M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 14, No. 1, Winter 2012, pp. 66–81
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs®
http://dx.doi.org/10.1287/msom.1110.0349
© 2012 INFORMS

Routing to Manage Resolution and Waiting Time in Call Centers with Heterogeneous Servers

Vijay Mehrotra
School of Management, University of San Francisco, San Francisco, California 94117, vmehrotra@usfca.edu

Kevin Ross
Jack Baskin School of Engineering, University of California, Santa Cruz, Santa Cruz, California 95064, kross@soe.ucsc.edu

Geoff Ryder
SAP Labs, Palo Alto, California 94304, geoff.ryder@sap.com

Yong-Pin Zhou
Michael G. Foster School of Business, University of Washington, Seattle, Washington 98195, yongpin@uw.edu

In many call centers, agents are trained to handle all arriving calls but exhibit very different performance for the same call type, where we define performance by both the average call handling time and the call resolution probability. In this paper, we explore strategies for determining which calls should be handled by which agents, where these assignments are dynamically determined based on the specific attributes of the agents and/or the current state of the system. We test several routing strategies using data obtained from a medium-sized financial service firm's customer service call centers and present empirical performance results. These results allow us to characterize overall performance in terms of customer waiting time and overall resolution rate, identifying an efficient frontier of routing rules for this contact center.

Key words: contact centers; call resolution; skill-based routing; performance management
History: Received: August 14, 2009; accepted: June 7, 2011. Published online in *Articles in Advance* October 7, 2011.

1. Introduction

Over the past two decades, customer service call centers have become a very important part of many companies' business operations; today, inbound call centers employ millions of agents across the globe and serve as a primary customer-facing channel for many different industries. There has also been a great deal of research interest in call center operations management, with the extensive and evolving literature thoroughly surveyed by Gans et al. (2003) and Aksin et al. (2007). Much of this research has focused on models for queueing, staffing, and performance analysis, which in turn provide critical input into personnel scheduling and rostering models.

For example, one common operational setting is a call center in which there is a single type of inbound call (referred to in this paper as a "single queue" model). In this setting, a key operational challenge is to determine how many agents to staff to achieve some waiting time objective. The two most common waiting time objectives are (a) a target mean waiting time (referred to as "average speed of answer" and commonly abbreviated as ASA) and (b) a target percentage of calls answered within some target time period (referred to as "service level" and commonly abbreviated as SL). Similarly, for a given level of staffing, steady-state queueing equations are typically used to estimate the ASA and SL values that will be achieved for a particular time period.

For models in which there are multiple types of inbound calls, the staffing and performance analysis problems become significantly more challenging when some or all of the agents are able to handle more than one type of call. This latter setting is often referred to as "skill-based routing," because calls are routed to different agents (or groups of agents) based on logic that depends at least in part on which agents are capable and/or particularly skilled in handling which types of calls. Good overviews of routing rules in these types of environments are given by L'Ecuyer (2006), whereas efficient rules are described by Armony (2005). Koole et al. (2003) analyze skill-based routing in a network where jobs not processed upon arrival are overflowed to other nodes (i.e., there is no queueing at the nodes).

The staffing challenge in this setting is to simultaneously determine how many agents should be staffed and which skills and priorities each agent should be assigned to achieve particular ASA or SL targets for each queue. Similarly, for a given combination of staffing level and routing logic, discrete-event simulation models are typically used to estimate the ASA and SL values that will be achieved for each particular inbound call.

Historically, the vast majority of the research literature has either used ASA and SL as the primary performance metric with which to judge a particular staffing configuration in both the single queue and the skill-based routing settings. This is because customer waiting time has historically been viewed as a key factor in determining a customer's satisfaction with the service delivered by the call center, because it is widely agreed that customers prefer to spend little or no time waiting for service.

More recently, some researchers have begun to model customer reneging (which in the call center context is typically referred to as "abandonment") and to include the customer abandonment rate (AR) as an important metric in evaluating operational performance (see Mandelbaum and Zeltyn 2007 for a good survey of the state of the art in this area). There are two main reasons for including customer abandonment in call center models. First of all, customers who abandon the queue are quite likely dissatisfied with the experience, and therefore this metric is an important one for call center managers who are focused on delivering high-quality customer service. Second, the effect of customer abandonment is to reduce the total traffic in the call center, and thus abandonment can have a significant impact on staffing needs and on customer waiting times.

It is important to note that the ASA, SL, and AR are all metrics that are based on a customer's waiting experience prior to service. However, it is well known in the marketing literature that the customer's experience *during* service is also a very strong determinant of customer satisfaction and loyalty. In the call center industry, this has led to a strong focus on additional operational performance metrics to understand (a) customers' perception of their service experiences and (b) the quality of service delivered by individual agents and by groups of agents. Two related and very important metrics are call resolution probabilities and call resolution rates.

The resolution probability (RP) is an attribute of an individual agent (or group of agents) for a particular type of call. We define the RP to be the likelihood that a call that is received by the call center and handled by a particular agent will be successfully resolved without requiring a subsequent phone call by the customer. For a specific type of call and an individual agent (or group of agents), the RP can be calculated from historical data as described in §4.1.5.

By contrast, the call resolution (CR) rate and first call resolution (FCR) rate are attributes associated with a particular type of call. Specifically, we define the CR rate as the overall proportion of inquiries that are successfully resolved without requiring a follow-up phone call. The related FCR rate is the proportion of inquiries that are resolved during the *very first* interaction a customer has with the call center. Therefore, for a particular call type, both the CR rate and the FCR rate are operational performance metrics that depend on the specific routing rules that determine how many calls of that type are handled by which specific agents (or groups of agents), and on the RP values for that call type associated with each of those specific agents (or groups of agents).

This important distinction between the RP input values and the CR and FCR performance metrics is made throughout this paper.

After a customer has received service from a call center agent on a particular issue, a subsequent call from that customer about the same issue is a clear sign that the issue had not been resolved during the previous service encounter, and this unsuccessful resolution of the issue that caused the customer to call is a strong sign of customer dissatisfaction. Thus, CR and FCR rates are very important customer-centric operational metrics in practice, though these have been largely absent from the academic literature on call center operations. Higher CR and FCR rates result in reduced system congestion (because of decreased callbacks and hence lower total call arrival rates) and subsequently lower staffing costs. There is a separate literature on retrial queues where jobs that leave the system *before* getting served may retry after some amount of time has elapsed (Falin and Templeton 1997, Aguir et al. 2004). The retrial rates in such settings also impact system congestion.

As data collection and analysis technologies for accurately measuring RP values begin to emerge, call center managers are increasingly focused on managing the CR and FCR metrics. As such, these metrics have been attracting more attention from call center industry leaders and researchers.

In many call centers, agents who have been trained to handle several different types of calls are also known to exhibit very different performance across specific types of calls, where performance is defined by both AHT and RP (Gans et al. 2010 empirically study the agents' heterogeneity in AHT, but not in RP). A challenge for call center managers is to determine how to make use of this information to determine which types of calls should be handled by which types of agents under which system conditions.

In this paper, we explore strategies for routing multiple types of calls to a large group of heterogeneous agents, where these assignments are made dynamically based on the specific attributes of the agents and/or the current state of the system. Our motivation for this research comes from the increasingly high-quality estimates for RP values that are available in many of today's call centers; our initial focus was on developing rules to turn this raw data into actionable information to support routing decisions and improved operational performance. As such, an important focus of our research are the empirical results presented in the case study in §4. Beyond this specific case, we also believe that this paper makes several important generalizable contributions to the call center operations management literature.

First of all, we explicitly model the relationship between RP and *effective* arrival rates to the call center by explicitly accounting for phone calls that occur as a result of previous calls that are not successfully resolved on previous attempts (we refer to these as "callbacks"). Using these effective arrival rates, we also develop an optimization model to maximize the CR rate under relatively general conditions. Second, we develop an efficient frontier of rules with respect to two call center performance measures (ASA and CR) and use this to evaluate the performance of different routing rules for an actual call center's operations. Finally, we conduct a set of simulation experiments to demonstrate that the numerical results from our case study hold empirically under very general conditions.

The remainder of this paper is organized as follows. In §2, we present a survey of the research literature on models that take into account RP and customer callbacks. In §3, we develop several routing strategies designed to minimize the overall ASA, maximize the overall CR rate, or to strike a balance between the two. In §4, we present an empirical case study that examines how our routing rules perform in simulation experiments based on agent data obtained from a financial service firm's customer service call centers. Next, in §5, we present results from a carefully designed set of numerical experiments that reveal that the relative performance of our rules is quite consistent across a diverse set of system parameter values. Finally, in §6, we provide a summary of the paper and its major findings, along with conclusions and possible directions for future research.

2. Literature Review

This paper is focused on call routing strategies to help manage both ASA and CR metrics. Although there has been research on both of these topics, there is surprisingly little work that examines the relationship between them.

Hart et al. (2006) provides a complete review of articles on FCR while also pointing out the importance of measuring and using FCR. This paper also discusses the existence of different operational definitions of FCR, lists various factors that impact RP and FCR (training, empowerment, technology), and explains how higher FCR rates can translate into lower costs and higher levels of customer satisfaction. Similarly, survey results presented by Read (2003) reveal that FCR rates drive customer satisfaction. Feinberg et al. (2002) state that FCR is not a significant determinant of customer satisfaction in the banking/financial services sector; however, these authors readily admit that their metric for customer satisfaction (percentage of customers who give "top box" evaluation) is a weak measure and may have accounted for the results. Cross (2000) cites the importance of the FCR metric, though he also warns against using FCR as the only performance measure. He argues that by focusing only on FCR, a manager may overlook opportunities to reduce the volume of non-value-added but simple-to-answer calls or possible ways to use callback or fax-back options to smooth demand.

Operations management researchers have paid comparatively little attention to models and methods for managing CR and FCR rates. However, there are many published papers that describe call routing and resource allocation rules for call centers, and below we discuss several that are relevant to our work.

Early work on routing in call centers considered either a homogeneous customer/call population or a homogeneous population of servers. Under those conditions, several important results are known about optimal allocation policies or maximal throughput policies under heavy traffic conditions. Most of these use queue backlog rather than waiting time as the control for deciding service allocation. A commonly accepted terminology differentiates between quality-driven (QD) and efficiency-driven (ED) regimes, emphasizing either utilization of servers or service quality. A balanced regime, referred to as quality and efficiency driven (QED), leads to the square root staffing rules as discussed by Halfin and Whitt (1981) and Borst et al. (2004), among others.

More recently, several researchers have extended routing models to consider a heterogeneous population of service agents. In this context, the maximum feasible arrival rate has been characterized by Armony and Bambos (2003), Dai and Lin (2005), and Stolyar (2004), and policies known as maximum pressure or cone policies are known to keep all queues stable whenever that is achievable. These policies essentially maximize the inner product of a service rate vector with the vector of queue states, routing calls with large backlogs to servers with high service rates. In Stolyar (2004), these policies are shown to optimize certain backlog-driven performance measures over time.

Gurvich et al. (2008) address the issue of how many servers are required and how to match them with customers to minimize staffing cost, subject to class-level quality-of-service constraints. They characterize priority control policies that are based on an idle-server-based threshold and show that these policies are asymptotically optimal as the service load grows to infinity. They also show good performance on relatively small systems.

Traditional routing algorithms try to match call types with agent skills subject to service constraints, but do not consider agents' preferences or performances for call types. Given the high rates of agent attrition found in the call center industry and the associated high cost of such turnover, Sisselman and Whitt (2007) propose routing algorithms that account for agent skills and preferences in an effort to balance customer waiting times with agent job satisfaction levels, which they refer to as "preference-based routing." They do so by assigning values to call type–agent combinations that incorporate management's judgment of the value of such pairings and each agent's preferences for the call types. Moreover, preference values are modeled generically and thus could be related to agents' call resolution probabilities in a framework similar to the one presented in this paper.

More closely related to our paper is de Véricourt and Zhou (2005). This paper considers RP in making call routing decisions. There is only one call type, but many agent groups. The agent groups are differentiated by their service rates and RP. These authors show that agent groups can be ranked by their *effective service rate* (the product of resolution probability and service rate), the so-called $p\mu$ rule. After defining their objective as minimizing the average total time to resolve a call, they show that there is always a preferred agent group (the one with the highest $p\mu$) to route the calls to, and when all agents in that class are busy, it is optimal, under certain conditions, to route to other classes following a state-dependent threshold rule. Using numerical tests, the authors show that a routing rule that overlooks the RP differences can perform poorly, which illustrates the importance of routing based on RP as well as service rates. To simplify the routing rule, they show numerically that the optimal state-independent rule already captures almost all the benefits of the state-dependent threshold rule. Moreover, routing solely based on the $p\mu$ index, without the use of thresholds, allows the call center to get most of the benefits.

Another rule proposed in the presence of heterogeneous servers with a single queue is the fastest servers first (FSF) rule in Armony (2005). The rule is described as a QED rule with heterogeneous servers and performs better than its homogeneous counterpart. As discussed below, the author also suggests that the methodology presented in that paper can be extended to prove the optimality of the so-called $p\mu$ rule in the Halfin–Whitt heavy traffic regime, though no formal proof is presented. The optimality of FSF-type rules has also been shown by Mandelbaum and Stolyar (2004), Dai and Tezcan (2008), and Gurvich and Whitt (2009) in slightly different settings, but to date we have been unable to extend such optimality to our skill-based routing setting with multiple call types and multiple unique types (or unique groups) of agents (Dai and Tezcan 2008 provide some insight into the challenge of proving such a result).

All of these papers, it should be noted, define system performance in terms of the waiting time distribution or some direct function of it. In contrast, one important distinguishing characteristic of this paper is that we examine metrics relating to both customer waiting time and call resolution rates. This distinction will be discussed further below.

3. Model Framework and Selected Routing Rules

A customer's experience during a service encounter consists of two parts: the time spent waiting for services and the service itself. Metrics that reflect that waiting time distribution, such as ASA and SL, deal with the first aspect, whereas measures such as CR and FCR rates deal with the second. In practice, as suggested by Cross (2000), most call centers must pay attention to both waiting time and CR rate metrics. However, when there is heterogeneity across different agents for a given type of call—some agents may, on average, handle calls more quickly, whereas others may be more likely to resolve the customer's issue—often there is an inherent trade-off in call routing decisions.

Thus, our goals are (a) to identify routing rules that achieve a balance between the two goals of low ASA values and high CR rates and (b) to empirically examine the trade-off between these two performance measures as a function of the choice of call routing rules.

3.1. Model Setting

Our setting features multiple call types (indexed by $i = 1, 2, \ldots, I$) and multiple agent groups (indexed by $j = 1, 2, \ldots, J$). Calls of type i arrive at a rate of λ_i. There are N_j agents in group j, with $N_j \in \mathbb{Z}^+$, and each agent in group j serves calls of type i with rate μ_{ij}. If agent group j is not capable of handling call type i, then $\mu_{ij} = 0$. When $\mu_{ij} > 0$, we say there is a "match" between call type i and agent group j. In addition, we assume that each agent of group j has a resolution probability p_{ij} for each call of type i, where $p_{ij} \in [0, 1]$.

Our model assumes that all interarrival times are independent of one another, that all service times are independent of one another, and that the outcome of each call (resolved or not) is independent of all other calls. Note that in developing our routing rules, we do not assume that arrivals follow a Poisson process or that the service times are exponentially distributed, although several of our routing rules are motivated by models that do make these assumption (e.g., Armony 2005, de Véricourt and Zhou 2005).

Below, $q_i(t)$ represents the number of type i customers waiting for service at time t, and $Z_j(t)$ is the number of busy agents from group j at time t, where $0 \le Z_j(t) \le N_j$, $\forall\, j$ and $\forall\, t \ge 0$.

We use the term "routing rule" to mean the logic that determines to which agent group an arriving call is assigned if there are no calls in the queue and some agents are free, and also the logic that determines which call an agent is assigned to handle when he/she becomes free when some calls are in the queue waiting for service.

It is important to note that the performance of this system is defined in terms of the steady-state ASA values and CR rates, and that these output metrics depend on the actual numerical values of the input parameters (arrival rates, service rates, call resolution probabilities, and staffing levels) and also on the chosen routing rule that is used to determine which call types are handled by which agents under what conditions. As a result, it is possible for a system to be "unstable"—that is, to have no steady-state mean waiting time—as a consequence of some combination of its input parameters and its routing rules. For example, in cases where the fastest (lowest AHT) agents also have very low resolution probabilities, the effect of the FSF rule from Armony (2005) would be to significantly increase the effective arrival rate due to customer callbacks. Similarly, a rule that routes each call to the available agent with the highest resolution probability regardless of handling times could have a similar effect by increasing the effective mean handle time. Both rules could result in a system load of more than 100%, leading to instability, whereas a more sensible rule could maintain the system load at below 100%. However, we note that neither our case study in §4 nor our simulation experiments in §5 feature this type of system instability.

3.2. Waiting-Centric Routing Rules

In this multiskill setting, the exact form of the optimal policy that minimizes customer waiting time is unknown, but as we stated earlier, the FSF rule in Armony (2005) performs well in the heterogeneous-server setting and is the best known routing rule in such a setting. We will use this rule, which we refer to in this paper as the **Max** μ rule, as our benchmark waiting-centric routing rule.

Rule 1: **Max** μ. When an agent of group j becomes free and there are matching calls waiting, select a call from of type i, where $i = \arg\max_{i:\, q_i(t)>0}\{\mu_{ij} \mid \mu_{ij} > 0\}$; therefore, an agent coming free will choose the matching call type for which she has the highest service rate. Similarly, if an arriving call of type i finds no calls of that type waiting for service and agents of one or more matching groups available, select an agent of group j, where $j = \arg\max_{j:\, Z_j(t)<N_j(t)}\{\mu_{ij} \mid \mu_{ij} > 0\}$; that is, the call will be routed to a matching agent group that has the highest service rate for that call type.

This routing rule ignores the fact that unresolved calls, regardless of how fast they might be handled initially, lead to increased load on the system as a result of callbacks. To address this potential problem, we also test the "$p\mu$ rule" from de Véricourt and Zhou (2005) that explicitly incorporates RP values and thus implicitly considers customer callbacks. We refer to the $p\mu$ metric as the *effective service rate*, and formally specify the "$p\mu$ rule" as follows:

Rule 2: **Max** $p\mu$. When an agent of group j becomes free and there are matching calls waiting, select a call of matching type i, where $i = \arg\max_{i:\, q_i(t)>0}\{p_{ij}\mu_{ij} \mid \mu_{ij} > 0\}$; that is, an agent coming free will choose the matching call type for which she has the highest effective service rate. Similarly, if an arriving call of type i finds no calls of that type waiting for service and agents of one or more matching groups available, select a matching agent group j, where $j = \arg\max_{j:\, Z_j(t)<N_j(t)}\{p_{ij}\mu_{ij} \mid \mu_{ij} > 0\}$; that is, the call will be routed to a free agent from the matching agent group that has the highest effective service rate for that call type.

Indeed, we conjecture that the **Max** $p\mu$ rule, which seeks to maximize the rate at which customers leave the system, actually minimizes the mean waiting time for the system under some conditions. This conjecture is based not only on our own extensive numerical experiments, but also on several papers we mentioned previously in the literature review that have shown the optimality of an FSF-type rule, each in slightly different settings.

We have also extensively tested several other waiting-centric rules (the interested reader is referred to Appendix A in the electronic companion for additional details). These rules have always produced higher ASAs than the **Max** $p\mu$ rule in our tests, while also being dominated on the CR dimension by the rules discussed below in §3.4. As such, we have not included these rules in our case study in §4 or in our subsequent simulation experiments in §5.

3.3. Resolution-Centric Routing Rules

Whereas the **Max** $p\mu$ rule is focused on minimizing the ASA, some call centers place a much higher priority on CR rates. Thus, in this section we discuss routing rules that explicitly prioritize CR rates and introduce an optimization-based routing rule that outperforms all others in terms of CR rates.

Just as managers who focus on minimizing customer wait times will use the **Max** μ rule, an analogous resolution-centric rule would be to route each arriving call to the available agent who has the highest resolution probability for that call. We refer to this as the **Max** p rule. Neither **Max** μ nor **Max** p achieves the intended goal, however, because they do not explicitly account for the RP and its impact on CR and waiting time. We will use **Max** p as the benchmark resolution-centric routing rule.

Rule 3: **Max** p. When an agent in group j becomes free and there are matching calls in queue, select $\arg\max_{i:\,q_i(t)>0}\{p_{ij} \mid \mu_{ij} > 0\}$; that is, that agent will be assigned a call of the type that she is *most likely to resolve*, regardless of waiting times and queue lengths. Similarly, if an arriving call of type i finds no calls of that type waiting for service and agents of one or more matching groups available, assign it to an agent of group j, where $j = \arg\max_{j:\,Z_j(t)<N_j(t)}\{p_{ij} \mid \mu_{ij} > 0\}$; that is, the call will be routed to an agent from the group that has the highest resolution probability for that call type.

However, in our setting there are a number of potential problems with this type of routing rule. For example, this rule may route substantially more traffic to some agent groups than to others. Moreover, this rule may not actually deliver the maximum CR rates because it is nonidling, which means when a call arrives and some matching agents are available, it will be routed to the agent type with the highest *p at that moment*. Similarly, when an agent becomes free, he/she will choose from the matching call types the one with the highest *p at that moment*. There is no consideration of possible future arrivals or service completions that may create a better match. Therefore, there may exist a gap between **Max** p and the maximum possible CR rate (we refer to this difference as the "CR gap"). Not only is there no guarantee that **Max** p will actually deliver the maximum CR rates, there is also no obvious way to quantify the CR gap.

In fact, we have extensively tested several such resolution-centric rules (the interested reader is referred to Appendix A in the electronic companion for a description of these rules), and indeed the CR rates produced by such rules were always lower than the CR rate produced by the optimization-based rule discussed below. In addition, these rules were dominated on both performance dimensions by the hybrid rules discussed below in §3.4. As such, we have not included these rules in our case study in §4 or in our subsequent simulation experiments in §5.

With these issues in mind, we next consider the problem of maximizing the overall expected CR rate for a given set of performance parameters and constraints. Once this optimization problem has been formulated and solved, the results will provide the basis for several subsequent routing rules. In addition, the optimal CR rate from our optimization model can also be used to estimate the CR gap for any other routing rules.

Our model seeks to identify the randomized call routing rule (that is, a routing rule that independently assigns each arriving call to a server according to fixed probability values) that maximizes the overall expected CR rate. To address the issues of stability and fairness across agent groups, we include constraints that specify minimum and maximum utilization targets for each of the agent groups. In addition, to ensure that each call type is treated somewhat fairly, we also include constraints that specify minimum and maximum effective utilization targets for each of the call types.

To determine these utilization levels, we must first calculate the *effective* arrival rate to each queue, taking into account callbacks due to unresolved earlier calls. Our model assumes that customers have no alternative to resolving their call through the call center, and hence all unresolved calls will return as future arrivals.

Denote λ_i to be the external arrival rate for (first-time) calls from customers of type i. Let π be any randomized routing rule, and let x_{ij} be the resulting proportion of calls of type i routed to agent group j under π. The effective arrival rate $\bar{\lambda}_i$, accounting for all the arrivals that result from unresolved calls, explicitly depends on the choice of the x_{ij} values, because they determine the percentage of customers who call back. In particular, we have

$$\begin{aligned}\bar{\lambda}_i &= \lambda_i + \lambda_i\biggl(\sum_j (1-p_{ij})x_{ij}\biggr) + \lambda_i\biggl(\sum_j (1-p_{ij})x_{ij}\biggr)^2 \\ &\quad + \lambda_i\biggl(\sum_j (1-p_{ij})x_{ij}\biggr)^3 + \cdots \\ &= \lambda_i \sum_{k=0}^{\infty}\biggl(\sum_j (1-p_{ij})x_{ij}\biggr)^k.\end{aligned}$$

The kth term in the summation corresponds to the expected number of customers who are making the kth call to resolve the same problem, $k = 1, 2, \ldots$. Now because $\sum_j (1-p_{ij})x_{ij} < 1$, we have

$$\bar{\lambda}_i = \frac{\lambda_i}{1 - \sum_j (1-p_{ij})x_{ij}}.$$

For an agent group j, their total arrival rate for calls of type i is $\bar{\lambda}_{ij} = \bar{\lambda}_i x_{ij}$, and hence their total utilization, which we denote ρ_j, can then be calculated as follows:

$$\rho_j = \sum_i \frac{\bar{\lambda}_{ij}}{N_j \mu_{ij}} = \biggl(\sum_i \frac{\bar{\lambda}_i x_{ij}}{\mu_{ij}}\biggr)\bigg/ N_j.$$

With this expression for the ρ_j values, we then include constraints that specify minimum and maximum allowable utilization levels for each of the agent groups, which we denote respectively as ρ_j^- and ρ_j^+.

We also seek to protect against one or more call types receiving insufficient attention by constraining the effective utilization associated with each call type i, that is, each call type i must be served at total utilization between lower bound τ_i^- and upper bound τ_i^+. To define these constraints, we must first define what we mean by utilization associated with call type i, which will require us to calculate the effective service attention given to calls of type i from all agent groups. For an agent group j, their total fraction of time spent serving queue i is $(\bar{\lambda}_{ij}/(N_j\mu_{ij}))/\sum_{i'}(\bar{\lambda}_{i'j}/(N_j\mu_{i'j}))$. Therefore the total service rate to calls of type i is $\bar{\mu}_i = \sum_{j=1}^{J} N_j\mu_{ij}((\bar{\lambda}_{ij}/\mu_{ij})/\sum_{i'}(\bar{\lambda}_{i'j}/\mu_{i'j}))$. The total effective utilization associated with call type i, which we denote τ_i, can then be calculated as follows:

$$\begin{aligned}\tau_i = \frac{\bar{\lambda}_i}{\bar{\mu}_i} &= \frac{\bar{\lambda}_i}{\sum_{j=1}^{J}(N_j\bar{\lambda}_{ij}/\sum_{i'}(\bar{\lambda}_{i'j}/\mu_{i'j}))}\\ &= \frac{\bar{\lambda}_i}{\sum_{j=1}^{J}(N_j\bar{\lambda}_i x_{ij}/\sum_{i'}(\bar{\lambda}_{i'} x_{i'j}/\mu_{i'j}))}\\ &= \frac{1}{\sum_{j=1}^{J}(N_j x_{ij}/\sum_{i'}(\bar{\lambda}_{i'} x_{i'j}/\mu_{i'j}))}.\end{aligned}$$

We note that as long as $\rho_j < 1$ for all agent groups j, we will also have $\tau_i < 1$ for all call types i. However, just as placing tighter bounds on ρ_j can ensure a *fair* distribution of workload between agent groups, bounds on τ_i can manage the relative treatment of different call types. For example, bounds on τ_i could avoid routing too many calls of type i to the busiest agent groups, and therefore improve the waiting time for calls of that type. This enables a contact center manager to set both fairness and prioritization constraints and meet call-type-specific performance targets, which are common in practice.

At this point, we are ready to formally present our optimization model, characterized by the parameters $0 \le \rho_j^- \le \rho_j^+ \le 1$ and $0 \le \tau_i^- \le \tau_i^+ \le 1$.

$$\begin{array}{rll}\text{Maximize} & \sum_{i,j}\bar{\lambda}_i p_{ij} x_{ij} & \text{(max total call resolution rate)}\\ \text{subject to} & 0 \le x_{ij} \le 1 \quad \forall\, i, j & \text{(fraction of calls bound)}\\ & \sum_{j} x_{ij} = 1 \quad \forall\, i & \text{(total calls routed to different agent groups)}\\ & \rho_j^- \le \rho_j \le \rho_j^+ \quad \forall\, j & \text{(utilization of each agent group)}\\ & \tau_i^- \le \tau_i \le \tau_i^+ \quad \forall\, i & \text{(utilization for each call type)} \quad (1)\end{array}$$

By ignoring the waiting time aspect of the problem and focusing solely on maximizing CR, formulation (1) represents a fluid version of the original problem. Assuming that there is a feasible solution to (1), solving (1) suggests a routing rule that will achieve the maximum CR; we define such a rule below. Conversely, if there is no feasible solution to (1), this demonstrates that the staffing level is not sufficient to handle its call volume, and in practice this would result in high waiting time and abandonment rates.

Note that formulation (1) is quite flexible. For example, the objective function for the optimization model (1) seeks to maximize the weighted sum of the CR rates across all call type and agent group combinations, where the weight placed on a particular call type i and agent group j is the product of the effective arrival rate $\bar{\lambda}_i$ and the proportion x_{ij} of type i calls routed to agent group j. However, by replacing $\bar{\lambda}_i$ with λ_i in the objective function, the resulting formulation can be used to solve the closely related problem of maximizing FCR, a change that also makes the objective function linear in the decision variables x_{ij}. Similarly, if a call center values higher CR rates for a particular call types i, additional weighting factors can be added in the objective function. Also, constraints on utilization can effectively be eliminated by setting $\rho_i^- = 0$ (and/or $\tau_i^- = 0$) and setting $\rho_i^+ = 1$ (and/or $\tau_i^+ = 1$).

Both the objective function and the utilization constraints are quadratic in decision variables, so that an optimal solution for this model can be efficiently obtained with any good commercial solver.

Our solution to this optimization problem provides us with an estimate for the expected value for the optimal CR rate, which can be used to help us quantify the CR gap. In addition, our solution identifies the x_{ij} values that maximize the expected CR rate, which we denote x_{ij}^*. The following routing rule uses these x_{ij}^* values to randomize the routing of each incoming call and produces the maximum CR rate in steady state.

Rule 4: **OptXRand**. Upon arrival, each call of type i is routed to agent group j with probability x_{ij}^*. Once routed to a particular agent group j, a call of type i waits in the FCFS queue for an agent of group j.

3.4. Hybrid Routing Rules

By construction, rule **OptXRand** focuses exclusively on CR rates and indeed has the virtue of achieving the optimal expected CR rate within the specified utilization bounds. However, by focusing solely on the CR rates, this routing rule also allows agents in one or more groups to be idle while calls in one or more groups are queued elsewhere, thereby removing the pooling benefits and thus resulting in potentially long waiting times. Similarly, whereas rule **Max**$p\mu$ does take RP values into account, it is not designed with any considerations for the achieved CR rate.

In this section, we propose two routing rules that seek to combine the waiting-centric **Max**$p\mu$ rule and the resolution-centric **OptXRand** rule.

Rule 5: **CallSwap**. Calls are routed to agent groups according to rule **OptXRand**. However, each queue has a predefined threshold above which the queue is considered "full." If the arrival of a call of type i to the queue for agent group j results in queue j becoming full, and if there are one or more agents free in one or more groups g free, where $g \neq j$, then route this call to an idle agent from the group that has the highest $p\mu$ value. Similarly, when an agent from group j comes free, (a) if there is a call waiting in queue j, then take the call at the head of that queue; (b) if the queue is empty, then check all other queues and take the call at the head of the full queue for which this agent has the highest $p\mu$ value; (c) if no queue is full, then the agent becomes idle.

In a sense, **CallSwap** is really a collection of routing rules that alternates between **OptXRand** and **Max** $p\mu$ depending on the state of the system and the value of the threshold parameter. The larger (smaller) the threshold parameter value, the more closely the rule adheres to **OptXRand** (**Max** $p\mu$). One of the major questions that we seek to address in the case study and numerical experiments that are described below is how the performance of rules of this class varies with the value of this threshold parameter.

The optimization problem (1) may have an optimal solution for which $x^*_{ij} = 0$ for multiple values of i and j. For such systems we propose another routing rule that combines the **Max** $p\mu$ rule and **OptXRand** rule by following the **Max** $p\mu$ rule after first restricting the choices to only combinations for which $x^*_{ij} > 0$. This rule is formally specified below.

Rule 6: **Max** $xp\mu$. When an agent of group j becomes free at time t, if there are available calls of matching types, then select a call of type i, where $i = \arg\max_{i:\, q_i(t)>0}\{p_{ij}\mu_{ij} \mid \mu_{ij} > 0,\, x^*_{ij} > 0\}$; that is, an agent coming free will choose the call type for which she has the highest effective service rate from among those agent groups for which $x^*_{ij} > 0$. Similarly, if an arriving call of type i finds no calls of that type waiting for service and agents of one or more matching groups available, select a matching agent group j, where $j = \arg\max_{j:\, Z_j(t)<N_j(t)}\{p_{ij}\mu_{ij} \mid \mu_{ij} > 0,\, x^*_{ij} > 0\}$; that is, a call of a particular type that arrives when one or more agents from multiple matching groups are free will be routed to a free agent from the matching agent group that has the highest effective service rate for that call type from among those agent groups for which $x^*_{ij} > 0$.

4. Call Center Case Study

Having suggested several routing rules in §3, we next seek to understand how well each of these rules performs by conducting a simulation case study utilizing an actual call center's data. In this case study and in the simulation experiments presented in §5, we examine the performance of these routing rules in terms of the two key performance metrics of overall ASA and aggregate CR rate. Although it is natural to try to identify which of these rules delivers the "best" operational performance, in our setting there is no clear answer to this question, in large part because different call center managers are likely to put different relative weights on each of the two key performance measures. Instead, we present our performance results on a two-dimensional graph, with the CR rate on the x-axis and the ASA on the y-axis. Our goal is to illustrate how different routing rules present the call center manager with different trade-offs.

Below we describe the operational input data in §4.1, discuss the simulation modeling platform in §4.2, and then present and discuss the numerical results in §4.3.

4.1. Database Characteristics and Preparation

Our case study is based on input data obtained from a financial services firm's call center. The database contains records associated with just over 2.7 million individual incoming customer phone calls, which constitute all of the calls for a single calendar year. Specifically, each record in the database contains the following five fields:

- the date and time of the call;
- the unique ID number for the agent who handled the call;
- the *call type* for that call;
- the time spent by the agent on the phone handling the call, or *handle time*;
- the resolution status of the call (see below).

The full database features more than 150 call types and over 500 individual agents. We elected to use only a subset of the call types and agents to focus on the core analytical questions of interest and to ensure that the run times for our simulations were fast enough to conduct extensive numerical experiments. The process of selecting and preparing the data to support our numerical experiments is described below.

4.1.1. Selection of Call Types. Analysis of the database revealed that a high concentration of workload came from just a few call types. Because the highest volume call types drive the overall performance, and the number of call types is a significant driver of simulation time, we chose the four largest call types. Collectively, these four call types comprised just over 25% of all call records in the data set, with each of them featuring significantly higher call volume than all other call types.

4.1.2. Selection of Agents. We filtered the database to include only those agents who handled at least 50 calls for each of the four major call types. Based on our conversations with the business owner of this database, we understood that this filtering had the effect of eliminating supervisors who would occasionally step in to handle phone calls as well as specialists who would be called upon only to deal with specific and advanced topics for a subset of our four call types. As a result, we selected a total of 228 agents for our study.

4.1.3. Creation of Agent Groups Through Cluster Analysis. Because contact centers typically operate in managed teams, we grouped the agents into a total of 20 groups. Our approach was to use cluster analysis to group the agents based on their performance, where our measures for performance were the p_{ij} and μ_{ij} values for each agent for each of the four queues. Using the JMP software, we implemented k-means clustering to divide the population of 228 agents into 20 groups of various sizes. The number of agents in each agent group used in our simulation are given in Appendix B in the electronic companion.

Our choice of 20 agent groups is somewhat arbitrary and is made only to limit the overall computing time associated with our simulations. We note that this is a conservative choice, because modeling each individual agent as a "group" of size one would have the effect of increasing the benefits of the routing rules from §3, because the rules would be able to take advantage of the differences across individual agents' parameter values, rather than just the differences in the group averages.

4.1.4. Arrival Rate Selection. For our numerical experiments, we assume that call arrival of each type i follows a Poisson process, and we choose arrival rates for each of the call types to maintain the same relative proportion t_i of expected calls of call type i as we found in the original database. For a given total call arrival rate across all four call types, which we denote by λ, the individual arrival rate for call type i is then $\lambda_i = t_i\lambda$. The t_i values are given in Appendix B in the electronic companion.

We chose a value for the total call arrival rate λ so that under a first-come, first-served routing rule the overall utilization level would be approximately 90%. This target is consistent with the performance of agents in an efficient call center and corresponds to a critical load where queue backlogs need to be carefully managed. This is calculated to be $\lambda = 2{,}160$ calls per hour.

4.1.5. Service Rate and Resolution Probability Estimation. For each combination of call type and agent group, we assume that service times are exponential. To estimate the service rates, we first calculate the AHT value for each call type by computing the mean of the handle time field. The service rate μ_{ij} for agent group j for call type i was then calculated as the reciprocal of the respective AHT value.

The process of calculating resolution probabilities was somewhat more involved. First, we note that the resolution status field in our database takes one of six different values:

1. the customer's inquiry was resolved to a satisfactory level, such that the customer need not call again regarding this particular issue;
2. this call is the first of two or more transactions needed to satisfy the customer;
3. this is one of potentially multiple intermediate customer follow-up calls about this issue;
4. this is the last call out of a repeated series of customer calls regarding the same issue;
5. the calling party was not recognized as a customer;
6. "other," a catchall category for odd call types.

For our analysis, we created the binary resolution field by mapping resolution status levels 1, 4, 5, and 6 to "successful" and mapping levels 2 and 3 to "unsuccessful."

From here, we calculated the resolution probabilities p_{ij} for agent group j for call type i by computing the proportion of successful calls from the binary resolution field.

All pairwise service rates and resolution probabilities used in our simulation experiments are presented in Tables 1 and 2 in Appendix B in the electronic companion, respectively.

4.2. Simulation Platform

Our experimental simulation platform consisted of a collection of Java programs that invoke the ContactCenters simulation library[1] (see Buist and L'Ecuyer 2005, L'Ecuyer and Buist 2006). All interarrival and service times in the simulation were exponential random variables.

4.2.1. Simulation Run Length. For each of the rules described in §3, we simulated a total of 15,000 realizations. For each realization, we used, in simulated time, a seven-hour warm-up period followed by one hour of data recording. The length of the warm-up period for each realization was determined by the randomization test described by Mahajan and Ingalls (2004) and Yucesan (1993) so that the mean values of output variables reported in our results accurately represent the system in steady state. More details about simulation run lengths and simulation parameters are given in Appendix B in the electronic companion.

[1] See http://www-etud.iro.umontreal.ca/~buisteri/contactcenters/ for more details on this open source library.

4.2.2. Simulating Callbacks. For all rules, we have assumed that unresolved calls result in immediate callbacks to the call center. In practice, there will typically be a random delay prior to subsequent calls, which can result in increased call volume during specific future periods that feature more (or less) congestion than the period in which the original call took place. Because our experiments focus on the steady-state behavior of the system (and arrival rates and staffing levels are stationary), we believe that the zero delay will not have a big impact on the validity of the results, just as shown numerically by de Véricourt and Zhou (2005).

4.3. Numerical Results and Discussion

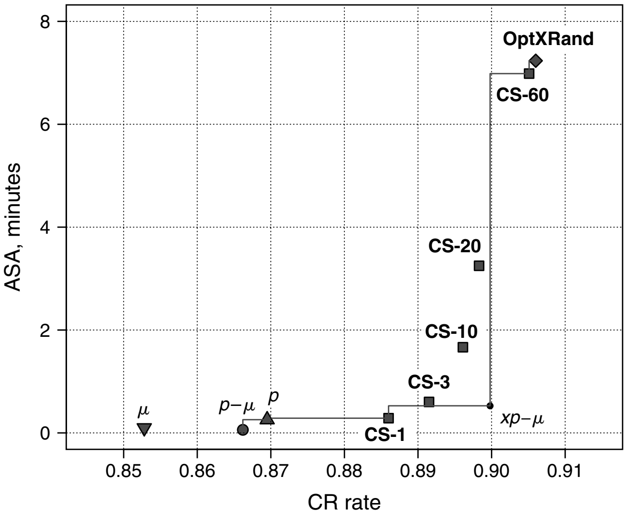
Figure 1 includes the ASA values and CR rates for all rules from §3. For this paper, we consider a routing rule to be on the *efficient frontier* of rules if *it is not dominated by another rule in terms of both ASA value and CR rate*. This differs from the common economic interpretation of a frontier, where convex combinations of rules form the frontier. Because there is no natural way to create a convex combination of two routing rules and expect to achieve the corresponding combination of performance, it is more natural to view the frontier as a set of discrete points representing the actual performance of individual rules.

The results presented in Figure 1 reveal some important insights. First of all, we see that our baseline **Max** μ rule, which myopically routes calls of a particular type to the available agent with the highest service rate for that call type, actually results in a higher ASA than the **Max** $p\mu$ rule, which instead routes calls of a particular type to the available agent with the highest *effective* service rate for that call type.

This suggests a very important operational insight: even the manager who cares only about achieving shorter customer waiting times (and not at all about the achieved call resolution rates) can benefit from considering resolution and using one of our intelligent routing rules. However, to realize these waiting time benefits, the call center must in turn make the commitment to measure resolution probabilities and utilize a routing rule that is based in part on this RP data.

Similarly, the myopic, resolution-centric **Max** p rule actually does not produce the maximum CR rate. Instead, by adopting the optimal allocation of calls to agents, **OptXRand** achieves the highest CR of all the policies, and the CR gap for **Max** p is 3.65%. **OptXRand** achieves this by keeping calls waiting in the agent queues they are assigned to, even though there may be another matching agent that is idle at that moment. This increase in CR rate comes at the expense of customer waiting time, however. We observe that **OptXRand** has the highest ASA value.

Figure 1 **Routing Rule Efficient Frontier**



	ASA (min.)	CR
Max μ	0.098	0.8528
Max $p\mu$[a]	0.060	0.8662
Max p[a]	0.257	0.8695
CS-1[a]	0.286	0.8860
CS-3	0.602	0.8915
CS-10	1.665	0.8961
CS-20	3.249	0.8983
CS-60[a]	6.985	0.9051
Max $xp\mu$[a]	0.527	0.8998
OptXRand[a]	7.235	0.9060

Note. **CS**-T, **CallSwap** with threshold T.
[a] Rules on the efficient frontier.

Next, we note that the **Max** $p\mu$ rule features the lowest ASA among all points on the efficient frontier at 0.06 minutes or 3.6 seconds. On the other end of the spectrum, we observe that the **OptXRand** rule features the highest CR rate at 90.6%. As expected, each of these rules performs well only on one of the performance measures: the **OptXRand** rule produces an ASA value of just over seven minutes, whereas the **Max** $p\mu$ rule results in a CR gap of 4% with **OptXRand**.

Bridging these two extreme points, we see that the **CallSwap**-class rules enable managers to express the relative importance placed on the ASA and CR rate metrics by selecting appropriate parameter values. In particular, for all parameter values, these rules result in significantly lower ASA values than the **OptXRand** rule. We note that the **CallSwap-1** rule delivers a CR rate of over 88.60% along with an ASA of approximately 18 seconds, almost 96% lower than that of **OptXRand**, while producing a CR rate that is two percentage points higher than the **Max** $p\mu$ rule (and a CR gap of two percentage points compared with **OptXRand**). As the "full" threshold increases (e.g., from **CallSwap-1** to **CallSwap-60**), the borrowing of calls becomes less common, and the **CallSwap** rule behaves more like the **OptXRand** rule. For example, the **CallSwap-20** rule delivers a CR rate of 0.73% below that of **OptXRand**, and its ASA increases significantly to slightly less than half of that of **OptXRand**. The **CallSwap-60** rule is much closer in performance to **OptXRand**, achieving 0.23% less in CR rate and a 25 second improvement in ASA. Thus, as the threshold increases, the **CallSwap**-class rules nicely build a bridge between the two anchoring rules **Max** $p\mu$ and **OptXRand**.

Finally, we note that the alternative hybrid rule **Max** $xp\mu$ shows very strong performance, with a better CR rate and ASA than many of the **CallSwap** rules, with a resolution gap of just 0.62% and an ASA of 32 seconds. This is especially important because although it does require the solution of the optimization, it does not carry the same overhead of tracking buffer overflow levels as in the range of **CallSwap** rules. For this case study, **Max** $xp\mu$ dominates the performance of the **CallSwap** rules with thresholds 3 to 20, which motivates further experiments in the next sections. This rule does not have the flexibility of the **CallSwap**-class rules, however, because it represents only a fixed point on the efficient frontier, whereas the **CallSwap**-class rules can cover a wide range of the frontier by adjusting the full threshold (allowing the call center a finer way to balance the two objectives).

These results clearly demonstrate that the choice of routing rules will have a significant impact on the operational performance of this particular call center. Moreover, the "best" rule clearly depends on the relative importance that the call center manager places on achieving high call resolution rates and short customer waiting times.

5. Additional Simulation Experiments

5.1. Motivation for Experiments

It is quite reasonable to question whether the results from §4 are merely a function of the specific values of the key model parameters, most notably the arrival rates λ_i, the service rates μ_{ij}, and the resolution probabilities p_{ij}. In this section, we present the results from a set of simulation experiments that are carefully designed to help us understand whether and under what conditions the results from §4 can be generalized.

In our setting, several parameter relationships are of particular interest. First of all, we consider the input parameters associated with our two dimensions for service quality. Specifically, we expect that the correlation between the service rates μ_{ij} and RP values p_{ij} will have a significant impact on the relative performance of different routing rules. For example, in our case study, the values had a correlation of 0.41.

A positive correlation indicates something about the calls or the agent training. On the one hand, easier calls—those that are more likely to be resolved—are also handled faster. On the other hand, agents who are more experienced or better trained can answer calls both faster and better, resulting in a higher service rate and higher resolution probability. A negative correlation, however, suggests that there is a trade-off between speed and resolution: when agents work faster (hence, a higher service rate), they are less likely to resolve problems (hence, a lower resolution probability).

In addition, we also study the effect of different utilization rates. Our intuition suggests that, all other things being equal, the value of intelligent call routing will decrease as the utilization rates ρ_j increases, because an increase in utilization means that there are fewer opportunities for our routing rules to "arbitrage" across differences between agent groups. In our tests, we seek to verify this hypothesis and quantify the relationship between utilization rate and the benefits available through our intelligent call routing rules.

5.2. Experiment Design

To test the effects of correlation and utilization values, we simulated a call center with three call types and three agent groups, each of size 20. Given the large number of input parameters, our simulation experiments were designed to provide as much clarity and insight as possible about the relative performance of our routing rules under different correlation and utilization levels.

The first set of experiments looked at the correlation values and compared the performance of different routing rules in the presence of correlation values of +1, 0, and −1. The second set of experiments controlled for correlation values and tested the performance of our routing rules under different utilization rates.

The same set of nine distinct p_{ij} values was used in each experimental scenario, ensuring that each case should have a similar potential CR rate. The service rates were initialized to a permutation of the same nine μ_{ij} values for each scenario, increasing with p_{ij} for the positive correlation case and decreasing for the negative correlation case. The service rates were then scaled to give the same value for the sum $\sum_{i,j} \mu_{ij} p_{ij}$, ensuring that each scenario had a similar overall average service rate. As in §4, the arrivals are assumed to be Poisson, and the service times are assumed to be exponential. For the correlation experiments in the following section, the arrival rates λ_i are selected to have the system utilized at 90% when the rule **Max** $p\mu$

is used. The specific values for λ_i, μ_{ij}, and p_{ij} for each experimental scenario are presented in Appendix C in the electronic companion.

The simulation software platform used to conduct the simulation experiments presented below is the same one that was defined in §4 above. In addition, the number of replications were chosen in the same manner as for the case study in the previous section, which is described in Appendix B in the electronic companion.

5.3. Numerical Results: Correlation Experiments

Figure 2 shows three summary plots for the performance of our routing rules, each with different correlation values. The first plot has a correlation of +1, the second has a correlation value of 0, and the third has a correlation value of −1. The extreme cases were created by assigning strictly increasing and strictly decreasing μ_{ij} values relative to the same p_{ij} set, whereas the zero correlation case used a varied permutation of the same values.

From an operational perspective, the first plot is the result of a strong positive correlation between the resolution probabilities and the service times, a case that arises in practice when the heterogeneity among p and μ values is largely due to the agents' experience and training levels (the better agents are able to complete all call types faster and with a higher probability of resolution) or the nature of the calls (easier calls are faster to handle and more likely to be resolved).

The second plot in Figure 2 represents the case when there is no correlation between the among p and μ values. The final plot in Figure 2 has a correlation value of −1, which is the case when there is a trade-off between speed and quality: the faster (slower) agents also tend to be less (more) careful and therefore resolve a lower (higher) portion of calls handled.

Looking at the plots in Figure 2, we take note of several significant results. First of all, we observe that the shape of the efficient frontier is consistent across all three correlation values and that this shape is also consistent with that observed in the case study in §4. Similarly, the relative performance of the different routing rules is largely consistent across all three correlation levels, and these relative positions are similar to what was observed in the results for the case study in §4.

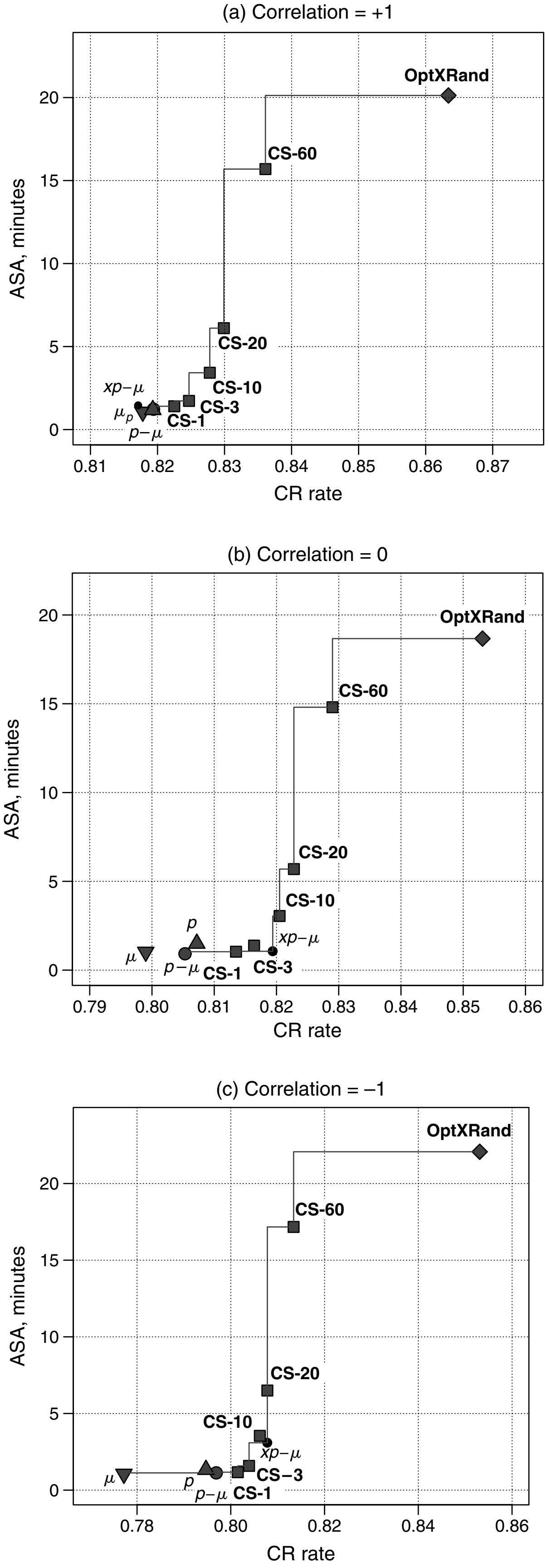
In our case study in §4, we observed the strong performance of the **Max**$p\mu$ rule, which takes into account both the service rates and the resolution probabilities when making its routing decisions, relative to the myopic rules **Max**μ and **Max**p. The results in Figure 2 further confirm this.

More significantly, we observe that when the correlation is strongly positive as in the first plot in Figure 2, there is very little difference in the performance of the rules **Max**$p\mu$, **Max**μ, and **Max**p. However, as the value of the correlation decreases, first to 0 in the second plot and then to −1 in the third plot, we see that both the ASA and the CR results for the **Max**$p\mu$ improve relative to the **Max**μ and **Max**p rules. Our interpretation of these results is as follows: when the correlation between the p and μ values is close to +1, the Maxμ and Max$p$ routing rules make very similar routing decisions, because routing on the basis of a high value for p (or μ) is largely equivalent to routing on the basis of the product $p\mu$. However, as the correlation between the p and μ value decreases, there is more benefit to taking into account both the resolution probabilities and the service rates when making a routing decision. This results from the fact that when the correlation value is lower, routing only on the basis of the highest resolution rates results in longer service times, whereas routing only on the basis of the fastest service rates results in lower resolution rates; in these cases, the **Max**$p\mu$ routing rule makes a better decision in terms of overall system throughput by choosing on the basis of the highest effective service rate.

Similarly, we take note of the resolution gap between the routing rule with the highest CR rate (which in all cases is the **OptXRand** rule) and the routing rule with the lowest resolution rate. In particular, we observe that this gap increases as the correlation decreases across three plots in Figure 2, ranging from 4.56% in the first plot when the correlation is +1 to 7.59% in the third plot when the correlation is −1. This suggests that the potential improvement in CR rate from better routing rules grows as the correlation decreases. We also note that in this experiment, as in our case study in §4, the waiting times associated with the **OptXRand** rule are extremely high, suggesting that this rule would be very unlikely to ever be implemented in practice; its primary purpose is in helping to quantify the resolution gap.

Finally, we take note of the uneven relative performance of the **Max**$xp\mu$ rule, which features the lowest CR rate for the case where the correlation is equal to +1 and a relatively high ASA for the case where the correlation is equal to −1. It appears that the performance of the **Max**$xp\mu$ rule depends heavily on the values of x_{ij} calculated through the optimization stage, which in turn is sensitive to the parameters of the agents in the contact center. We note that if $x_{ij} = 0$ for some (i, j) pair, then agent group j is never sent call type i. This may induce some idling, increasing the achieved ASA while increasing the CR rate by ensuring that calls only go to the preferred agent groups. In particular, if there were *no* (i, j) pairs for which $x_{ij} = 0$, then this rule would be equivalent to the **Max**$p\mu$ rule, but if there were mostly zeros, then

Figure 2 Efficient Frontiers with Different Correlation Values



	ASA (min.)	CR
Max μ	1.062	0.8178
Max $p\mu$[a]	1.224	0.8194
Max p	1.188	0.8190
CS-1[a]	1.405	0.8225
CS-3[a]	1.726	0.8247
CS-10[a]	3.423	0.8278
CS-20[a]	6.109	0.8299
CS-60[a]	15.692	0.8361
Max $xp\mu$	1.457	0.8171
OptXRand[a]	20.128	0.8634

[a] Rules on the efficient frontier.

	ASA (min.)	CR
Max μ	1.024	0.7990
Max $p\mu$[a]	0.922	0.8053
Max p	1.500	0.8072
CS-1[a]	1.037	0.8135
CS-3	1.379	0.8164
CS-10[a]	3.047	0.8205
CS-20[a]	5.690	0.8228
CS-60[a]	14.805	0.8290
Max $xp\mu$[a]	1.061	0.8194
OptXRand[a]	18.760	0.8531

[a] Rules on the efficient frontier.

	ASA (min.)	CR
Max μ[a]	1.053	0.7772
Max $p\mu$[a]	1.125	0.7969
Max p	1.128	0.7959
CS-1[a]	1.170	0.8015
CS-3[a]	1.581	0.8039
CS-10	3.541	0.8062
CS-20[a]	6.498	0.8078
CS-60[a]	17.168	0.8134
Max $xp\mu$[a]	3.09	0.8078
OptXRand[a]	22.072	0.8531

[a] Rules on the efficient frontier.

it would be similar to **OptXRand**. Each of our scenarios lie between these extreme possibilities, impacting the degree to which the rule sacrifices ASA for CR and, in turn, whether or not the rule appears on the efficient frontier.

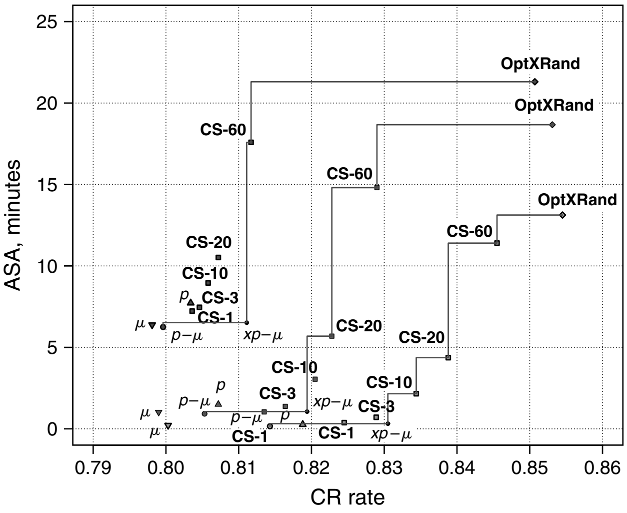
5.4. Numerical Results: Utilization Experiments

Figure 3 shows the impact of utilization on performance. To isolate the impact of utilization, we began by setting the correlation value to zero, just as we had for the second plot of Figure 2. From here, we varied the utilization of the system to values of 85%, 90%, and 95% by appropriately scaling the external arrival rate vector.

Several observations can be made from the plots in Figure 3. Just as in the three plots Figure 2, one key observation is the consistent shape of all three lines in Figure 3, along with the consistent relative position of the different routing rules for each of these different utilization levels.

Once again, we see that in all the scenarios, the **OptXRand** rule is superior to all others in terms of resolution rate but has the worst waiting time performance by a significant margin. Whereas **OptXRand** anchors one end of the ASA–CR efficient frontier, **Max** $p\mu$ anchors the other end, once again dominating the myopic **Max** μ and **Max** p routing rules at all utilization levels.

Figure 3 **Efficient Frontiers with Different Utilization Values**



	Low		Medium		High	
	ASA	CR	ASA	CR	ASA	CR
Max μ	0.215	0.8003	1.024	0.7990	6.382	0.7981
Max $p\mu$[a]	0.154	0.8143	0.922	0.8053	6.250	0.7996
Max p	0.266	0.8188	1.500	0.8072	7.723	0.8034
CS-1	0.380	0.8245	1.037	0.8135	7.228	0.8036
CS-3	0.709	0.8289	1.379	0.8164	7.459	0.8046
CS-10	2.155	0.8344	3.047	0.8205	8.954	0.8058
CS-20	4.368	0.8388	5.690	0.8228	10.519	0.8072
CS-60[a]	11.402	0.8455	14.805	0.8290	17.587	0.8117
Max $xp\mu$[a]	0.316	0.8305	1.061	0.8194	6.520	0.8111
OptXRand[a]	13.126	0.8545	18.760	0.8531	21.306	0.8507

Note. **CS**-*T*, **CallSwap** with threshold *T*.

[a] Rules on the efficient frontier (for all three utilization levels).

In between these two extremes, the class of **CallSwap** rules fill the gap, with the lower threshold rules closer to **Max** $p\mu$ and the higher threshold rules closer to **OptXRand**. For example, going from **Max** $p\mu$ to **CallSwap-1** generates a significant gain in CR rate at all utilization levels. Increasing threshold parameter values for the **CallSwap** rule corresponds to increased server idling (which increases ASA), but also enables better matching of calls to agents in terms of RP values (so that the overall CR rate also increases). This trade-off moves the rule closer to a work-conserving rule and eliminates unnecessary idling; at the same time, this comes at a cost in terms of the CR rate achieved, because the CallSwap policies by construction deviate from the optimal proportion of each type of call being handled by each agent group.

Next, we note that each of the routing rules produces not only higher ASA values but also lower achieved CR rates as the system utilization increases. This inverse relationship between system utilization in CR rates results from the decrease in agent idle time; for nonidling policies such as **Max** $p\mu$, there are simply less opportunities for agents (calls) to choose call types (agents) for which they have a higher RP, whereas policies based on the optimization model (1) are less able to capitalize on discrepancies in RP values because of lower levels of agent idle time.

Similarly, we observe that the **CallSwap**-class rules are all near the frontier for lower threshold values when the utilization level is relatively low (i.e., 85%), but not when the utilization level is high (i.e., 95%). This is reasonable because as the overall system utilization increases, queues will more frequently reach the value of the "full" threshold parameter for a **CallSwap** rule, which in turn leads to more deviation in the actual handling of calls from the optimal proportion x^*_{ij} of each call type i to be routed to agent group j based on (1).

Based on all of our numerical results, our recommendation is that **CallSwap**-type rules be used when utilization is at or below 90%, for under these conditions we observe that significant improvements in CR rates can be achieved with relatively small increases in ASA. Furthermore, we note that the bulk of the improvement in CR rates that the **CallSwap** rules can deliver is captured with parameter values from 1 to 10, whereas **CallSwap** rules with higher parameter values yield relatively small improvements in CR rates, and these only at the cost of significant increases in ASA values.

As utilization increases, the performance of the **Max** $xp\mu$ (which is on the efficient frontier for all levels of utilization) comes to dominate all of the **CallSwap** rules by an increasing amount and for increasing threshold levels. This rule achieves a significantly greater CR rate by ensuring that calls are only handled by the preferred agents, and it sacrifices less in ASA at higher utilization because there is very little idling possible in a fully utilized call center.

6. Conclusions

In this paper, we have examined a complex problem that has not been addressed previously in the research literature. Whereas nearly all the papers in the literature define system performance as a (one-dimensional) function of the waiting time distribution, we model the performance of our call center system in terms of both the length of a customer's wait and the effectiveness of the agent who ultimately handles the call in resolving a customer's issue. In this context, we examine routing rules to improve operational performance on one or both of these output dimensions. Using a data set from a financial services call center, we have conducted a set of simulation experiments to examine the relative performance of these routing rules. In addition, we have conducted additional simulation experiments to extend our understanding of the relative performance of these routing rules along our two performance dimensions.

Our simulation results provide several insights about call routing rules. Our intelligent hybrid routing rules deliver near-optimal CR rates with relatively low ASA values for a variety of parameter values and relationships. For each set of input parameters, we also construct an efficient frontier that is intended to help managers understand the trade-offs between ASA and CR rates.

In closing, we propose several extensions to the research presented in this paper. Whereas our efficient frontier explicitly illustrates the trade-off between ASA and CR rate, one natural next step would be to identify the optimal rule for a given call center based on the cost of customer waiting and the value of successful call resolution. Also, although there has been a significant amount of research on skill-based routing and agent pooling, to our knowledge this research has not considered the impact of such rules on CR rates when different agent groups have different AHT and RP values for different call types. In addition, there is a significant literature on agent learning and attrition (e.g., Pinker and Shumsky 2000, Gans and Zhou 2002, Whitt 2006, Ryder 2009) that suggests that the choice of routing rules may also have an impact on the learning and turnover effects experienced by different agent groups; this is an area that we feel may be worthy of investigation. Finally, we have taken the arrival rates, service rates, and staffing levels as time-independent inputs to our model, though in practice these parameters vary by time of day. This suggests several avenues for future research including methods for call forecasting (taking into account CR rates and customer callback delay time distributions) and agent scheduling (taking into account not only waiting times but also routing rules and resolution rates).

Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://msom.journal.informs.org/.

Acknowledgments

Yong-Pin Zhou's research was supported by the National Science Foundation [Grant CMMI-0645075]. Vijay Mehrotra's research was supported by a faculty development grant from the University of San Francisco. Geoff Ryder's research was supported by the National Science Foundation [Grant EEC-0946373] to the American Society for Engineering Education.

References

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

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call-center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

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

Armony, M., N. Bambos. 2003. Queueing dynamics and maximal throughput scheduling in switched processing systems. *Queueing Systems* **44**(3) 209–252.

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

Buist, E., P. L'Ecuyer. 2005. A java library for simulating contact centers. *Proc. 2005 Winter Simulation Conf.*, IEEE, Piscataway, NJ, 556–565.

Cross, K. F. 2000. Call resolution: The wrong focus for service quality? *Quality Progress* **33**(2) 64–67.

Dai, J. G., W. Lin. 2005. Maximum pressure policies in stochastic processing networks. *Oper. Res.* **53**(2) 197–218.

Dai, J. G., T. Tezcan. 2008. Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* **59**(2) 95–134.

de Véricourt, F., Y.-P. Zhou. 2005. A routing problem for call centers with customer callbacks after service failure. *Oper. Res.* **53**(6) 968–981.

Falin, G., J. G. C. Templeton. 1997. *Retrial Queues*. Chapman and Hall/CRC, London.

Feinberg, R. A., L. Hokama, R. Kadan, I. Kim. 2002. Operational determinants of caller satisfaction in the banking/financial services call center. *Inter. J. Bank Marketing* **20**(4/5) 174–180.

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

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

Gans, N., N. Liu, A. Mandelbaum, H. Shen, H. Ye. 2010. Service times in call centers: Agent heterogeneity and learning with some operational consequences. *Borrowing Strength: Theory*

Powering Applications—A Festschrift for Lawrence D. Brown, IMS Collections, Vol. 6. Institute of Mathematical Statistics, Beachwood, OH, 99–123.

Gurvich, I., W. Whitt. 2009. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* **11**(2) 237–253.

Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service level differentiation in call centers with fully flexible servers. *Management Sci.* **54**(2) 279–294.

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

Hart, M., B. Fichtner, E. Fjalestad, S. Langley. 2006. Contact centre performance: In pursuit of first call resolution. *Management Dynam.* **15**(4) 17–28.

Koole, G., A. Pot, J. Talim. 2003. Routing heuristics for multi-skill call centers. *Proc. 2003 Winter Simulation Conf.*, IEEE, Piscataway, NJ, 1813–1816.

L'Ecuyer, P. 2006. Modeling and optimization problems in contact centers. *Proc. 3rd Internat. Conf. Quant. Eval. Systems (QEST'06)*, IEEE, Piscataway, NJ, 145–154.

L'Ecuyer, P., E. Buist. 2006. Variance reduction in the simulation of call centers. *Proc. 2006 Winter Simulation Conf.*, IEEE, Piscataway, NJ, 604–613.

Mahajan, P. S., R. G. Ingalls. 2004. Evaluation of methods used to detect warm-up period in steady state simulation. *Proc. 2004 Winter Simulation Conf.*, Vol. 1, IEEE, Piscataway, NJ, 663–671.

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

Mandelbaum, A., S. Zeltyn. 2007. Service engineering in action: The Palm/Erlang-a queue, with applications to call centers. D. Spath, K.-P. Fähnrich, eds. *Advances in Services Innovations*. Springer, Berlin, 17–46.

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

Read, B. B. 2003. Call center checkup. *Call Center Magazine* (June 1), http://www.icmi.com/Resources/Articles/2003/June/Call-Center-Checkup.

Ryder, G. 2009. Routing to develop expertise in customer contact centers. Doctoral dissertation, University of California, Santa Cruz, Santa Cruz.

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

Stolyar, A. 2004. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.* **14**(1) 1–53.

Whitt, W. 2006. The impact of increased employee retention on performance in a customer contact center. *Manufacturing Service Oper. Management* **8**(3) 235–252.

Yucesan, E. 1993. Randomization tests for initialization bias in simulation output. *Naval Res. Logist. Quart.* **40**(5) 643–663.