quality, and customer satisfaction.

Finally, every call center manager is acutely aware that phone conversations between customers and agents are interactions between human beings. This suggests that the psychological issues associated with the agents' experience can have a major impact on both customer satisfaction and overall system performance. Although these types of issues have been researched extensively by behavioral scientists, operations management researchers have only recently begun to explicitly include such factors in richer analytic models.

Given the size of the call center industry and the complexity associated with its operations, call centers have emerged as a fertile ground for academic research. A relatively recent survey paper (Gans, Koole, and Mandbelbaum 2003) cites 164 papers associated with call center-related problems, and an expanded on-line bibliography (Mandbelbaum 2004) includes over 450 papers along with dozens of case studies and books. In addition, there have been several more specialized surveys associated with call center operations, including that of Koole and Mandelbaum (2002), who focused on queueing models for call centers; L'Ecuyer (2006), who focused on optimization problems for call centers; and Koole and Pot (2006) and Aksin, Karaesmen, and Ormeci (2007), who both focused on multiskill call centers.

This survey seeks to provide a broad perspective on both traditional and emerging call center management challenges and the associated academic research. The specific objectives and major contributions of this paper are as follows:

1. To provide a survey of the academic literature associated with traditional call center problem areas such as forecasting, queueing, capacity planning, and agent scheduling over the past few years;

2. To identify several key emerging phenomenon that affect call center managers and to catalog the

academic research that has been done in response to these developments;

3. To recognize new call center operations management paradigms that consider the role of the call center in helping firms to attract, retain, and generate revenue from customers and to propose some important implications of these new paradigms on future research;

4. To chronicle research on psychological aspects of call center agent experience, survey recent operations management papers that have incorporated some of these ideas into their modeling, and suggest ways in which such work can be incorporated into future operations management research; and

5. To highlight gaps in the current literature on call center operations management and opportunities areas for future research.

The remainder of the paper is organized as follows. In Section 2, we survey recent work on traditional call center operations management problems. Section 3 reviews research that considers demand modulation as an alternative to supply side management. In Section 4, we look at the research literature that emerged as a result of technology-driven innovations, including multi-site routing and pooling, the design of multi-skill call centers, the blending of inbound calls with other types of workflow such as outbound calls and emails, and increased call center outsourcing. In Section 5, we examine several key human resources issues that affect call centers and chronicle recent operations management research that sought to incorporate some of these factors into their models. In Section 6, we explore research that integrates call center operations with sales and marketing objectives, focusing on cross-selling and long-term customer relationship management. In each of the above sections, we suggest specific opportunities for future research. Concluding comments are provided in Section 7.

2. Managing Call Center Operations: The Traditional View

Traditional operations management challenges for call center managers include the determination of how many agents to hire at what times based on a long-term forecast of demand for services ("resource acquisition") and the scheduling of an available pool of agents for a given time period based on detailed short-term forecasts for a given time period ("resource deployment"). In addition, once initial resource deployment decisions have been made, there may be additional shorter-term decisions to be made, including forecast updating, schedule updating, and real-time call routing.

Resource acquisition decisions must be made several weeks and sometimes months ahead of time because of lead times for hiring and training agents. Also, because most call centers have fairly high employee turnover and absenteeism levels, models that support resource acquisition decisions must explicitly account for random attrition and absenteeism.

Resource deployment decisions are typically made 1 or more weeks in advance of when the calls actually arrive. A cost-effective resource deployment plan attempts to closely match the supply of agent resources with the uncertain demand for services. The (highly variable) demand for resources is expressed in terms of *call forecasts*, which are typically composed of call arrival distributions and service time distributions, both of which vary over time. This variability means that both forecasting and queueing models play an important role in modeling resource deployment decisions. From a scheduling perspective, agents can typically be assigned to a range of shift patterns, and the process of determining an optimal (or near-optimal) schedule has a significant combinatorial complexity.

In addition, as new data about forecasts and agent availability becomes available for a given day or week, this information can be used to modify both the near-term call arrival forecasts and the agent schedules that are driven by them. Finally, as calls actually arrive, there may be specific decisions to be made about queuing policies or call routing.

In this section, we begin our survey by looking at recent work on these call center operations management problems. We focus on call forecasting in Section 2.1, resource acquisition in Section 2.2, and performance evaluation, staffing, scheduling, and routing in Section 2.3. Next, we consider the basic problems of staffing, scheduling, and routing when arrival rates are random in Section 2.4. Finally, Section 2.5 provides a brief overview of developments in performance evaluation models for call centers, reflecting some of the newer characteristics of modern call centers.

2.1. Call Forecasting

Call forecasts are defined by (a) the specific queue or call type associated with the forecast; (b) the time between the creation of the forecast and the actual time period for which the forecast was created (often referred to as the forecasting "lead time"); and (c) the duration of the time periods for which the forecasts are created, which can range from monthly (to support resource acquisition decisions) to short time frames, such as 15-, 30-, or 60-minute periods (to support resource deployment decisions). Over the years, there have been relatively few papers that focused on forecasting call volumes, prompting Gans et al. (2003) to assert that call forecasting was "still in its infancy."

However, in the past few years, there have been a handful of important developments in the call fore-

casting field, driven by increased availability of historical databases of call volumes and by utilization and adaptation of new techniques that have been applied to similar forecasting problems in other application areas.

Weinberg, Brown, and Stroud (2007) propose a multiplicative effects model for forecasting Poisson arrival rates for short intervals, typically 15, 30, or 60 minutes in length, with a 1-day lead time. In their setting, the call arrival rate for a given time interval of a particular day of the week is modeled as the product of the forecasted volume for that day of the week and the proportion of calls that arrive in that time interval plus a random error term. To estimate the model's parameters, the authors adopt a Bayesian framework, proposing a set of prior distributions, and using a Monte Carlo Markov chain model to estimate the parameters of the posterior distribution.

Although computationally intensive, the methodology proposed by Weinberg, Brown, and Stroud (2007) is quite valuable from an operational perspective. In particular, because the model produces forecasts of Poisson arrival rates on an intra-day interval basis, these results can be used in conjunction with performance models and agent scheduling algorithms. In addition, the authors propose a modification of this method to allow for intra-day forecast updating, which can in turn be used to support intra-day agent schedule updating. The paper includes a forecasting case study in which data from a large North American commercial bank's call centers are used to test both the 1-day-ahead forecasts and intra-day forecast updates, with very promising results.

Soyer and Tarimcilar (2007) introduce a new methodology for call forecasting that draws on ideas from survival analysis and marketing models of customer heterogeneity. Specifically, this paper models call arrivals as a modulated Poisson process, where the arrival rates are driven by advertisements that are intended to stimulate customers to contact the call center. The parameters for the call intensity associated with each particular type of advertisement and future time interval are modeled by a Bayesian framework, using a Gibbs sampler (Dellaportes and Smith 1993) to approximate the posterior distributions. The authors also test their methodology by conducting numerical experiments using call volume data from a call center for which all calls can be traced directly to specific advertisements, with the forecasts being created for single- and multi-day time periods.

Shen and Huang (2007) develop a statistical model for forecasting call volumes for each interval of a given day and also provide an extension of their core modeling framework to account for intraday forecast updating. Their model is based on the use of singular value decomposition to achieve a substantial dimensionality reduction, and their approach also decomposes predictive factors into inter- and intra-day features. For the empirical cases presented, the methodology produces forecasts that are more accurate than both the (highly unsophisticated) standard industry practice and the results from Weinberg, Brown, and Stroud (2007); the methodology is also significantly less computationally intensive than the Monte Carlo Markov chain methods of Weinberg, Brown, and Stroud (2007).

Taylor (2007) presents an empirical study that compares the performance of a wide range of univariate methods in forecasting call volumes for several UK bank call centers as well as for the Israeli bank call center data from Brown et al. (2005), considering lead times ranging from 1 day to 2 weeks. Taylor's performance comparison includes methods that have appeared previously in the call center literature, such as seasonal Auto Regressive Moving Average modeling (Andrews and Cunningham 1995) and dynamic harmonic regression (Tych et al. 2002), as well as several other models that have not previously been used for call center forecasting. The latter group includes an exponential smoothing model for double seasonality that was originally developed for forecasting short-term electric utility demand (Taylor 2003); a periodic Auto Regressive model; and a model based on robust exponential smoothing based on exponentially weighted least absolute deviations (Cipra 1992). The empirical comparison showed no clear "winner," because different methods proved to be more effective under different lead times and different workloads.

2.2. Personnel Planning: Resource Acquisition

The call center resource acquisition problem has been studied by a handful of researchers. Gans and Zhou (2002) model a process in which agents are hired and experience both learning and attrition over time, demonstrating that a threshold policy for hiring agents is optimal in their setting. Ahn, Righter, and Shanthikumar (2005) look at a general class of service systems and demonstrate that under the assumption of continuous number of agents who can be hired and fired at will, the optimal policy is of a "hire-up-to/fire-down-to" form. Bordoloi (2004) combines control theory and chance-constrained programming techniques to derive steady-state workforce levels for different knowledge groups and a hiring strategy to achieve these targets. Bhandari, Harchol-Balter, and Scheller-Wolf (2007) consider both the hiring of regular workers and the contracting of part-time workers along with the operational problem of determining how many part-time workers to deploy under different load conditions. Ryder, Ross, and Musacchio (2008) examine the impact of different routing strategies on employee learning in a multi-skill environment in an

attempt to understand the connection between routing, learning, and overall staffing needs.

Given the importance of the resource acquisition decision, there is significant need for additional research in this area, including models for long-term forecasting, personnel planning for general multi-skill call centers, and resource acquisition planning for increasingly complex networks of service providers (as described by Keblis and Chen 2006, for example).

2.3. Personnel Planning: Staffing, Scheduling, and Routing

The traditional approach to call center resource deployment decisions is to attempt to build an agent schedule that minimizes costs while achieving some customer waiting time distribution objectives. As such, targeted staffing levels for each period of the scheduling horizon are typically key inputs to the scheduling and rostering problems. These targets depend on both how much work is arriving into the call center at what times (as estimated by the call volume forecasts and the forecasted mean service times) and how quickly the call center seeks to serve these customers (estimated by some function of the customer waiting time distribution). Once the forecasts and waiting time goals have been established, queueing performance evaluation models are used to determine the targeted number of service resources to be deployed. The actual performance obtained from the deployed resources also depends on the operational problem of allocating incoming calls to these resources dynamically, known as the call routing problem. Our review follows the same hierarchical order that would be followed in the resource deployment problem for call centers: we first review staffing problems, then provide an overview of scheduling and rostering problems, and finally demonstrate how the call routing problem interacts with them.

2.3.1. Staffing Problems. Simulation models and analytic queueing models are the two alternatives to performance evaluation. Mehrotra and Fama (2003) provides an overview of the inputs required for building a call center simulation model, while Koole and Mandelbaum (2002), and Mandelbaum and Zeltyn (2006) are good sources for a detailed overview of queueing models of call centers.

The simplest queueing model of a call center is the M/M/s queue, also known as an Erlang-C system. This model ignores blocking and customer abandonments. The Erlang-B system incorporates blocking of customers. The Erlang-C model is further developed to incorporate customer impatience in the Erlang-A system (Garnett, Mandelbaum, and Reiman 2002). Performance measures and approximations for the Erlang-A system are discussed by Mandelbaum and Zeltyn (2007b). Sensitivity of this model to changes in its parameters is analyzed by Whitt (2006c), where it is demonstrated that performance is relatively insensitive to small changes in abandonment rates.

For most inbound call centers, the management objective is to achieve relatively short mean waiting times and relatively high agent utilization rates. Gans et al. (2003) refer to such an environment as a "Quality and Efficiency Driven" regime. In this context, let R be the system-offered load measured in terms of the mean arrival rate times and the mean service time. The so-called "square-root safety-staffing rule" stipulates that if R is large enough then staffing the system with $R + \beta\sqrt{R}$ servers (for some parameter β) will achieve both short customer waiting times and high server utilization.

This rule was first observed by Erlang (1948) and was later formalized by Halfin and Whitt (1981) for the Erlang-C model (i.e., an M/M/s queue). Its practical accuracy was tested for service systems by Kolesar and Green (1998). This rule was further supported by Borst, Mandelbaum, and Reiman (2004) and Maglaras and Zeevi (2003) under various economic considerations. This rule has since been demonstrated to be robust with respect to model assumptions such as customer abandonment (Garnett, Mandelbaum, and Reiman 2002; Zeltyn and Mandelbaum 2005), an inbound call center with a call-back option (Armony and Maglaras 2004a,b), and call centers with multiple queues and agent skills (Gurvich, Armony and Mandelbaum 2006, Armony and Mandelbaum 2004), which will be discussed in more detail below.

Borst, Mandelbaum, and Reiman (2004) have also identified two other operating regimes: the quality driven and the efficiency driven (ED) regimes, which are rational operating regimes under certain costs structures. In the ED regime server utilization is emphasized over service quality; however, with customer abandonment, this regime can also result in reasonable performance as measured by expected waiting time and fraction of customer abandonment (Whitt 2004b). Whitt has proposed fluid models for system approximation under the ED regime (Whitt 2006a,b) and has shown its applicability in staffing decisions under uncertain arrival rate and agent absenteeism.

Most of the early literature on staffing deals with these problems in settings with a single pool of homogenous agents (see references in Gans et al. 2003; Garnett, Mandelbaum, and Reiman 2002; Borst, Mandelbaum, and Reiman 2004; Atlason, Epelman, and Henderson 2004; and Massey and Wallace 2006). Recent literature on staffing models focuses on multi-skill settings, that is, in call centers where calls of different types are served using service representatives with different skills (Pot, Bhulai, Koole, 2007; Bhulai, Koole, and Pot, 2007; Cezik and L'Ecuyer,

2006; Chevalier and Van den Schrieck, 2006; Harrison and Zeevi, 2004, Wallace and Whitt, 2005, Armony, 2005, Bassamboo, Harrison, and Zeevi 2005, 2006). A different setting with homogeneous agents serving various customer types to whom differentiated service is provided is analyzed by Gurvich, Armony, and Mandelbaum (2006). Aksin, Karaesmen, and Ormeci (2007), Koole and Pot (2006), and L'Ecuyer (2006) survey recent research on multi-skill call center problems.

Typically staffing formulations seek to determine the number of full-time equivalent employees needed given an objective function and some constraints. The most widely used is a staffing cost minimization objective with service level constraints (see, for example, Atlason, Epelman, and Henderson 2004; Cezik and L'Ecuyer, 2006; Bhulai, Koole, and Pot 2007; Jagerman and Melamed, 2004; Mandelbaum and Zeltyn 2007a), although staffing problems with profit maximization objectives have also been proposed (Aksin and Harker, 2003; Koole and Pot, 2005; Helber, Stolletz, and Bothe 2005; Baron and Milner, 2006). Armony et al. (2007) establish convexity properties and comparative statics for an M/M/s queue with impatience, demonstrating the relationship between abandonments and optimal staffing. Koole and Pot (2005a) show that these convexity properties fail to hold when the buffer size is also a decision variable. Canon et al. (2005) formulate the staffing problem as a deterministic scheduling problem.

2.3.2. Shift Scheduling and Rostering. Taking the results from the staffing problem as inputs, typically on an interval-by-interval basis, the shift scheduling problem determines an optimal collection of shifts to be worked, seeking to minimize costs while achieving service levels or other labor requirements. Closely related to the scheduling problem, the rostering problem combines shifts into rosters and provides the actual matching between employees and rosters. The scheduling problem and the rostering problem have been studied extensively, both in the context of call centers (see references in Gans et al. 2003) and in more general contexts (Ernst et al. 2004 chronicles over 700 papers on these topics). In this section, rather than attempt an extensive survey of the scheduling and rostering literature, we instead describe several different approaches to these problems, along with illustrative recent papers and some fruitful directions for future research.

The traditional approach to the scheduling problem is to formulate and solve a mathematical program to identify a minimum cost schedule. Although variants of this approach have been widely utilized, both in the research literature and in industrial applications, over the years several issues have also been identified with this basic method. For large call centers with a single queue of call arrivals and a homogeneous pool of agents, each with several possible shift and break combinations and associated restrictions, the size of the mathematical program grows very rapidly. This issue is addressed by several researchers, most notably Aykin (1996, 2000), who models flexible break constraints for each shift and tests the proposed methodology with several large test problems.

Another problem with the traditional mathematical programming approach is that it requires as input a target agent staffing level for each time interval. This concept of target staffing level is in turn based on the assumption that all agents are able to handle all incoming calls. However, in a multi-queue/multi-skill environment, this assumption is clearly violated, and much of the work in recent years has sought to address this specific shortcoming of the traditional methodology. Fukunaga et al. (2002) propose a hybrid method that combines scheduling heuristics with simulation to simultaneously solve both the scheduling and the rostering problem and discuss a commercial implementation of this method that is used by over 1,000 call centers today. Similarly, Cezik and L'Ecuyer (2006) propose a methodology that combines linear programming with simulation to determine a schedule. Avramidis et al. (2007) develop search methods that use queueing performance approximations to produce agent schedules for a multi-skill call center.

Another stream of research in the area of call center scheduling focuses on eliminating approximations that result from the traditional separation between the staffing and the scheduling problems described above. Motivated by the dependency of adjacent time intervals' waiting time distributions, which is ignored by traditional scheduling algorithms, Atlason, Epelman, and Henderson (2004) use subgradient information for the objective function along with simulation in order to determine agent schedules. In a similar spirit, noting traditional methods assume that service level goals are "hard constraints" that must be met during each interval, Koole and van der Sluis (2003) instead develop a scheduling methodology that seeks to meet only an overall service level objective over the course of an entire scheduling period (typically a day or a week). Ingolfsson, Cabral, and Wu (2003) note that the traditional staffing methods use steady-state staffing models for individual intervals and seek to eliminate errors induced by this approximation by using transient results on a period-by-period basis, which they refer to as the "randomization method," along with integer programming to create agent schedules. Motivated by the potential impact of understaffing on call abandonment, Saltzman (2005) and Saltzman and Mehrotra (2007) develop and test a scheduling methodology that combines linear programming, tabu search, and simulation while including costs to staff,

waiting times, and abandoned calls in the objective function.

The separation of shift scheduling from the actual rostering process presents another potential problem with the traditional approach. In practice, the mismatch between the (ideal) optimal shifts and the (actual) assignment of shifts to individual agents can have a major negative impact on the overall performance of the call center, and this impact is often exacerbated by updates to call forecasts and schedules that result from new information being obtained after the initial schedule has been created. Because of the complexity associated with the coordination of individual agents' preferences and restrictions, many large call centers and multi-site call center operations require agents to "bid" on particular shifts sequentially, with the order of bidding based on factors such as seniority and previous quality of service delivered. Building on this practice (known in the call center industry as "shift bidding"), Keblis, Li, and Stein (2007) investigate an auction-based approach to the problem of matching labor supply with labor demand in a call center, allowing agents to bid competitively for different shifts. In particular, this type of bidding mechanism suggests a method for pricing services for part-time "work at home" agents, while also facilitating real-time schedule adjustments as a result of updated call forecasts. The issue of real-time schedule adjustments in service operations has also been addressed by Hur, Mabert, and Bretthauer (2004), Easton and Goodale (2005), and Mehrotra, Ozluk, and Saltzman (2006).

2.3.3. The Call Routing Problem. The routing problem is a control problem that involves assigning incoming calls to specific agents or pools of agents and then scheduling calls when several are waiting for the same agent pool. This problem has attracted a lot of attention as a call center application and more generally as a challenging queueing control problem (Ormeci, Burnetas, and Emmons 2002; Ormeci, 2004; Gans and Zhou, 2003; Koole, Pot, and Talim 2003; Atar, Mandelbaum, and Reiman 2004a,b; Mandelbaum and Stolyar, 2004; Harrison and Zeevi, 2004b; Armony, 2005; de Vericourt and Zhou, 2006; Bhulai, 2005; Koole and Pot, 2006; Bassamboo, Harrison, and Zeevi 2005; Tezcan, 2005; Atar, 2005a, 2005b; Jouini et al. 2006, Tezcan and Dai, 2006, Gurvich and Whitt, 2007).

The problems of staffing, scheduling, and routing exhibit hierarchical dependency. The call routing problem in multi-skill call centers is also known as skills-based routing. In multi-skill settings, how well calls are routed determines the effectiveness of staff usage, while the staffing problem constrains the routing decision. These problems interact, as explained via examples in Aksin, Karaesmen, and Ormeci (2007) and Koole and Pot (2006), and further interact with the flexibility design problem (Aksin and Karaesmen 2003; Aksin, Karaesmen, and Ormeci 2007). The hierarchical dependency, as well as the close interaction between staffing and routing, make these problems challenging from an operations research perspective (Cezik and L'Ecuyer 2006; Harrison and Zeevi 2005; Armony and Maglaras 2004a, Wallace and Whitt 2005; Bhulai, Koole, and Pot 2007; Gurvich, Armony, and Mandelbaum 2006, Bassamboo, Harrison, and Zeevi 2006; Chevalier and Van den Schrieck 2006). Even when treated in isolation and ignoring important interdependencies, obtaining optimal solutions poses a challenge. Deterministic linear programming, diffusion, or fluid approximations have been proposed to overcome this problem in large-scale centers (Armony and Maglaras 2004a,b; Armony and Mandelbaum 2004; Harrison and Zeevi 2004, 2005; Bassamboo, Harrison, and Zeevi 2006; Whitt 2006a,b; Tezcan and Dai 2006; Gurvich and Whitt 2007). Other papers use simulation in combination with optimization (Atlason, Epelman, and Henderson 2003; Atlason, Epelman, and Henderson 2004; Cezik and L'Ecuyer 2006), loss system, or other approximations (Koole and Talim, 2000, Chevalier and Tabordon, 2003; Koole, Pot, and Talim 2003; Shumsky 2004; Chevalier, Shumsky, and Tabordon 2004; Koole and Pot 2005b; Chevalier and Van den Schrieck 2006; Franx, Koole, and Pot 2006; Avramidis et al. 2006) to enable analysis.

Despite the large number of papers discussed in this section, we believe that there are significant research opportunities with these classical problems. In particular, capturing more of the dependency and interaction among staffing, scheduling, and routing is a promising direction for further research.

2.4. Personnel Planning under Arrival Rate Uncertainty

Historically, most of the papers in the call center literature have modeled the arrival process to be a time-inhomogeneous Poisson process and, thus, forecasting call volumes is in most cases (implicitly or explicitly) equivalent to estimating the time-dependent Poisson arrival rates. This assumption is in many cases quite reasonable. For example, Brown et al. (2005) conducted an extensive empirical study of historical data from an Israeli bank's call center operations and conclusively failed to reject the hypothesis that the call arrivals follow a time-inhomogeneous Poisson process; however, in the same study, after using call type, time of day, and day of week to build an empirical model to forecast the call arrival rates for short time intervals, the authors concluded that the Poisson arrival rates are not easily predictable.

Because of the difficulty of accurately forecasting call arrival rates, several researchers have explored the

implications of modeling call arrivals with a random arrival rate. Whitt (1999b) suggests a particular form of a random arrival rate for capturing forecast uncertainty. Chen and Henderson (2001), Avramidis, Deslauriers, and L'Ecuyer (2004), Brown et al. (2005), and Steckley, Henderson, and Mehrotra (2005) point out the randomness of arrivals in real call centers, a feature that is ignored in most of the literature. Steckley, Henderson, and Mehrotra (2005), Harrison and Zeevi (2005), Robbins et al. (2006), and Torzhkov and Armony (2007) analyze call center performance under random arrivals. Thompson (1999) and Jongbloed and Koole (2001) provide methods for determining target staffing when the arrival rate is random. Ross (2001, Chapter 4) offers extensions to the square-root staffing rule to account for a random arrival rate. Robbins et al. (2007) consider the question of cross-training a subset of agents from different queues to meet demand in the presence of uncertain arrival rates. Other recent papers that focus on planning problems in the presence of random arrivals are those by Steckley, Henderson, and Mehrotra (2007), Whitt (2006e), Baron and Milner (2006), Bassamboo and Zeevi (2007), and Aldor-Noiman (2006).

Another traditional call center modeling assumption is that the arrivals during one time period within a planning horizon are independent of the arrivals in the other time periods for purposes of determining staffing levels and agent schedules. Green, Kolesar, and Soares (2001, 2003) have dubbed this the *stationary, independent, period by period* method. However, several empirical studies have demonstrated that for many call centers there is significant correlation in call volumes across time periods. Brown et al. (2005) develop a non-linear least squares model in which a previous day's call volume is an independent variable in predicting the subsequent day's call volume, producing roughly a 50% reduction in the variability of the forecasted daily volumes. Motivated by empirical analysis of a large telecommunication firm's call centers that demonstrates both greater-than-Poisson variability and strong correlation across time periods within the same day, Avramidis, Deslauriers, and L'Ecuyer (2004) develop and test several analytic models in which the arrival *rate* for each interval of the day is a random variable that is correlated with the arrival rates of the other intervals. Steckley, Henderson, and Mehrotra (2005) analyze data from several call centers and identify significant cross-period correlation in call volumes; motivated by these results, Mehrotra, Ozluk, and Saltzman (2006) present a framework for intraday forecast and schedule updating that utilizes the call arrival model of Whitt (1999b) to model cross-period correlation.

We believe that this points to at least two interesting and important areas for future research. First, there is a need for research into additional performance analysis models under different arrival rate variability assumptions, as well as for more validation of such assumptions with operational data. Second, reconsidering the scheduling and rostering problems under the more general assumption that arrival rates are random variables is another very promising area that is just now beginning to receive attention from researchers. For example, Robbins and Harrison (2007) view arrival rate variability as a fundamental component of the agent scheduling problem and propose a stochastic programming solution to determine the best combination of agents and shifts that explicitly accounts for the risk inherent in the arrival rate uncertainty.

2.5. Performance Evaluation for Modern Call Centers

As call centers have evolved in terms of size and configuration, and as more empirical analysis has shed light on the features of typical queueing model primitives like arrivals, abandonment, and service times in these centers, new performance evaluation models have been developed and analyzed. These models are motivated by different features of modern call centers, as well as empirically observed characteristics of queueing model primitives. The latter analysis has been initiated by a research collaboration between researchers at The Technion and The Wharton School that has provided a clean source of customer call-based call center data from several sources, which has subsequently been developed into a complete platform for data-based analysis of call center problems (a description of the DataMOCCA Project can be obtained from http://iew3.technion.ac.il/serveng/References/DataMOCCA). The important distinction of the data provided in this project is that unlike typical call center data that averages data over time intervals, these data are on a per-call basis, thus enabling deeper analysis as well as a more natural tie to marketing- or human resource-related analyses. Further use of this type of data to explore the links between call center operational problems and human resource and customer related issues is a promising direction for future research.

Large call centers have motivated the analysis of heavy traffic limits as useful approximations of queueing models (see, for example, Halfin and Whitt 1981, Garnett, Mandelbaum, and Reiman 2002, Jennings et al. 1996, Whitt 2004a,b). Motivated by recent empirical studies demonstrating that service times and abandonment times are not necessarily exponentially distributed (Mandelbaum, Sakov, and Zeltyn 2000; Brown et al. 2005), models with general service times and general abandonment times have been analyzed and approximations for their performance developed

(Whitt 2004b, 2005, 2006c; Reed 2005, Zeltyn and Mandelbaum 2005, Jelenkovic, Mandelbaum, and Momcilovic 2004, Mandelbaum and Momcilovic 2007, Gamarnik and Momcilovic 2007, Kaspi and Ramanan 2007). Mandelbaum and Zeltyn (2004) explore a linear relationship between the probability to abandon and the waiting time in queue in an Erlang-A model. Although such linearity should not exist in the presence of general impatience distributions, empirical evidence by Brown et al. (2005) suggests a similar linear relationship. Mandelbaum and Zeltyn (2004) analyze the problem both theoretically and empirically and demonstrate that, over realistic parameter values, general impatience distributions result in performance that resembles the Erlang-A model. This is an important result, supporting the robustness of the Erlang-A model, even in settings with non-exponential impatience times. Similarly, as reviewed in more detail in Section 2.4, Steckley, Henderson, and Mehrotra (2005), Harrison and Zeevi (2005) and Torzhkov and Armony (2007) analyze call center performance under random arrivals.

Blocked or abandoned calls may redial later, which is a feature ignored in most models. This type of retrial behavior and its influence on performance is modeled by Mandelbaum et al. (1999) and Aguir et al. (2004). Approximations, in particular a fluid approximation, perform very well for such systems. The use of fluid approximations in the presence of time-varying parameters is also supported by Ridley, Fu, and Massey (2003) and Jimenez and Koole (2004). The need to manage multi-skill call centers has led to performance evaluation models for systems with flexible servers (Chevalier and Tabordon 2003; Shumsky 2004; Stolletz and Helber 2004; Whitt 2006a; Franx, Koole, and Pot 2006).

We believe that performance evaluation will continue to provide research opportunities, particularly in light of the developments described in Sections 3 and 4 below.

3. Demand Modulation

Many call centers face highly unpredictable demand that is also time-varying. The time-varying element is relatively easy to handle by adjusting staffing levels. Papers by Jennings et al. (1996), Massey (2002), Ridley, Fu, and Massey (2003), Feldman et al. (2005), and Green, Kolesar, and Whitt (2007) are examples of papers that consider the staffing problem under time-varying demand. But when call volume is unpredictable, limited flexibility in adjusting staffing levels may lead to situations of over- or under-staffing, at least temporarily. This section deals with means of modulating demand as a way of ensuring load balancing and higher level of predictability. Demand modulation is also used to reduce operating costs by encouraging callers to obtain service through other channels, such as the Internet, that are more scalable or less expensive.

The simplest form of demand modulation that may be used in call centers is call admissions. The most primitive form of call admission is a busy signal that customers encounter every time all lines are busy. Given costs of infrastructure, such busy signals are very rare in medium to large call centers and non-bursty call volume. A more sophisticated form of call admission can be done by selectively admitting calls according to their relative importance to the organization (Ormeci, 2004). This practice is also very unusual in call centers. Bassamboo, Harrison, and Zeevi (2006) demonstrate that under some circumstances it is beneficial not to admit less profitable customers so as to reduce the chances of losing more profitable ones later on.

Regardless of whether a call center regulates its calls through an admission control mechanism, one fact that call center managers must face is that callers are inherently impatient. If a customer call is not answered within a certain time, the customer will hang up (abandon) and subsequently may either retry later or not. Generally, call center managers strive to minimize the number of abandonments, because of the premise that abandonments are associated with a negative waiting experience and might lead to loss of goodwill and even to churn. However, abandonments also have a positive component associated with them, because they provide a natural mechanism for load balancing. To wit, when the system is heavily loaded impatient customers tend to abandon, alleviating the workload and hence shortening the waiting times of the more patient callers.

Because of the importance of abandonment in determining staffing levels, there has been a stream of literature that focuses on understanding customer abandonment (Hassin and Haviv 1995; Mandelbaum and Shimkin 2000; Zohar, Mandelbaum, and Shimkin 2002; Shimkin and Mandelbaum 2004) and its impact on system performance (Garnett, Mandelbaum, and Reiman 2002; Mandelbaum and Zeltyn 2004; Zeltyn and Mandelbaum 2005; Armony, Plambeck, and Seshadri 2007; Mandelbaum and Zeltyn 2006; Baron and Milner 2006; Mandelbaum and Zeltyn 2007b).

Acknowledging that overloaded situations and abandonments will always exist, researchers have proposed that notifying callers of their anticipated delay as soon as they call would cause impatient customers to leave right away (balk), whereas the more patient customers are likely to wait until their call is answered. Whitt (1999a) has demonstrated that the overall average waiting time of all customers is reduced if delay announcement is accurate. Guo and Zipkin

(2006, 2007a) have identified cases in which information improves performance, but have also demonstrated that such information can actually hurt the service providers or the customers under exponential service time and more general phase-type distributions. Guo and Zipkin (2007b) noted that the effect of information on total throughput depends on the shape of the distribution describing the customers' sensitivity to delay. In their analysis, Guo and Zipkin compare a system with delay information to a system in which the decision on whether to join the queue is based on expected steady-state delay equilibrium. This equilibrium analysis is similar to the approach taken by Whitt (2003), where it is assumed that the balking decision is based on expected steady-state delay equilibrium, and it is demonstrated how the demand scales with respect to the number of servers. Jouini and Dallery (2006) consider how to estimate callers' waiting time and what information to announce to callers in a multiple-customer setting with a fixed priority sequencing rule. The above papers all assume that if a customer has decided to stay given the announced information, he will subsequently remain in the system until his service ends.

Given that delay announcements in a stochastic environment are inevitably inaccurate, it is plausible that callers may abandon the system even if initially they decided to stay and wait for their service. Armony, Shimkin, and Whitt (2006) propose a model in which callers may balk in response to a delay announcement, but provided they do not balk their time-to-abandon distribution is also dependent on the same announcement. Armony, Shimkin, and Whitt (2006) proposes as a delay announcement scheme the delay of the last caller to enter service, which is numerically shown to be very accurate in large overloaded systems. A closely related scheme of announcing the delay of the first customer in line has been proposed by Nakibly (2002). Similar to Armony, Shimkin, and Whitt (2006), in a single class setting, Jouini, Dallery, and Aksin (2007b) consider a model where customers are allowed to abandon subsequent to delay announcements. The possibility of announcing different percentiles of the delay distribution is proposed and the relationship between performance and announcement precision is explored. The paper demonstrates that announcements with higher precision are not universally preferred. Finally, in the context of delay announcement in call centers Jouini, Dallery, and Aksin (2007a) published the first paper to model delay announcement in a multiple-customer class setting with priorities. In this setting, future arrivals to the higher priority class may increase the delay of lower priority callers.

In addition to abandonment, load balancing can also be done by encouraging callers to use an alternative service channel when the system is overloaded. Such an alternative channel could be a Web site or an e-mail service request, but could also come in the form of suggesting to the customer to call at a less busy time or to leave a number and be called back later. Armony and Maglaras (2004a,b) propose a model in which callers are given a choice of whether to wait on line for their call to be answered or to leave a number and be called back within a specified time. They show that this call-back scheme allows the system to both increase throughput and reduce average waiting times.

Most call center papers consider the call volume to be an exogenous factor, an external stream of calls. However, many calls are in fact redials of callers who have been blocked (busy signal) or abandoned or have not had their call resolved. A generic name for such calls is retrials. Recognizing the significance of call resolution on overall customer satisfaction and on the system load, many call centers include in their compensation schemes to their customer service representatives (CSRs) a number-of-resolved-calls component. For example, de Vericourt and Zhou (2005) consider a system in which agents differ with respect to two quality dimensions: service speed (rate μ) and probability of call resolution (p). A caller whose call has not been resolved will call the center again with the same concern. In this paper, the authors consider the problem of routing calls to CSRs to minimize the total number of calls in the system. They show that the routing policy that routes calls to the CSR with the highest product $p\mu$ is optimal under certain conditions. Armony (2007) considers this problem for a system with many servers (who are grouped in multiple pools consistent with their service rate and call resolution probability) and demonstrates that the same $p\mu$ policy is asymptotically optimal in the sense that it minimizes the queue length and waiting times in steady state. Mehrotra, Ross, and Zhou (2007) consider such an environment with multiple pools of agents and multiple classes of customers and examine several routing policies in an attempt to simultaneously maximize call resolution rates and minimize customer waiting times.

Aguir et al. (2004) consider how retrials impact the performance of call centers. They propose a fluid model to approximate the queue length process, which tends to be accurate for large overloaded systems. Using numerical analysis they demonstrate that erroneously considering retrials first-time calls can lead to very significant distortions in forecasting and staffing decisions. In a subsequent paper, Aguir et al. (2007) demonstrate that, surprisingly, ignoring retrials by considering them first-time calls can lead to under- or over-staffing with respect to the optimal staffing level, depending on the forecasting assumptions.

Our discussion thus far with respect to load balanc-

ing has focused on the overloaded periods and how to postpone some of the load for a later, less congested period. But load balancing could also be done by staffing to meet peak load demand and doing other necessary work (see, e.g., Gans and Zhou 2003; Bhulai and Koole 2003) when call volume is low. One such activity that has become an integral part of many call centers in recent years is cross-selling. Cross-selling may be defined as selling a product to the caller, which is not the primary reason the caller has contacted the center for. Although cross-selling can be regarded as a load-balancing activity, its primary purpose in most cases is to generate revenue. Hence, we will review the cross-selling literature as part of Section 6.

As is the case with any service system, consumer psychology also plays a big role in call centers. There is a rich literature on consumer psychology and particularly on customers' delay perception when they wait in queue. Some examples include the papers by Maister (1985), Hui and Tse (1996), Hui and Zhou (1996), Carmon and Kahenman (2002), and Munichor and Rafaeli (2006).

Incorporating findings from such behavioral research and performing further behavioral experiments to confirm modeling assumption is another underexplored research direction that could potentially lead to more practical demand modulation schemes. For example, Munichor and Rafaeli (2006) demonstrate that callers are more satisfied when delay announcements are made, especially if these are made periodically, and give callers a sense of progression in terms of their position in line. Investigating the operational impact of multiple delay announcements during a caller's wait is a promising direction for future research. Other promising directions are to incorporate findings from real call center data (e.g., Feigin 2006) and customer choice models from the Economics literature (e.g., Gonzales-Simental and Pines 2006).

4. Technology-Driven Innovations and Challenges

Over the past decade, several technological advances have had a profound impact on the call center industry. The deregulation of the telecommunications industry has increased competition, leading to increased network capacity, improved quality, and lower costs for both domestic and international traffic. In addition, automatic call distributor and computer telephony integration technology has grown cheaper, more reliable, and increasingly sophisticated. Finally, as businesses have increasingly focused on their respective core competencies, these advances in telecommunications infrastructure have made it easier for companies to contract all or part of their call center operations to third-party firms known as outsourcers, many of whom have all or part of their operations in an offshore location in another country

In this section, we examine some of the key operations management implications of these industry changes. In Section 4.1, we discuss multi-site call centers and survey the research literature associated with the management decisions that are required under this type of operational structure. In Section 4.2, we look at pooling and design issues in call centers in which multiple types of phone calls or other types of customer traffic (e-mails, outbound calls, etc) are present. In Section 4.3, we explore the phenomenon of call center outsourcing, focusing on recent research associated with contract structures and incentives.

4.1. Multi-Site Operations

Once the decision has been taken to operate an in-house call center, important design issues must be addressed. The most basic of these pertains to the number of sites to establish. Most large companies opt for a multi-site structure, where multiple sites allow for geographic risk mitigation and enable tapping different labor pools. The decision of a single-site versus multi-site structure is typically a strategic one and has not been addressed in the operations management literature. In a multi-site structure, a further decision pertains to the possibility of virtually pooling these sites. Investing in appropriate technology will enable virtual pooling, thus making the operations of a multi-site center virtually identical to those of a single-site one. Some call sharing and routing problems for a multi-site call center where technology for virtual pooling between sites is not available have been analyzed by Aguir (2004). Aguir demonstrates that good routing policies result in performance close to what can be obtained through virtual pooling. Tezcan (2005) provides further evidence that smart routing policies in distributed call centers can achieve performance optimizing and load balancing results approaching that of a virtual call center.

Because the specifics of multi-site routing problems are determined by the technology in place, they tend to be application specific and have been mostly analyzed by practitioners. An interesting interaction exists between routing of calls and site utilization. Call routing schemes that send more calls to sites with higher efficiency (e.g., the faster-server-first scheme proposed by Armony (2005)) will lead to higher utilization of servers at those sites. This is not always desirable from a human resource perspective, as further elaborated upon in Section 5 below. Servers at such a site will feel overloaded, whereas those at the less efficient sites will have less opportunity to learn. Understanding this relationship among call routing,

site utilization, and human resource well-being and learning is an interesting future research area.

4.2. Pooling and Design of Multi-Skill and Blended Call Centers

The problem of pooling in queueing systems has a long history in the operations literature (see, for example, Buzacott 1996; Mandelbaum and Reiman 1998). Rather than review the literature on pooling in queueing systems, here we point out some recent papers that have explored the pooling issue specifically in the context of call centers. Pooling is also related to the design of flexibility and to skill-based routing problems and is further reviewed in those contexts below.

Pooling several specialist groups into larger pools with cross-trained agents, and its performance effects are analyzed by Tekin, Hopp, and van Oyen (2004). Numerical analysis explores the sensitivity to pool size and call parameters. This analysis is in line with earlier studies, demonstrating that under certain conditions pooling queues has operational advantages. Jouini, Dallery, and Nait-Abdallah (2006) look at the opposite problem, that of partitioning pooled structures into specialized teams. Using the case of a real call center, they illustrate the benefits of this type of partitioning from an organizational behavior and management perspective. They further indicate that the disadvantage from a pooling perspective can be overcome by allowing a limited amount of flexibility in the specialized team structures. Hu and Benjaafar (2006) demonstrat that in settings with customer classes having non-homogeneous service requirements and the possibility for rush hour-induced peaks, server partitioning is beneficial, albeit at the expense of some classes.

Changing characteristics of call centers in terms of functionality, customer types, and agent skills has generated a large interest in multi-class/multi-skill call center problems. In the design of such call centers, one of the key questions for an operations manager is to determine the appropriate type and level of flexibility. More specifically, the flexibility design problem investigates skill set design for flexible call center employees, as well as the right mix of flexible and specialized agents. The flexibility design problem and associated literature is reviewed in detail by a recent review article on cross-training in call centers (Aksin, Karaesmen, and Ormeci 2007). Most of this literature builds on the analysis of a product-plant network in the article by Jordan and Graves (1995) focusing on a manufacturing setting. Aksin and Karaesmen (2003, 2007) consider the problem in the context of call centers and demonstrate that certain flexibility principles also hold in this setting. These principles pertain to the benefits of flexibility and are that limited flexibility is almost as good as full flexibility; skill-sets should be established to form long-chain structures such that neighboring skill sets share a skill, allowing calls to be offloaded during times of congestion; and in systems with balanced arrival rates and revenues, skill sets should be balanced as well. The two-skill structures of Wallace and Whitt (2005) and Mazzuchi and Wallace (2004) provide additional support for these principles in call center settings.

Determining the appropriate mix of specialized and flexible servers is the second major issue that has been addressed within the flexibility design problem. Aksin, Karaesmen, and Ormeci (2005) present some results for the Jordan and Graves framework and demonstrate that the marginal value of an additional cross-trained server is decreasing in the number of existing cross-trained servers, indicating a trade-off between value and cost that must be managed. Pinker and Shumsky (2000) address this question for a call center, where additionally the quality trade-off is modeled. Chevalier, Shumsky, and Tabordon (2004) suggest a 20% cross-training rule of thumb, which through numerical examples they demonstrate to be quite robust to different cost parameters. Through a simulation study, Robbins et al. (2007) demonstrate the same type of diminishing returns property in a call center setting with uncertain demand and service level constraints. These papers consider the extremes of full flexibility and specialists, and further exploration of this question in settings with limited flexibility remains to be done. The result of Jouini, Dallery, and Nait-Abdallah (2006), demonstrating that a fully pooled structure's performance can be achieved by specialist groups that can handle a small proportion of calls from other teams, seems to provide additional support to Chevalier, Shumsky, and Tabordon (2004). In technical support centers or some medical call centers, calls flow between flexible and specialized agents in a hierarchical fashion, starting with gatekeepers and escalating higher up to specialists (Shumsky and Pinker 2003). In this regard, the staffing and routing problem analyzed by Hasija, Pinker, and Shumsky (2005) addresses the question pertaining to the mix of flexible versus specialized servers in such multi-tier call centers.

The flexibility design problem is closely related to the staffing and routing problems described in Section 2. This interaction, as well as the interaction with human resource management (reviewed in Section 5), is elaborated by Aksin, Karaesmen, and Ormeci (2007). Further analyses, characterizing how the flexibility design question is answered in conjunction with staffing and control and how the skill-set design interacts with human resource well-being and performance, will constitute valuable additions to research as well as important contributions to call center management.

A multi-skill call center in which inbound and outbound calls or phone and e-mail calls are combined is known as a "blended" operation. The key distinction of problems with blending comes from the fact that e-mail calls or callbacks have less urgency and can be inventoried to some extent, relative to phone calls. The call blending problem has led to research on performance evaluation (Bernett, Fischer, and Masi 2002; Pichitlamken et al. 2003; Deslauriers et al. 2007) and analysis of blending policies (Gans and Zhou 2003; Bhulai and Koole 2003; Armony and Maglaras 2004a,b). Keblis and Chen (2006) consider a staffing problem in a setting with blending as well as "co-sourcing," which is defined in the next section.

4.3. Call Center Outsourcing and Service Contracting

Call centers can be managed in-house or within shared service organizations sometimes run as separate business entities (Aksin and Masini 2006). Increasingly, call center operations are outsourced to companies that specialize in running other companies' call centers. Partial outsourcing is also common, where some calls are kept in-house and others are outsourced. This is known as co-sourcing (Aksin, de Vericourt, and Karaesmen 2006; Ren and Zhou 2006). Some of this outsourcing is directed to companies or sites that are abroad, thus taking the form of offshoring. Companies outsource all or part of their calls for economic or strategic reasons: to lower costs, to benefit from economies of scale, to obtain additional capacity and flexibility, or to benefit from the technological capabilities of the sourcing firm. The vast interest in outsourcing and offshoring has motivated some recent call center research.

Like other supplier relationships, the success of call center outsourcing projects hinges on the contracts in place and their implementation. Understanding and modeling existing contracts, as well as proposing new ones that overcome problems of prevailing contract types, constitutes a fruitful area for research, in which some initial steps have been taken. Aksin, De Vericourt, and Karaesmen (2006), Ren and Zhou (2006), and Milner and Olsen (2006) address different problems related to call center outsourcing contracts. The difficulty of embedding queueing models in a contract analysis problem are overcome by approximating call center performance via fluid approximations or by considering heavy-traffic approximations. Gans and Zhou (2007) focus on an implementation problem that addresses the question of how precisely calls will be shared in a co-sourcing contract between a client and a contractor firm. Keblis and Chen (2006) propose a solution to Amazon.com's large-scale capacity planning problem, featuring several internal call centers as well as co-sourcing agreements with several external service providers.

Motivated by a real example, Aksin, de Vericourt, and Karaesmen (2006) analyze two outsourcing contracts offered to a user firm by a price-setting contractor firm: one in which a flexible volume of calls can be outsourced to the contractor to whom payment is made for utilized capacity on a per call basis and another in which a fixed level of capacity is reserved and paid for, irrespective of subsequent utilization. Optimal capacity decisions by both parties and optimal pricing decisions are characterized in a multi-period setting with uncertain demand. The paper investigates the economic or operational settings that lead to a preference for complete outsourcing or partial outsourcing in the form of co-sourcing. Contract preferences for each firm are shown to change depending on cost parameters, as well as demand uncertainty within and between time periods.

Ren and Zhou (2006) explore settings in which the contractor determines staffing and exerts effort that influences service quality, operationalized as the number of calls that are served and resolved. After showing that both a piece-meal and a pay-per-call-resolved contracts can coordinate the staffing decision, however, fall short of the system optimum for the service quality dimension, the authors propose contracts that will coordinate on the quality dimension. They highlight the importance of service quality contractability in call center outsourcing settings.

Hasija, Pinker, and Shumsky (2007) examine a wide variety of contract structures used by a large firm that makes extensive use of outsourcers. In particular, they look at how different contract terms (including pay-per-time, pay-per-call, and constraints on service levels and abandonment rates) can be utilized by the contracting firm in the presence of information asymmetry about workers' productivity.

Rather than comparing different contract structures, Milner and Olsen (2006) explore the role service level constraints in outsourcing contracts play in settings where the contractor firm has both contractual and non-contractual clients. They demonstrate that with such contracts, it is rational for the contractor firm to provide service priority to the calls of the contract client, mostly during its own off-peak times, ensuring the satisfaction of service levels on the average but violating them during its own peak times. Contracts that include measures of the variability of delay are shown to alleviate this problem. Baron and Milner (2006) define a period-based service level to overcome this problem and explore how this type of a constraint affects the call center staffing problem. Although not in the context of outsourcing, Koole (2005) identifies the problematic nature of the typical service level measure in call centers and proposes a different waiting

time performance metric that measures the fraction of time that waiting exceeds an acceptable level.

The contracting literature focusing on outsourcing contracts in call centers assumes common information that is shared with all involved parties. Hasija, Pinker, and Shumsky (2007) relax this assumption and allows for information asymmetry about worker productivity between a client firm and a service provider. Different contract forms in practice are analyzed in this setting, particularly focusing on the issue of coordination between the two parties.

At a more operational level, Gans and Zhou (2007) analyze the routing problem faced between a user company and contractor. The setting is one where the user company is co-sourcing, treating high value customers' calls in-house and outsourcing low-value customers' calls. This firm's objective is to maximize low-value customers throughput subject to a high-value customer service level constraint. The outsourcing company minimizes staffing costs subject to the low-value customers' service level constraint. The paper compares the performance of four call routing schemes that differ in their complexity in terms of both technological requirements and coordination needs between the firms. Keblis and Chen (2006) consider the capacity planning problem for a multi-site, multi-vendor, phone and e-mail blending call center using a mathematical programming approach. Their formulation explicitly takes different costs and constraints imposed by various co-sourcing agreements with their contractors into account. This indicates a rich area of future research in call center staffing for call centers with co-sourcing, in which traditional staffing problems are analyzed within the economics of outsourcing or co-sourcing contracts.

5. Human Resource Issues in Call Centers

Whereas the operations management literature examines personnel planning problems from the perspective of the call center manager or the firm, all of these decisions, i.e., staffing, shift scheduling, rostering, and routing control, affect the employees of a call center as well. These effects, in turn, together with employee incentives, influence call center performance and are typically ignored in operations management models. We next view the human resource management of call centers from an organizational behavior perspective, identifying different human resource practices and how they relate to call center performance. Holman (2005) provides an extensive review of the literature in this domain. Our emphasis will be on illustrating the ties between the operations management and organizational behavior perspectives on human resource issues in call centers.

The key trade-off between customer service and efficiency faced by an operations manager in a call center is also the central tension that a human resource manager must manage. According to Houlihan (2002), "This tension unmasks a series of conflicts: between costs and quality, between flexibility and standardization and between constraining and enabling job design." Whereas traditional call center human resource strategies are characterized by control oriented practices, there is some evidence of commitment strategies in the literature (see, for example, Houlihan 2002; Deery and Kinnie 2002; Batt 2002; Batt and Moynihan 2002). Tayloristic practices characterize a control-oriented call center. High involvement practices, such as selective hiring and extensive training, job designs that include individual discretion and allow for ongoing learning, and incentives such as training, security, high pay levels, and trust building performance measurement systems, characterize a commitment strategy (Batt 2002). Batt and Moynihan (2002) and Houlihan (2002) point out various alternative production models for call centers driven by different market segments or internal needs.

Employees of a call center feel the tension between control and commitment in part through performance measurement systems. Call centers monitor both quantitative (calls per hour, average call times, time between calls, etc.) and qualitative (content, style, adherence to policies, etc.) aspects of calls answered by an employee. Target setting is extensively used to ensure performance along both the quantitative and the qualitative dimensions (Bain et al. 2002). This type of incentive system is typically associated with a Tayloristic control-oriented view of work. Quantitative and qualitative targets may furthermore be conflicting, thus creating additional pressure on employees. This conflict combined with the intensity of monitoring is believed to lie at the root of call center employee burnout, leading to negative effects like turnover, absenteeism, and quality problems. The Harvard Business School case (9–694–047), "A Measure of Delight: the Pursuit of Quality and AT&T Universal Card Services," documents the complexity and associated tensions created by call center performance measurement systems. Holman, Chissick, and Totterdell (2002) provide some evidence that monitoring that enables employees to enhance their skills and service performance has a positive effect on their well-being, whereas the intensity of monitoring or the perception thereof has a strong negative effect on well-being. Other characteristics of commitment strategies have also been shown to lead to positive performance outcomes. There is empirical evidence that high involvement practices lead to higher sales and quality and lower quit rates (Batt 2002). Teams that are able to create a collaborative environment are shown to have

better knowledge sharing capabilities, thus leading to better service (Moynihan and Batt 2001). Nevertheless, in practice, control-oriented models dominate call center management practices, and call center employees exhibit a high incidence of burnout.

The most common definition of burnout considers emotional exhaustion, depersonalization, and diminished personal accomplishment the three components of this stress syndrome (Maslach and Jackson 1981). High emotional exhaustion, characterized by fatigue and a feeling of lack of emotional resources, tends to be related to jobs involving frequent and intense interpersonal contacts (Maslach and Jackson 1981; Cordes and Dougherty 1993). More specifically, CSRs have experienced emotional exhaustion (Cordes and Dougherty 1993; Singh, Goolsby, and Rhoads 1994). Role ambiguity or conflict in an employee's job description and overload in terms of not having the required skills or not having the required time to complete a task further contribute to emotional exhaustion.

As described above, call centers naturally provide most of the antecedents of burnout. A CSR in a call center is in contact with a large number of customers. These customers call with problems in many cases, thus making the contact one in which the customer has a negative attitude and may be aggressive (Grandey et al. 2004). CSRs in a call center are closely monitored for speed and quality. In many environments, speed and quality provide conflicting goals, although a CSR is expected to perform well on both dimensions. Because operational effectiveness of a call center is very important, a CSR will typically receive one call after another, with very little or no time between calls, thus resulting in a general sense of overload. It has also been demonstrated empirically that call center employees are susceptible to emotional exhaustion (Singh, Goolsby, and Rhoads 1994; Von Emster and Harrison 1998; Deery, Iverson, and Walsh 2002; Witt, Andrews, and Carlson 2004). According to Cordes and Dougherty (1993), emotional exhaustion is the first stage of burnout.

Some consequences of burnout in call centers are turnover, absenteeism, increased rework in certain settings, and inability to meet quantitative volume requirements or targets (Tuten and Niedermeyer 2004; Deery, Iverson, and Walsh 2002; Workman and Bommer 2004; Witt, Andrews, and Carlson 2004). Especially, turnover and absenteeism have important and direct economical implications. Turnover not only increases hiring costs, but also affects performance because of the presence of learning curves for new employees (Batt 2002). Absenteeism results in unplanned under-staffing, leading to bad customer service and additional fatigue of those who are present. According to Mercer Human Resource Consulting, call centers experienced an average 33% of turnover in 2003 (www.incoming.com/statistics). A survey of 658 call center workers by Australian Service Union revealed that 88% found their work stressful and almost a third took time off from work because of stress (www.incoming.com/statistics).

Bakker, Demerouti, and Schaufeli (2003) provide evidence that job demands (like work pressure or changes in tasks) are important predictors of health problems leading to absenteeism, and job resources (like social support, coaching, performance feedback) are the predictors of involvement, determining turnover intentions. Combined with earlier cited evidence relating monitoring, job design, or other human resource practices to the well-being and resulting performance of call center employees, it is clear that a good understanding of the relationship between human resource practice and performance outcomes in call centers will enable better management of the quality–efficiency trade-off. In addition, Batt and Moynihan (2002) indicate the need for a better understanding of the call center operations management literature by researchers in organizational behavior to fully understand and manage this trade-off.

It is not surprising that planning and control of human resources interact with other human resource practices and jointly influence performance outcomes. For example, according to a proposition stated by Cordes and Dougherty (1993), "High levels of work demands are the primary determinants of emotional exhaustion. These demands include work overload, role conflict, and direct, intense, frequent or lengthy interpersonal contacts." By adjusting staff levels or by differentiating the type of work through call blending or better skills-based routing, call center managers can control the workload of servers, thus influencing one of the most important reasons for burnout. Similarly, the attainment of some targets by employees depends on the staffing planning and control dimension. Target measures like calls per hour, etc., can be affected by staffing and rostering decisions (Bain et al. 2002). High call abandonment as a result of high absenteeism or bad rostering may result in missing targets. Overstaffing may also prevent certain quantitative targets from being met because of the lack of a sufficient number of calls per employee. Further empirical research that explores the relationship between operational planning and control and human resource performance is necessary to enable future modeling work at this interface.

Some recent research has sought to take elements of this interaction between human resources and operations into account. The trade-off between efficiency and quality was first modeled explicitly by Pinker and Shumsky (2000). More specifically, the authors model the trade-off between cost efficiency due to economies

of scale resulting from cross-trained staff and quality benefits from experience-based learning in specialists. Experience-based learning, as well as employee turnover, is present in the models of Gans and Zhou (2002), Easton and Goodale (2002), and Whitt (2006d). Whitt (2006d) notes satisfaction leads to better retention, resulting in higher experience levels, leading to better performance. Both Gans and Zhou (2002) and Easton and Goodale (2002) consider the staffing problem, albeit at different time scales, with learning and turnover. Easton and Goodale (2002) further allow for random absenteeism of servers. Whitt (2006e) analyzes the staffing problem with absenteeism and random demand.

The quality–efficiency trade-off has also been addressed in the context of call routing. De Vericourt and Zhou (2005) develop routing schemes in which efficiency in the form of response times and quality in the form of service failures are explicitly taken into account. Mehrotra, Ross, and Zhou (2007) explore these issues in an environment with multiple types of agents and multiple classes of customers. Sisselman and Whitt (2005, 2007) provide a link between the call routing problem and employee preferences. Together with the paper by Whitt (2006d,e), these constitute the most direct models of the human resource planning and control interface in call centers and provide promising directions for future research.

Incentive issues arising from the quality–efficiency tension are considered by Shumsky and Pinker (2003) and Gunes and Aksin (2004). Shumsky and Pinker (2003) analyze referral incentives for gatekeepers in multi-tier call centers. Gunes and Aksin (2004) focus on customer service representatives that must perform selling in addition to their basic duty of service provision and design incentive schemes that enable the balance between these two tasks, while ensuring overall call center effectiveness. This incentive dimension provides another potentially fruitful research direction to enhance understanding at the human resource–operations management boundary.

6. The Interface between Operations and Marketing

The operations management literature for call centers has traditionally focused on minimizing customer waiting times and agent staffing costs. The role of the call center in maintaining customer satisfaction and loyalty—which are crucial to most businesses—has historically been overlooked by most researchers. In addition, the one-to-one interaction between call center agents and the customers with whom they interact has the potential to reveal customer needs that a company can meet through other products, an activity known in the call center industry as "cross-selling." In this section, we provide a brief survey of existing papers that examine these types of interactions between call center operations and marketing activities and suggest future directions for research into this increasingly important organizational interface.

6.1. Integrating Cross-Selling Activities into Call Center Operations

Cross-selling activities are now a prevalent practice in call centers. By cross-selling activities we refer to attempts to sell a product or a service to a calling customer that are initiated by the CSR rather than the customer. Cross-selling can be categorized as a marketing activity, but it has some significant operational implications; first, cross-selling attempts necessarily increase call handling times and, in turn, unless staffing levels are appropriately adjusted, waiting times also increase. Second, because agents tend to know more about the caller's buying potential than the system manager, providing the right incentives to agents to make the right decisions with respect to cross-selling is key. Third, there might be various ways by which one can segment the customer population depending on the information available (or acquired), so deciding on the right degree and timing of segmentation matters. Finally, other decisions such as inventory level and degree of customization are also relevant.

Aksin and Harker (1999) study the implications of adding sales functions to a service-oriented banking call center. They highlight the fact that beyond the visible costs of training and technology, adding sales function also adds significantly to system congestion and hence has cost implications in terms of service quality. The authors examine two different scenarios for sales: specialization and non-specialization. In the former, salespeople form their own center, which handles sales calls only. In the latter scenario, some agents are cross-trained to perform both service and sales function. These two scenarios differ with respect to their implications on congestion effects. Finally, the authors conclude that in order to successfully introduce sales into a service-focused center, staffing levels must be adjusted, and the right processes and human resource practices should be adopted.

Gunes and Aksin (2004) consider a situation in which the server can observe the realization of the value-generation potential of customers that are not observable to the manager. The manager, who is interested in maximizing expected profit, is concerned with providing the right incentives to the server so she will attempt to cross-sell to customers who are profitable and not waste time on cross-selling to less profitable customers. The authors identify characteristics of appropriate incentive schemes and demonstrate how they interact with market-segmentation and service-level choices.

Ormeci and Aksin (2004) consider the dynamic cross-selling control problem in a multi-server call center with customers who vary with respect to their revenue generating potential. They propose a static cross-selling heuristic that is based on having preferred customers (those who always generate a cross-selling attempt, regardless of the state of the system) and those that never generate those attempts. They establish sufficient conditions under which the static heuristic performs nearly as well as the optimal dynamic cross-selling control policy. The authors also perform an extensive numerical examination to test under which conditions the call center will benefit the most from dynamic cross-selling decisions.

Byers and So (2007) consider a single-server call center with cross-selling capability. For such a center they study various threshold type policies that determine whether to attempt to cross-sell to a customer based on the queue length and the customer probability of purchasing the product. They demonstrate that it is worthwhile using customer and queue length information, especially in environments with moderate utilization and high customer heterogeneity. They also show that using only queue length information generally outperforms using only customer identity information when the system is highly congested, but the opposite is true when there is high variability in customer profiles. Similar results are demonstrated by the same authors for the multi-server case in the article by Byers and So (2004).

The joint problem of staffing and cross-selling control is studied by Armony and Gurvich (2006) and Gurvich, Armony, and Maglaras (2006). The former explores the single-customer-class case, and the latter focuses on the multi-class case. Armony and Gurvich (2006) determine that in great generality a threshold cross-selling policy in which cross-selling opportunities are exercised whenever the number of customers in the system is below a certain threshold is asymptotically optimal as the system size grows large. They identify two major operating regimes: the cross-selling driven regime and the service driven regime. In the former most customers are being cross-sold to, whereas in the latter only a small fraction of the customers are subject to cross-selling. Interestingly, if staffing levels are appropriately adjusted, the introduction of cross-selling does not necessarily add to customer waiting times, although it extends their service times.

The model of Armony and Gurvich (2006) is generalized by Gurvich, Armony, and Maglaras (2006) in several ways: (a) heterogeneity of customer population, (b) customers' willingness to listen to a cross-selling offer is sensitive to the delay they experience, and (c) the inclusion of product customization (manifested through the asking price) as another control dimension. The paper shows that in this scenario the marketing decisions (such as customer segmentation and pricing) can be decoupled from the operational decisions (expressed in terms of the staffing and cross-selling control) in the sense that sequential decision making (marketing first, operations second) leads to the same results as simultaneous optimization. Finally, whereas market segmentation is always beneficial, using the customer segment information is extremely valuable in making cross-selling and product customization decisions and less valuable for call routing decisions. This last insight may no longer hold if multi-skill agents handle the calls, especially if they differ with respect to their sales capabilities. In this case, routing decisions are much more critical. Modeling such multi-skill systems and understanding how to optimally operate them is an important direction for further exploration.

6.2. Customer Satisfaction and Call Center Operations Management

The relationship between customer satisfaction and profitability and stock price has been well established empirically through the American Customer Satisfaction Index Fornell et al. (2006). In addition, the impact of the customer service experience on customer satisfaction and retention has been studied by several researchers, including Johnston (1998), Goodman and Newman (2002), and Chebat, Davidow, and Codjovi (2005). One of the key findings from this stream of research is that the vast majority of dissatisfied customers do not complain but are nevertheless at much greater risk of abandoning their relationship with the company as a consequence of their unhappiness. The importance of the call center in this relationship is underscored by a recent study that asserted that 80% of a firm's interaction with its customers is through call centers, and 92% of customers form their opinion about a firm based on their experience with call centers (Anton, Setting, and Gunderson 2004).

Gans (2002) explores this phenomenon in the context of repeated customer interactions with a group of competing suppliers, modeling customer choice as a function of the quality of previous interactions with a given supplier. Given this Bayesian updating of customer preferences, the quality of customers' experience with a particular firm will have a major impact on that firm's long-term market share.

Recognizing the importance of service quality on customer satisfaction, de Vericourt and Zhou (2005) model a call center in which calls that are not handled successfully cause the customer to call back. This paper examines heterogeneous agents, each of which has potentially different call handling times and call resolution rates, and develops a strategy for routing the

two different classes of calls across different agent groups.

Over the past several years, there has been extensive investment by call centers in customer relationship management (CRM) systems that capture and store information about customers and their interaction with the company. A great challenge for managers, both in the call center and in other parts of the organization, is to determine how to leverage the contents of these systems to reduce costs and improve the company's relationship with its customers. Mehrotra and Grossman (2006) describe process improvement methods for a consumer software company's technical support call center. Utilizing CRM data captured during customer phone calls, analysts were able to quantify the impact of specific issues on call volumes and work with the product marketing, engineering, and documentation groups to eliminate specific problems from future software releases. The result of these processes was a lower per-customer call arrival rate, as well as increased customer satisfaction. Sun and Li (2006) use CRM data about service durations and customer retention in conjunction with an adaptive customer learning model (Sun, Li, and Zhou 2006) to suggest policies for distributing calls from different types of customers across heterogeneous on- and offshore call centers within the same network, while considering both short-term and long-term customer economic implications for the firm.

We see several opportunities for research in this area. First, because of the relationship between successful call resolution and customer satisfaction—and because of the potentially significant impact of retrials by customers whose calls are not handled successfully—there is a need for more performance models that include first call resolution rates and customer callbacks. Second, as CRM systems capture and store increasing volumes of call history, this data provides an opportunity for segmenting customers into distinct groups based on value and preferences; similarly, for routing purposes, agents can be segmented into groups based on their performance characteristics. Finally, in addition to direct costs such as agent wages, many existing staffing and scheduling models can be extended to include longer-term financial effects from customer retention and loss, with these parameters becoming easier to estimate over time as a result of better, more accessible databases.

Concluding Remarks

As the global call center industry continues to grow, the range of operations management challenges that call centers face has become broader and more complex. The call center industry's growth has been driven by many factors, including evolving management practices, decreased telecommunications costs, and increasingly powerful information technology. In addition, several other factors have also contributed to increased operational breadth and complexity, including firms' awareness of call centers as a powerful customer channel, not only for service delivery but also for customer satisfaction, sales opportunities, and relationship management. Increased outsourcing and globalization of service delivery have also played a major part in both the industry's growth and the increase in operational complexity.

In this paper, we have surveyed recent call center research and examined many of the challenges presented by changes in the industry. Our focus has been mainly on operations management of inbound call centers, but we have sought to highlight research in other disciplines and the interfaces between these areas and operations management research. Although there has been progress in many different directions, we also see significant opportunities and needs for additional research, and throughout the paper we have outlined research directions that have the potential to significantly improve the way in which call centers are managed.

To conclude this survey, we present some "macro" research themes that we believe are important for future call center operations research.

First, researchers can benefit from improving the way in which the tension between efficiency and quality of service is modeled. Historically, most research on call center operations has equated service quality with customer waiting times. However, there are numerous studies that demonstrate that customers place a high value on other dimensions of their experience, including factors such as first call resolution and perceived agent competency, as well as less tangible measures such as politeness and friendliness. As such, there is a need for effectively modeling service quality in a manner more consistent with these customer values. Also, given that efficiency and speed often conflict with other broader measures of service quality, there are inherent challenges in measuring agent performance and establishing compensation structures that are more likely to produce the desired efficiency and quality outcomes, while reducing the tension felt by agents. Similarly, firms would benefit from a better understanding of the relationship between customers' service experiences and their repeat purchase behavior, loyalty to the firm, and overall demand growth in order to make better decisions about call center operations.

Second, we believe that there is still significant work to be done on traditional call center operations management problems, including both theoretical and empirical research. Forecasting models will continue to play an important role in operations, serving as a

critical input for both resource acquisition and resource deployment decisions. In addition, there is an opportunity for increased integration of forecasting, hiring, staffing, scheduling, and routing decisions, ultimately leading to better resource utilization and lower customer waiting times. Also, as multi-queue and multi-site call center operations become more common, the queueing models used for staffing and performance analysis play an increasingly important role. In this context, we see a need for understanding the robustness of more advanced models while also exploring which modeling assumptions are essential for what types of analyses (and which assumptions can be safely relaxed for particular types of operations).

Agent skill set design is another potentially fruitful area for investigation. Although the agents' skill sets can have a big impact on staffing and scheduling decisions, the design of these skill sets has historically been treated as inputs rather than variables that can be controlled for operational advantage. In addition, the migration of agents through different skill configurations has a significant effect on agent learning, career paths, job satisfaction, and attrition rates, all of which have an affect on operational performance.

The increased use of call center outsourcing firms and work-at-home agents has created another rich area that researchers have just begun to explore. The structure of outsourcing contracts plays a major role in determining how long arriving calls must wait and, more generally, how well customers are ultimately served. The ongoing development of new performance metrics on which the effectiveness of different contracts can be evaluated is a useful direction to explore. In support of this, more analysis is needed about the different types of contracts and performance metrics used in practice. We also see great value in additional insights about the implications of various contract structures for outsourcers who must deliver service for multiple firms and similarly for firms who contract for services from multiple outsourcers. Understanding the impact of outsourcing contracts (and the availability of work-at-home agents) on the scheduling and management of in-house personnel is another fruitful direction.

Finally, there is a need to more closely examine the behavioral issues that influence call center operations. In particular, there is much to learn about call center customer behavior, including an understanding of their patience and abandonment behavior, their reaction to different types of waiting time information, their response to promotions and other marketing efforts through the call center, their quality expectations and perceptions in terms of both quantitative and qualitative call center metrics, and the connection between their service experience and their long-term purchase practices and loyalty. A better understanding of agent behavior is also important, including an understanding of how different staffing, scheduling, and routing practices impact key outcomes such as agent turnover, absenteeism, and service quality. Although some of these behavioral issues can be examined through lab experiments and simulations, we believe that in many cases empirical analyses based on historical data can be extremely valuable in providing insight into these questions.

Acknowledgments

We thank George Shanthikumar and David Yao for the invitation to write this paper. We are also grateful to Tommy Mermelshtayn for extensive literature search and many helpful comments and to Leslie Culpepper for invaluable assistance with the bibliography. Finally, we greatly appreciate suggestions made by Avi Mandelbaum in various stages of the manuscript's evolution.

Uncited References

This section comprises references that occur in the reference list but not in the body of the text. Please position each reference in the text or, alternatively, delete it. Any reference not dealt with will be retained in this section: Canon, C., J. Billaut, J. Bouqard (2005), Gonzalez-Simental, M. T., E. Pines (2006), Henderson S., A. J. Mason (1998), Maister, D. H. (1985), Mandelbaum, A., W. A. Massey, M. I. Reiman, A. Stolyar, B. Rider 2002, Mandelbaum, A., S. Zeltyn (2005), Sun, B., S. Li (2006).

Uncited References

This section comprises references that occur in the body of the text but not in the reference section. Please position each reference in the reference section or delete it. Any references not dealt with will be retained in this section: Armony et al.(2005); Avramidis et al. (2006), Maister (1984), Grandey et al. (2003).

References

Aguir, M. S. 2004. *Modeles stochastique pour l'aide a la decision dans les centres d'appels*. PhD Thesis, Ecole Centrale Paris.

Aguir, M. S., O. Z. Aksin, F. Karaesmen, Y. Dallery. 2007. On the interaction between retrials and sizing of call centers. *European Journal of Operational Research*, forthcoming.

Aguir, M. S., F. Karaesmen, O. Z. Aksin, F. Chauvet. 2004. The impact of retrials on call center performance. *OR Spectrum* **26**(3) 353–376.

Ahn, H.-S., R. Righter, J. G. Shanthikumar. 2005. Staffing decisions for heterogeneous workers with turnover. *Mathematical Methods of Operations Research* **62**(3) 499–514.

Aksin, O. Z., F. de Vericourt, F. Karaesmen. 2006. Call center outsourcing contract analysis and choice. *Management Science*, forthcoming.

Aksin O. Z., P. T. Harker. 1999. To sell or not to sell: Determining the tradeoffs between service and sales in retail banking phone centers. *Journal of Service Research* **2**(1) 19–33.

Aksin, O. Z., P. T. Harker. 2003. Capacity sizing in the presence of a common shared resource: Dimensioning an inbound call center. *European Journal of Operational Research* **147**(3) 464–483.

Aksin, O. Z., F. Karaesmen. 2003. Designing flexibility: Characterizing the value of cross-training practices. Working Paper, Koc University.

Aksin, O. Z., F. Karaesmen. 2007. Characterizing the performance of

process flexibility structures. *Operations Research Letters* **35**(4) 477–484.

Aksin O. Z., F. Karaesmen, E. L. Ormeci. 2005. On the interaction between resource flexibility and flexibility structures in *Proceedings of the Fifth International Conference on Analysis of Manufacturing Systems—Production Management*, Zakynthos, Greece, May 2005.

Aksin, O. Z., F. Karaesmen, E. L. Ormeci. 2007. A review of workforce cross-training in call centers from an operations management perspective in *Workforce Cross Training Handbook*, D. Nembhard (ed.), CRC Press, Boca Raton, FL.

Aksin, O. Z., A. Masini. 2006. Effective strategies for internal outsourcing and offshoring of business services: An empirical investigation. *Journal of Operations Management*, forthcoming. doi:10.1016/j.jom.2007.02.003.

Aldor-Noiman, S. 2006. *Forecasting demand for a telephone call center: Analysis of desired versus attainable precision*. M. Sc. Thesis, Technion, Haifa, Israel.

Andrews, B. H., S. M. Cunningham. 1995. L.L. Bean improves call-centre forecasting. *Interfaces* **25**(6) 1–13.

Anton, J., T. Setting, C. Gunderson. 2004. Offshore company call centers: A concern to U.S. consumers. Technical Report, Purdue University Center for Customer-Driven Quality.

Apte, U. M., R. O. Mason. 1995. Global disaggregation of information-intensive services. *Management Science* **41**(7) 1250–1262.

Armony, M. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* **51**(3–4) 287–329.

Armony, M., I. Gurvich. 2006. When promotions meet operations: Cross-selling and its effect on call-center performance. Submitted for publication.

Armony, M., C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527–545.

Armony, M., C. Maglaras. 2004b. On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Operations Research* **52**(2) 271–292.

Armony, M., A. Mandelbaum. 2004. Design, staffing, and control of large service systems: The case of a single customer class and multiple server types. Working paper. Technion—Israel Institute of Technology.

Armony, M., E. L. Plambeck, S. Seshadri. 2007. Sensitivity of optimal capacity to customer impatience in an unobservable M/M/S queue (Why you shouldn't shout at the DMV). *Manufacturing and Service Operations Management*, forthcoming.

Armony, M., N. Shimkin, W. Whitt. 2006. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, forthcoming.

Atar, R. 2005a. A diffusion model of scheduling control in queuing systems with many servers. *Annals of Applied Probability* **15**(1b) 820–852.

Atar, R. 2005b. Scheduling control for queuing systems with many servers: Asymptotic optimality in heavy traffic. *Annals of Applied Probability* **15**(4) 2606–2650.

Atar, R., A. Mandelbaum, M. Reiman. 2004a. A Brownian control problem for a simple queuing system in the Halfin–Whitt regime. *Systems Control Letters* **51**(3–4) 269–275.

Atar, R., A. Mandelbaum, M. Reiman. 2004b. Scheduling a multiclass queue with many exponential servers: Asymptotic optimality in heavy-traffic. *Annals of Applied Probability* **51**(3) 1084–1134.

Atlason, J., M. A. Epelman, S. G. Henderson. 2003. Using simulation to approximate subgradients of convex performance measures in service systems in *Proceedings of the 2003 Winter Simulation Conference*, New Orleans, LA. S. Chick, P. J. Sánchez, D. Ferrin, D. J. Morrice, (eds.), 1824–1832.

Atlason, J., M. A. Epelman, S. G. Henderson. 2004. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research* **127**(1–4) 333–358.

Avramidis, A. N.,W. Chan, P. L'Ecuyer. 2007. Staffing multiskill call centers via search methods and a performance approximation. *IIE Transactions*, forthcoming.

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

Aykin, T. 1996. Optimal shift scheduling with multiple break windows. *Management Science* **42**(4) 591–602.

Aykin, T. 2000. A comparative evaluation of modeling approaches to the labor shift scheduling problem. *European Journal of Operational Research* **125**(2) 381–397.

Bain, P., A. Watson, G. Mulvey, P. Taylor, G. Gall. 2002. Taylorism, targets and the pursuit of quantity and quality by call center management. *New Technology, Work and Employment* **17**(3) 170–185.

Bakker, A. B., E. Demerouti, W. B. Schaufeli. 2003. Dual processes at work in a call center: An application of the job-demands-resources model. *European Journal of Work and Organizational Psychology* **12**(4) 393–428.

Baron, O., J. Milner. 2006. Staffing to maximize profit for call centers with alternate service level agreements. Working Paper, University of Toronto.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2005. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems: Theory and Applications* **51**(3–4) 249–285.

Bassamboo, A., J. M. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Operations Research* **54**(3) 419–435.

Bassamboo, A., A. Zeevi. 2007. On a data-driven method for staffing large call centers. *Operations Research*, forthcoming.

Batt, R. 2002. Managing customer services: Human resource practices, quit rates and sales growth. *Academy of Management Journal* **45**(3) 587–599.

Batt, R., L. Moynihan. 2002. The viability of alternative call center production models. *Human Resource Management Journal* **12**(4) 14–34.

Bernett, H. G., M. J. Fischer, D. M. B. Masi. 2002. Blended call center performance analysis. *IT Professional* **4**(2) 33–38.

Bhandari, A., M. Harchol-Balter, A. Scheller-Wolf. 2007. An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers. *Management Science*, forthcoming.

Bhulai, S. 2005. Dynamic routing policies for multi-skill call centers. Technical Report WS2004–11, Vrije Universiteit Amsterdam.

Bhulai, S., G. Koole. 2003. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control* **48**(8) 1434–1438.

Bhulai, S., G. Koole, A. Pot. 2007. Simple methods for shift scheduling in multi-skill call centers. *Manufacturing & Service Operations Management*, forthcoming.

Bordoloi, S. K. 2004. Agent recruitment planning in knowledge-intensive call centers. *Journal of Service Research* **6**(4) 309–323.

Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* **100**(469) 36–50.

Buzacott, J. A. 1996. Commonalities in reengineered business processes: Models and issues. *Management Science* **42**(5) 768–782.

Byers, R. E., K. C. So. 2004. The value of information-based cross-sales policies in telephone service centers. Working Paper, Graduate School of Management, University of California, Irvine.

Byers, R. E., K. C. So. 2007. A mathematical model for evaluating cross-sales policies in telephone service centers. *Manufacturing and Service Operations Management* **9**(1) 1–8.

Canon, C., J. Billaut, J. Bouqard. 2005. Dimensioning an inbound call center using constraint programming in *Principles and practice of constraint programming*, P. van Beek (ed.), Springer, Berlin.

Carmon, Z., D. Kahneman. 2002. The experienced utility of queueing: Experience profiles and retrospective evaluations of simulated queues. Working Paper, INSEAD.

Cezik, T., P. L'Ecuyer. 2006. Staffing multiskill call centers via linear programming and simulation. *Management Science*, forthcoming.

Chebat, J. C., M. Davidow, I. Codjovi. 2005. Silent voices: Why some dissatisfied consumers fail to complain. Journal of Service Research **7**(4) 328–342.

Chen, B., S. G. Henderson. 2001. Two issues in setting call center staffing levels. *Annals of Operations Research* **108**(1–4) 175–192.

Chevalier P., R. A. Shumsky, N. Tabordon. 2004. Routing and staffing in large call centers with specialized and fully flexible servers. Working Paper, Universite Catholique de Louvain.

Chevalier, P., N. Tabordon. 2003. Overflow analysis and cross trained servers. *International Journal of Production Economics* **85**(1) 47–60.

Chevalier, P., J-C. Van den Schrieck. 2006. Optimizing the staffing and routing of small size hierarchical call-centers. Working paper, Universite Catholique de Louvain.

Cipra, T. 1992. Robust exponential smoothing. *Journal of Forecasting* **11**(1) 57–69.

Cordes, C. L., T. W. Dougherty. 1993. A review and an integration of research on job burnout. *Academy of Management Review* **18**(4) 621–656.

de Vericourt, F., Y-P. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968–981.

de Vericourt, F., Y.-P. Zhou. 2006. On the incomplete results for the heterogeneous server problem. *Queueing Systems* **52**(3) 189–191.

Deery, S., R. Iverson, J. Walsh. 2002. Work relationships in telephone call centres: Understanding emotional exhaustion and employee withdrawal. *Journal of Management Studies* **39**(4) 471–196.

Deery, S., N. Kinnie. 2002. Call centres and beyond: A thematic evaluation. *Human Resource Management Journal* **12**(4) 3–13.

Dellaportes, P., A. F. M. Smith. 1993. Bayesian inference for generalized linear and proportional hazards models via gibbs sampling. *Applied Statistics* **42**(3) 443–459.

Deslauriers, A., P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, A. N. Avramidis. 2007. Markov chain models of a telephone call center with call blending. *Computers and Operations Research* **34**(6) 1616–1645.

Easton, F. F., J. C. Goodale. 2002. Labor scheduling with employee turnover and absenteeism. Working Paper, Syracuse University.

Easton, F. F., J. C. Goodale. 2005. Schedule recovery: Unplanned absences in service operations. *Decision Sciences* **36**(3) 459–488.

Eichfeld, A., T. D. Morse, K. W. Scott. May 2006. Using call centers to boost revenue. *The McKinsey Quarterly.*

Erlang, A. K. 1948. On the rational determination of the number of circuits in *The life and works of A. K. Erlang*. E. Brockmeyer, H. L. Halstrom, A. Jensen (eds.), Copenhagen: The Copenhagen Telephone Company.

Ernst, A. T., H. Jiang, M. Krishnamoorthy, B. Owens, D. Sier. 2004. An annotated bibliography of personnel scheduling and rostering. *Annals of Operations Research* **127**(1–4) 21–144.

Feigin, P. 2006. Analysis of customer patience in a bank call center. Working Paper in preparation, Technion Israel Institute of Technology.

Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2005. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, forthcoming.

Fornell, C., S. Mithas, F. Morgeson, M. S. Krishnan. 2006. Customer satisfaction and stock prices: High returns, low risk. *Journal of Marketing* **70**(1) 3–14.

Franx G. J., G. M. Koole, S. A. Pot. 2006. Approximating multi-skill blocking systems by hyperexponential decomposition. *Performance Evaluation* **63**(8) 799–824.

Fukunaga, A., E. Hamilton, J. Fama, D. Andre, O. Matan, I. Nourbakhsh. 2002. Staff scheduling for inbound call centers and customer contact centers in *Eighteenth National Conference on Artificial Intelligence* (Moncton, Alberta, Canada, July 28–August 1, 2002), R. Dechter, M. Kearns, R. Sutton, (eds.), American Association for Artificial Intelligence, Menlo Park, California, 822–829.

Gamarnik D., P. Momcilovic. 2007. Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. Working paper, University of Michigan.

Gans, A. 2002. Customer loyalty and supplier quality competition. *Management Science* **48**(2) 207–221.

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

Gans, N., Y-P. Zhou. 2002. Managing learning and turnover in employee staffing. *Operations Research* **50**(6) 991–1006.

Gans, N., Y-P. Zhou. 2003. A call-routing problem with service-level constraints. *Operations Research* **51**(2) 255–271.

Gans, N., Y-P. Zhou. 2007. Call-routing schemes for call-center outsourcing. *Manufacturing and Service Operations Management* **9**(1) 33–50.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Operations Management* **4**(3) 208–227.

Goodman, J. A., S. Newman. 2003. Understanding customer behavior and complaints. *Quality Progress* **36**(1) 51–55.

Grandey, A. A., D. N. Dickter, H. Sin. 2004. The customer is not always right: Customer aggression and emotion regulation of service employees. *Journal of Organizational Behavior* **25**(3) 397–418.

Green, L. V., P. J. Kolesar, J. Soares. 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* **49**(4) 549–564.

Green, L. V., P. J. Kolesar, J. Soares. 2003. An improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management* **12**(1) 46–61.

Green L.V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16**(1) 13–39.

Gunes, E. D., O. Z. Aksin. 2004. Value creation in service delivery: Relating market segmentation, incentives, and operational performance. *Manufacturing and Service Operations Management* **6**(4) 338–357.

Guo, P., P. Zipkin. 2006. The effects of information on a queue with balking and phase-type service times. Working paper, Duke University.

Guo, P., P. Zipkin. 2007a. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970.

Guo, P., P. Zipkin. 2007b. Information and congestion in a service system with balking. Working paper, Duke University.

Gurvich, I., M. Armony, C. Maglaras. 2006. Cross-selling in a call center with a heterogeneous customer population. *Operations Research*, forthcoming.

Gurvich I., M. Armony, A. Mandelbaum. 2006. Service level differ-

entiation in call centers with fully flexible servers. *Management Science,* forthcoming.

Gurvich, I., W. Whitt. 2007. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. Working Paper, Columbia University.

Halfin, S, W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3) 567–588.

Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime. *Operations Research* **52**(2) 243–257.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management* **7**(1) 20–36.

Hasija, S., E. J. Pinker, R. A. Shumsky. 2005. Staffing and routing in a two-tier call center. *International Journal Operational Research* **1**(1–2) 8–29.

Hasija, S., E. J. Pinker, R. A. Shumsky. 2007. Call center outsourcing contracts under information asymmetry. *Management Science*, forthcoming.

Hassin, R., M. Haviv. 1995. Equilibrium strategies for queues with impatient customers. *Operations Research Letters* **17**(1) 41–5.

Helber, S., R. Stolletz, S. Bothe. 2005. Erfolgszielorientierte Agentenallokation in inbound call Centern. *Zeitschrift fur Betriebswirtschaftliche Forschung* (February), 3–32.

Holman, D. 2005. Call centers in *The essential of the new workplace: A guide to the human impact of modern work practices*. D. Holman, T. D. Wall, C. W. Clegg, P. Sparrow, A. Howard (eds.), Wiley, New York.

Holman, D., C. Chissick, P. Totterdell. 2002. The effects of performance monitoring on emotional labor and well being in call centers. *Motivation and Emotion* **26**(1) 57–81.

Houlihan, M. 2002. Tensions and variations in call centre management strategies. *Human Resource Management Journal* **12**(4) 67–85.

Hu, B., S. Benjaafar, 2006. On server partitioning in queueing systems during rush hour. Working Paper, University of Minnesota.

Hui, M. K., D. K. Tse. 1996. What to tell consumers in waits of different lengths: An integrative model of service evaluation. Journal of Marketing **60**(2) 81–90.

Hui, M. K., L. Zhou. 1996. How does waiting duration information influence customer's reactions to waiting for services? *Journal of Applied Social Psychology* **26**(19) 1702–1717.

Hur, D., V. A. Mabert, K. M. Bretthauer. 2004. Real-time work schedule adjustment decisions: An investigation and evaluation. *Production and Operations Management*, **13**(4) 322–339.

Ingolfsson, A., E. Cabral. X. Wu. 2003. Combining integer programming and the randomization method to schedule employees. Technical Report, University of Alberta.

Jagerman, D. L., B. Melamed. 2004. Models and approximations for call center design. *Methodology and Computing in Applied Probability* **5**(2) 159–181.

Jelenkovic P., A. Mandelbaum, P. Momcilovic. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems: Theory and Applications* **47**(1–2) 53–69.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Science* **42**(10) 1383–1394.

Jimenez, T., G. Koole. 2004. Scaling and comparison of fluid limits of queues applied to call centers with time carrying parameters. *OR Spectrum* **26**(3) 413–422.

Johnston, R. 1998. The effect of intensity of dissatisfaction on complaining behaviour *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior* **11** 69–77.

Jordan, W. C., S. C. Graves. 1995. Principles on the benefits of manufacturing process flexibility. *Management Science* **41**(4) 577–594.

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

Jouini, O., Y. Dallery. 2006. Estimating and announcing waiting times in multiple customer class call centers *Proceedings of INCOM* **2** 371–376.

Jouini, O., Y. Dallery, O. Z. Aksin. 2007a. Queueing models for multiclass call centers with real-time anticipated delays. Working Paper, Koc University.

Jouini, O., Y. Dallery, O. Z. Aksin. 2007b. Modeling call centers with delay information. Working Paper, Koc University.

Jouini, O., A. Pot, Y. Dallery, G. Koole. 2006. Real-time dynamic scheduling policies for multiclass call centers with impatient customers. Working paper, Ecole Centrale Paris.

Jouini, O., Y. Dallery, R. Nait-Abdallah. 2006. Analysis of the impact of team-based organizations in call center management. *Management Science*, forthcoming.

Kaspi, H., K. Ramanan. 2007. Law of large numbers limits for many-server queues. Working Paper, Technion—Israel Institute of Technology.

Keblis, M., M. Chen. 2006. Improving customer service operations at amazon.com. *Interfaces* **36**(5) 433–445.

Keblis, M., Y. Li, W. E. Stein. 2007. Real-time staffing of virtual call centers. Working Paper, Texas A&M University.

Kolesar, P., L. Green. 1998. Insights on service system design from a normal approximation to Erlang's delay formula. *Production and Operations Management* **7**(3) 282–293.

Koole, G. 2003. Redefining the service level in call centers. Technical Report, Department of Stochastics, Vrije Universiteit Amsterdam.

Koole, G., A. Mandelbaum. 2002. Queueing models of call centers: An introduction. *Annals of Operations Research* **113**(1–4) 41–59.

Koole, G., A. Pot. 2005a. A note on profit maximization and monotonicity for inbound call centers. Technical report, Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands.

Koole G., A. Pot. 2005b. Approximate dynamic programming in multi-skill call centers in *Proceedings of the 37th Conference on Winter Simulation*, 576–583.

Koole, G., A. Pot. 2006. An overview of routing and staffing algorithms in multi-skill customer contact centers. Technical Report, Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands.

Koole, G., A. Pot, J. Talim. 2003. Routing heuristics for multi-skill call centers in *Proceedings of the 35th Conference on Winter Simulation*, 1813–1816.

Koole, G., J. Talim. 2000. Exponential approximation of multi skill call centers architecture. *Proceedings of QNETs*, **23** 1–10.

Koole, G., E. van der Sluis. 2003. Optimal shift scheduling with a global service level constraint. *IIE Transactions* **35** 1049–1055.

L'Ecuyer, P. 2006. Modeling and optimization problems in contact centers in *Proceedings of the Third International Conference on the Quantitative Evaluation of Systems* (QEST 2006), University of California, Riversdale, IEEE Computing Society, 145–154.

Maglaras C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* **49**(8) 1018–1038.

Maister, D. H. 1985. The psychology of waiting lines in *The service encounter: managing employee/customer interaction in service businesses*, J. A. Czepiel, M. R. Solomon, C. F. Suprenant. (eds.), D. C. Heath and Company, Lexington Books, Lexington, Massachusetts.

Mandelbaum, A., W. A. Massey, M. I. Reiman, B. Rider. 1999. Time varying multiserver queues with abandonments and retrials in

Proceedings of the 16th International Teletraffic Conference, 355–364.

Mandelbaum, A. Momcilovic P. 2007. Queues with many servers: The virtual waiting-time process in the QED regime. Working paper, Technion Israel Institute of Technology.

Mandelbaum A., M. I. Reiman. 1998. On pooling in queueing networks. *Management Science* **44**(7) 971–981.

Mandelbaum A., A. Sakov, S. Zeltyn. 2000. Empirical analysis of a call center. Technical Report, Technion Israel Institute of Technology.

Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems: Theory and Applications* **36**(1–3) 141–173.

Mandelbaum, A., A. L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized c\mu-rule. *Operations Research* **52**(6) 836–855.

Mandelbaum, A., S. Zeltyn. 2004. The impact of customers' patience on delay and abandonment: Some empirically-driven experiments with the M/M/n + G queue. *OR Spectrum* **26**(3) 377–411.

Mandelbaum, A., S. Zeltyn. 2005. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Systems* **51**(3–4) 361–402.

Mandelbaum, A., S. Zeltyn. 2006. Service-engineering of call centers: Research, teaching, practice. IBM Business Optimization and Operations Research Workshop, Haifa, Israel.

Mandelbaum, A., S. Zeltyn. 2007a. Service engineering in action the palm/erlang A queue, with application to call centers in *Advances in services innovations*, D. Spath, K.-P. Fahnrich (eds.), Springer, Berlin.

Mandelbaum, A., S. Zeltyn. 2007b. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. Working Paper, Technion—Israel Institute of Technology.

Maslach, C., S. E. Jackson. 1981. The measurement of experienced burnout. *Journal of Occupational Behavior* **2**(2) 99–113.

Massey, W. A. 2002. The Analysis of queues with time-varying rates for telecommunication models. *Telecommunications Systems* **21**(2–4) 173–204.

Massey, W. A., R. B. Wallace. 2006. An optimal design of the M/M/C/K queue for call centers. *Queueing Systems*, forthcoming.

Mazzuchi, T.A., R. B. Wallace. 2004. Analyzing skill-based routing call centers using discrete-event simulation and design experiment in *Proceedings of the 36th Conference on Winter Simulation*, Washington, DC, R. G. Ingalls, M. D. Rossetti, J. S. Smith, B. A. Peters (eds.), 1812–1820.

Mehrotra, V., J. Fama. 2003. Call center simulation modeling: Methods, challenges, and opportunities in *Proceedings of the 35th conference on Winter Simulation*, New Orleans, LA, S. Chick, P. J. S'anchez, D. Ferrin, D. J. Morrice (eds.), **1**, 135–143.

Mehrotra, V., T. Grossman. 2006. New processes enhance cross-functional collaboration and reduce call center costs. Working Paper, Department of Decision Sciences, San Francisco State University.

Mehrotra, V., O. Ozluk, R. Saltzman. 2006. Intelligent procedures for intra-day updating of call center agent schedules. Working Paper, Department of Decision Sciences, San Francisco State University.

Mehrotra, V., A. Ross, Y.-P. Zhou. 2007. Call center routing strategies in the presence of servers with heterogeneous performance attributes. Working Paper, University of California—Santa Cruz.

Milner, J. M., T. L. Olsen. 2006. Service level agreements in call centers: Perils and prescriptions. *Management Science*, forthcoming.

Moynihan, L., R. Batt. 2001. Knowledge sharing and performance of teams in call centers. Working Paper, Cornell University.

Munichor, N., A. Rafaeli. 2006. Numbers or apologies? Customer reactions to tele-waiting time fillers. *Journal of Applied Psychology* **92**(2) 511–518.

Nakibly, E. 2002. *Predicting queueing delays for multiclass call centers*. Ph.D. thesis, Technion Israel Institute of Technology.

Ormeci, E. L. 2004. Dynamic admission control in a call center with one shared and two dedicated service facilities, *IEEE Transactions on Automatic Control* **49**(7) 1157–1161.

Ormeci, E. L., O. Z. Aksin. 2004. Revenue management through dynamic cross-selling in call centers. Technical Report, Koc University.

Ormeci, E. L., A. N. Burnetas, H. Emmons. 2002. Admission policies for a two class loss system with random rewards, *IIE Transactions* **34**(9) 813–822.

Pichitlamken, J., A. Deslauriers, P. L'Ecuyer, A. N. Avramidis. 2003. Modeling and simulation of a telephone call center in *Proceedings of the 37th Conference on Winter Simulation*, New Orleans, LA, **2**, 1805–1812.

Pinker, E. J., R. A. Shumsky. 2000. The efficiency-quality trade-off of cross-trained workers. *Manufacturing and Service Operations Management* **2**(1) 32–48.

Pot A., S. Bhulai, G. Koole. 2007. A simple staffing method for multi-skill call centers. *Manufacturing and Service Operations Management*, forthcoming.

Reed, J. E. 2005. The G/GI/N queue in the Halfin-Whitt regime. Working Paper.

Ren, Z. J., Y-P. Zhou. 2006. Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, forthcoming.

Ridley, A. D., M. C. Fu, W. A. Massey. 2003. Customer relations management: call center operations: Fluid approximations for a priority call center with time-varying arrivals in *Proceedings of the 35th Conference on Winter Simulation,* New Orleans, LA, **2,** 1817–1823.

Robbins, T. R., Harrison, T. P. 2007. A stochastic programming model for scheduling call centers with global service level agreements. Working Paper.

Robbins, T. R., Harrison, T. P., Medeiros, D. J. 2007. Partial cross training in call centers with uncertain arrivals and global service level agreements in *Proceedings of the 2007 Winter Simulation Conference*, Washington, DC. 2252–2258.

Robbins, T. R., Medeiros, D. J., Dum, J. 2006. Evaluating arrival rate uncertainty in call centers in *Proceedings of the 2006 Winter Simulation Conference*, Monterey, CA. 2180–2187.

Ross, A. M. 2001. Queueing systems with daily cycles and stochastic demand with uncertain parameters. PhD thesis, University of California, Berkeley, Berkeley, California.

Ryder, G. S., K. G. Ross, J. T. Musacchio. 2008. Optimal service policies under learning effects, *International Journal of Services and Operations Management* **4**(6) forthcoming.

Saltzman, R. 2005. A hybrid approach to minimizing the cost of staffing a call center. *International Journal of Operations and Quantitative Management* **11**(1) 1–14.

Saltzman, R. M., V. Mehrotra. 2007. Managing trade-offs in call center agent scheduling: Methodology and case study in *Proceedings of the 2007 Summer Computer Simulation Conference*, San Diego, CA. G. A. Wainer and H. Vakilzadian (eds.), 643–651.

Shen, H., J. Z. Huang. 2007. Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management*, forthcoming.

Shimkin, N., A. Mandelbaum. 2004. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems: Theory and Applications* **47**(1–2) 117–146.

Shumsky, R. A. 2004. Approximation and analysis of a queueing

system with flexible and specialized servers. *OR Spectrum* **26**(3) 307–330.

Shumsky, R. A., E. J. Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839–856.

Singh, J., J. R. Goolsby, G. K. Rhoads. 1994. Behavioral and psychological consequences of boundary spanning burnout for customer service representatives. *Journal of Marketing Research* **31**(4) 558–569.

Sisselman, M. E., W. Whitt. 2005. Empowering customer-contact agents via preference based routing. Seatlink White Paper. Accessed at http://www.seatlink.net/whitepapers.asp.

Sisselman, M. E., W. Whitt. 2007. Value-based routing and preference-based routing in customer contact centers. *Production and Operations Management* **16**(3) 277–291.

Soyer, R., M. Tarimcilar. 2007. Modeling and analysis of call center arrival data: A bayesian approach. *Management Science*, forthcoming.

Steckley, S. G., S. G. Henderson, V. Mehrotra. 2005. Performance measures for service systems with a random arrival rat in *Proceedings of the 37th Conference on Winter Simulation*. Orlando, FL. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, J. A. Joines (eds.), 566–575.

Steckley, S. G., S. G. Henderson, V. Mehrotra. 2007. Forecast errors in service systems. *Probability in the Engineering and Information Sciences*, forthcoming.

Stolletz, R., S. Helber. 2004. Performance analysis of an inbound call center with skill-based routing: A priority queueing system with two classes of impatient customer and heterogeneous agents. *OR Spectrum* **26**(3) 331–352.

Sun, B., S. Li. 2006. Improving effectiveness of customer service in a cost-efficient way with an empirical application to the call allocation decisions with out-sourced centers. Working Paper, Tepper School of Business, Carnegie Mellon University.

Sun, B., S. Li, C. Zhou. 2006. Adaptive' learning and 'proactive' customer relationship management. *Journal of Interactive Marketing* **20**(3–4) 82–96.

Taylor, J. W. 2003. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of Operational Research Society* **54**(8) 799–805.

Taylor, J. W. 2007. A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, forthcoming.

Tekin, E., W. J. Hopp, M. P. van Oyen. 2004. Pooling strategies for call center agent cross-training. Working Paper, University of North Carolina, Chapel Hill.

Tezcan, T. 2005. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Mathematics of Operations Research*, forthcoming.

Tezcan, T., J. Dai. 2006. Dynamic control of N-Systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. Working Paper, Georgia Institute of Technology.

Thompson, G. M. 1999. Setting staffing levels in pure service environments when the true mean daily customer arrival rate is a normal random variate. Unpublished manuscript.

Torzhkov, A., M. Armony. 2007. Staffing of service systems with arrival rate forecast evolution. Working paper.

Tuten, T. L., P. E. Neidermeyer. 2004. Performance, satisfaction, and turnover in call centers: The effects of stress and optimism. *Journal of Business Research* **57**(1) 26–34.

Tych, W., D. J. Pedregal, P. C. Young, J. Davies. 2002. An unobserved component model for multi-rate forecasting of telephone call demand: The design of a forecasting support system. *International Journal of Forecasting* **18**(4) 673–695.

von Emster, G. R., A. A. Harrison. 1998. Role ambiguity, spheres of control, burnout, and work-related attitudes of teleservice professionals. *Journal of Social Behavior and Personality* **13** 375–385.

Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management* **7**(4) 276–294.

Weinberg. J., L. D. Brown, J. R. Stroud. 2007. Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association*, **102,** 1185–1199.

Whitt, W. 1999a. Improving service by informing customers about anticipated delays. *Management Science* **45**(2) 192–207.

Whitt, W. 1999b. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* **24**(5) 205–212.

Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. *Operations Research* **51**(4) 531–542.

Whitt, W. 2004a. A diffusion approximation for the G/Gl/n/m queue. *Operations Research* **52**(6) 922–941.

Whitt, W. 2004b. Efficiency driven heavy traffic approximations for many server queues with abandonment. *Management Science* **50**(10) 1449–1461.

Whitt, W. 2005. Engineering solution of a basic call-center model. *Management Science* **51**(2) 221–235.

Whitt W. 2006a. A multi-class fluid model for a contact center with skill-based routing. *International Journal of Electronics and Communications* **60**(2) 95–102.

Whitt, W. 2006b. Fluid models for multiserver queues with abandonments. *Operations Research* **54**(1) 37–54.

Whitt, W. 2006c. Sensitivity of performance in the Erlang A model to changes in the model parameters. *Operations Research* **54**(2) 247–260.

Whitt, W. 2006d. The impact of increased employee retention upon performance in a customer contact center. *Manufacturing and Service Operations Management* **8**(3) 221–234.

Whitt, W. 2006e. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* **15**(1) 88–102.

Witt, L. A., M. C. Andrews, D. S. Carlson. 2004. When conscientiousness isn't enough: Emotional exhaustion and performance among call center customer service representatives. *Journal of Management* **30**(1) 149–160.

Workman, M., W. Bommer. 2004. Redesigning computer call center work: A longitudinal fields experiment. *Journal of Organizational Behavior* **25**(3) 317–337.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-servers asymptotics of the m/m/n+g queue. *Queueing Systems: Theory and Applications* **51**(3–4) 361–402.

Zohar, E., A. Mandelbaum, N. Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science* **48**(4) 566–583.