SERVICE SCIENCE
Vol. 7, No. 2, June 2015, pp. 132–148
ISSN 2164-3962 (print) | ISSN 2164-3970 (online)

http://dx.doi.org/10.1287/serv.2015.0101

# Experience-Based Routing in Call Center Environments

Thomas R. Robbins
East Carolina University, Greenville, North Carolina, 27858, robbinst@ecu.edu

In this paper we examine some assumptions commonly made in modeling call centers. In particular, we evaluate the assumption that agents are homogeneous, statistically equivalent servers. We examine empirical data to highlight the issues that create heterogeneity between agents. We explore a call center environment where agents increase their productivity over time, but eventually leave the organization. We consider the implication of this heterogeneity and explore a routing policy that attempts to exploit this heterogeneity and improve long-term call center performance. We consider the application of experience-based routing; that is, routing to agents based on their availability and experience relative to other available agents. We examine policies where calls are routed to the most experienced agents when the call center is busy, to facilitate efficiency, and to the least experienced agent when the call center is slow, to facilitate learning. We investigate the potential improvement in performance that can be achieved by considering agent experience when making routing decisions and characterize the conditions under which the improvement is most significant. We find that routing to agents based on experience can yield substantial improvements over a wide range of conditions.

*Keywords*: service operations management; call center operations; dynamic routing; simulation analysis
*History*: Received August 27, 2013; Received in final revised form March 31, 2014; Accepted April 29, 2015.

## 1. Introduction

Call centers are examples of queuing systems: calls arrive, wait in a virtual queue, and are then serviced by a server/agent. Call centers are often analyzed using standard queuing models that assume that all servers are homogeneous, and that calls are routed to an agent who has been idle for the longest period of time. However, call centers are service systems that differ from standard queuing models in several fundamental ways. Call center servers are human beings that differ from their co-workers in terms of ability, skills, and productivity. The assumption of homogeneous servers is suspect in any service delivery system, but particularly in call center applications where the service being offered can vary widely from call to call and the efficiency of the service is highly dependent on the agent's knowledge. Consider, for example, the case of a technical support help desk; agents must resolve a wide range of technical issues and the speed at which they can diagnose and resolve those issues is greatly influenced by their experience handling similar issues. Under these conditions the assumption of homogeneous servers is quite questionable.

In this analysis we will challenge the assumption of agent homogeneity and examine the implications of agent heterogeneity. While many things create differences between agents, a significant differentiation occurs because of agent experience. If we assume that call center agents are subject to learning curve effects, then agents will become more productive with experience. A call can then be processed most efficiently if the most experienced, and therefore the most productive agent handles it.

In this paper we examine the implications of what we call *experience-based routing*, a process where calls are routed to agents based on the relative experience levels of the available agents and the congestion level of the call center. We consider call center environments where agents become more productive with time but eventually quit. We examine policies where calls are routed to the most experienced agents when the call center is busy, to facilitate efficiency, and to the least experienced agent when the call center is slow, to facilitate learning. Our goal is to determine the impact these policies can have on call center performance and identify the conditions for which the policy is most beneficial.

The remainder of this paper is organized as follows. In §2 we review the associated literature. In §3 we investigate empirical issues that motivate our model. In §4 we discuss our simulation model and the experimental design we develop to evaluate it. In §5 we review the results of our experiments. We conclude in §6 with summary observations and identify future research questions.

## 2. Associated Literature

### 2.1. Call Center Queuing Models

Call centers are generally modeled as queuing systems. Queuing models are used to estimate system performance so that the appropriate staffing level can be determined to achieve a desired performance metric. By far the most common queuing model used for inbound call centers is the Erlang C model known in general queuing terminology as the *M/M/N* queue (Gans et al. 2003, Brown et al. 2005). The Erlang C model is a very simple multiserver queuing system. Calls arrive according to a Poisson process at an average rate of $\lambda$. By nature of the Poisson process, interarrival times are independent and identically distributed exponential random variables with mean $\lambda^{-1}$. Calls enter an infinite length queue and are serviced on a first come, first served (FCFS) basis. All calls that enter the queue are serviced by a pool of $n$ homogeneous (statistically identical) agents. Calls are most often routed to an agent based on a longest-idle-service-first (LISF) policy. Service times are assumed to follow an exponential distribution with a mean service time of $\mu^{-1}$, so the queue is serviced at an average rate of $n\mu$.

The steady state behavior of the Erlang C queuing model is easily characterized, see for example Gans et al. (2003). The offered utilization is defined as $\rho = \lambda/(n\mu)$. The offered utilization represents the proportion of available agent time spent handling calls under the assumption that all calls are serviced. Given the assumption that all calls are serviced, the traffic intensity must be strictly less than one or the system becomes unstable, i.e., the queue grows without bound. The proportion of callers that must queue prior to service, or the *proportion waiting*, is a basic measure of system performance. Another relevant performance measure for call center managers is the *average speed to answer* (ASA): the average time a call spends in queue. The *telephone service factor* (TSF) is the proportion of calls answered within a goal time, for example, the proportion of calls answered within 30 seconds. In addition, we may track the *total time in system*: the time a caller is on the phone including wait time and service time.

A key metric in practice is the *abandonment rate*: the proportion of all calls that leave the queue (hang up) prior to service. Abandonment rates cannot be estimated directly using the Erlang C model because the model assumes no abandonment occurs. Abandonment can dramatically alter queuing system behavior (Gans et al. 2003). Since high abandonment levels are common in many call centers, many researchers advocate the use of the more complex Erlang A model (Gans et al. 2003, Brown et al. 2005). Erlang A assumes that each caller has a finite willingness to wait, and will abandon the queue or hang up if their wait time exceeds their patience. Whereas the Erlang C calculations are relatively straight forward, Erlang A calculations are much more complicated and often require approximations (Mandelbaum and Zeltyn 2007). This complexity is one reason that Erlang C is more widely used.

Both Erlang models assume that calls arrive according to a Poisson process with a known and constant arrival rate and that the queuing system is operating in a steady state. In practice, arrival rates are typically not constant, but instead are time-varying. Since queuing models assume steady state behavior with a constant arrival rate, some approximation is required to apply stationary queuing models to time varying arrival rates. There are several relatively simple approximations that can be considered with time-varying arrival rates (Jennings et al. 1996). Solutions include the simple stationary approximation (SSA) and the pointwise stationary approximation (PSA) (Green and Kolesar 1991, 1997). The solution that is most widely applied in practice is the stationary independent period by period (SIPP) approach. In this approach the day is divided into a number of discrete periods, for example 30 minutes. For each period the arrival rate is assumed to be constant and the queuing system operating in steady state, independent of the conditions in other periods. A series of papers analyze these assumptions and evaluate their validity. The accuracy of this method is analyzed in Green et al. (2001), in which the authors assume a model of sinusoidal varying arrival rates with no abandonment. The authors perform detailed numerical analysis and show that SIPP can lead to poor approximations in cases where the relative amplitude of the sine wave is large, planning periods are long, service rates are low, or the system is large. They present several alternatives to the basic SIPP model including SIPP max, SIPP min, and various lagged versions. The implications of staffing under time-varying demand are analyzed by Green et al. (2005). Liu and Whitt (2012) develop staffing algorithms to stabilize abandonment rates when arrival rates are time varying.

Finally, both Erlang models assume that agents are homogeneous. More precisely, it is assumed that the service times follow the same statistical distribution independent of the specific agent handling the call. Empirical evidence supports the notion that some agents are more efficient than others and the distribution of call time is dependent on the agent to whom the call is routed. In particular, more experienced agents typically handle calls faster than newly trained agents (Armony and Ward 2010).

## 2.2. Learning

The notion that individuals improve their efficiency with repetition dates back as far as the late 19th century when Hermann Ebbinghaus performed an experimental psychology study of human memory. (His work is reprinted in Ebbinghaus 1964.) Ebbinghaus studied subject's ability to memorize three letter nonsense syllables. He found that learning occurs with repetition, rapidly at first, but with a declining rate. Although Ebbinghaus did not use the term, the phenomenon of improved productivity with experience is commonly referred to as the *learning curve*.

In the 1930s the learning curve was applied to the production costs of aircraft engines with a focus on direct labor costs (Wright 1936). Learning curve theory was widely studied in military applications during World War II and the immediate post-war era. A detailed review of military applications in this era is provided in the paper by Asher (1956). Significant academic work on learning curves was done during the 1960s and 70s. A detailed and comprehensive review is provided in Yelle (1979). More recent analysis of learning in industrial scenarios is summarized in Argote and Epple (1990) and Argote et al. (1990), works that focus on organizational differences in learning. Argote's considerable work on learning curves is summarized in her book (Argote 1999).

More recent work has applied learning curve theory to service applications. A number of papers analyze learning curves in medical applications (Pisano et al. 2001, Kaul et al. 2006, Passerotti et al. 2009, Charland et al. 2011). Recent work has also applied learning curve theory to call centers. Kim et al. (2012) use data from a university computing call center to perform detailed econometric modeling of learning curve effects. Robbins and Harrison (2011) demonstrated a statistically significant learning curve effect in an IT help desk environment and evaluated the effect that the learning curve has during the launch of a new service when all agents are inexperienced.

Although many different functional forms have been used for learning curves, the standard learning curve model assumes that task time is reduced by a set percentage with each doubling of cumulative output (see for example Argote 1999). A standard mathematical formulation is

$$Y_x = Kx^n \tag{1}$$

where $K$ is the effort (typically person-hours) required to produce the first unit, and $Y_x$ is the hours required to produce the $x$th unit. The exponent $n$ is the learning index, and is defined as $n = \log(b)/\log(2)$ where $b$ is the learning rate, and $1 - b$ is the progress rate. For example, if the learning rate $b$ is 80%, the total effort required to produce a unit will be reduced by 20% for each doubling of output. The learning curve is a convex curve on a linear scale, but a straight line on a log–log scale.

### 2.3. Turnover

Compounding the learning curve effect in call centers is the issue of employee turnover. Employee turnover, or *wastage* as it sometime referred to in the literature, has long been a topic of quantitative analysis. Bartholomew (1971) and Bartholomew and Forbes (1979) are classic foundational works on the statistical analysis of employee turnover.

Basic turnover models are developed in Bartholomew and Forbes (1979). A common modeling objective is to use statistical techniques to analyze historical turnover patterns and use that to make predictions of future turnover. These models typically include some independent variables that attempt to segregate employees to homogeneous groups. Independent variables could include age, job level, gender, or job function. A common choice is to separate the workforce based on date of hire into cohorts; i.e., employee groups with roughly the same hire date (Bartholomew and Forbes 1979).

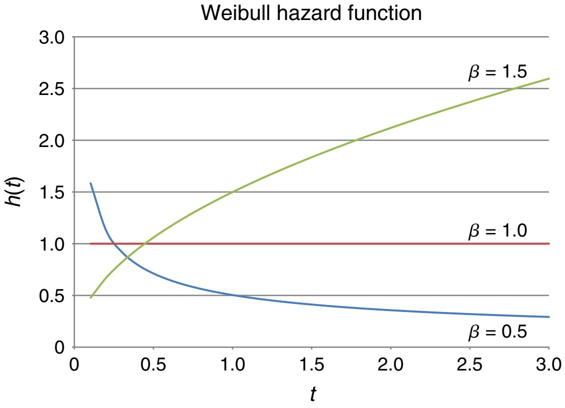
Statistical analysis of turnover is quite similar to reliability models and some of the terminology carries over. The statistical distribution of employee tenure can be described in multiple, equivalent formats. The survivor function $G(t)$ defines the probability that an individual survives (remains employed) for a time $t$. Its complement $F(t)$ is the distribution of the completed length of service. If we treat time as a continuous variable, then $f(t)$ represents the probability density. In a continuous time model the *hazard rate* (or *failure rate*) function can be defined as

$$r(t) = \frac{f(t)}{1 - F(t)}. \tag{2}$$

The hazard rate represents the time-instantaneous likelihood of failure as a function of time.

The three alternative representations (density function, survivor function, hazard function) are all mathematically equivalent; given one the other two can be derived. As in reliability analysis, it is the hazard rate formulation that is most illustrative of the underlying turnover dynamic. A constant hazard rate implies that an employee's likelihood of quitting remains constant throughout their employment. With an increasing failure rate the propensity to quit increases with length of service. With a decreasing failure rate the propensity to quit declines with length of service. Most empirical analyses of turnover suggest a decreasing failure rate (Bartholomew and Forbes 1979). Robbins (2007) analyzed data on 1,400 employee separations from a call center outsourcing firm and found a declining hazard rate.

**Figure 1.** Weibull Hazard Function for Various Values of $\beta$



When analyzing turnover, empirical data is often fit to a theoretical distribution. Many statistical distributions have been proposed to model wastage rates. Bartholomew and Forbes (1979) discuss the exponential model, which has a constant failure rate. While the constant failure rate simplifies analysis, it is not always a good fit to observed data. Bartholomew and Forbes (1979) also present a mixed exponential model, along with the lognormal model. Another distribution commonly applied in reliability analysis is the Weibull distribution, which is perhaps the most widely used distribution for lifetime analysis (Lawless 2003). The Weibull distribution has two parameters, $\beta$ and $\lambda$. The density function of the Weibull distribution is as follows:

$$f(t) = \lambda\beta(\lambda t)^{\beta-1}e^{-(\lambda t)^{\beta}} \quad t > 0. \tag{3}$$

The corresponding hazard rate function is

$$h(t) = \lambda\beta(\lambda t)^{\beta-1}. \tag{4}$$

The Weibull distribution provides a flexible model for reliability because with a $\beta$ greater than one the failure rate is increasing, with $\beta$ less than one the failure rate is decreasing, and when $\beta$ equals one the failure rate is constant. Note that when $\beta$ equals one, the Weibull distribution simplifies to the exponential distribution with a constant hazard rate of $\lambda$. Figure 1 shows the Weibull hazard function for various values of $\beta$.

Distribution fitting is challenging because we are generally forced to deal with censored data; we do not know the final length of service for workers still employed. Lawless (2003) provides an overview of multiple techniques that can be used to fit a distribution to censored data.

Turnover is a key issue in call center capacity management because call centers have a reputation, often times well deserved, for high levels of employee turnover. Each employee separation takes capacity from the system that must be replaced. Given the knowledge intensive nature of many call center environments, such as technical support, the lead time to replace that capacity, including training time, can be quite long. If the call center is subject to significant learning effects, the time required to replace the capacity includes recruiting, training, and on-the-job learning.

A number of papers in management and organization/human resources literature address the work environment and turnover problem in general, and several address call center-specific issues. Witt et al. (2004) examine issues of emotional exhaustion in a survey of 92 call center agents. Specifically, they examine the relationship between exhaustion and performance. Singh et al. (1994) survey 377 agents and find burnout levels among customer service agents are high relative to other high stress occupations. Singh (2000) examines how burnout affects productivity and quality. Cordes and Dougherty (1993) review the literature on job burnout. Holman (2002) surveys 577 call center agents to assess several measures of employee well-being. Cotton and Tuttle (1986) perform a meta-analysis of the papers available at the time that examined turnover across industries. Abelson and Baysinger (1984) develop a conceptual model of turnover and argue that the optimal level of turnover balances retention and turnover costs. Their argument is that some positive level of turnover is desirable, but they provide no hard data as to what level is optimal. Robbins (2007) analyzed data from a call center outsourcing firm over a two-year period and reported annualized attrition rates consistently above 20% and often above 35% per year.

### 2.4. Empirical Analysis of Call Centers

The differences between the behaviors of call center systems as predicted by queuing models, and the actual behavior observed in practice has become a popular topic of research. These differences are identified by empirical analysis of call center data. Brown et al. (2005) perform a detailed statistical analysis of data from an Israeli bank's call center. Their analysis focuses on key queuing model parameters such as the arrival, service, and abandonment processes. Their analysis exposes several substantial deviations from queuing model assumptions. These include highly variable and uncertain arrival times, talk times that are log-normally distributed (as opposed to the exponential distribution assumed in Erlang models), and patience hazard rates that are not constant (as assumed in Erlang A models).

Gans et al. (2010) analyze data from a four call center network supporting a U.S.-based banking operation. This analysis focuses on agent heterogeneity and learning effects. Their analysis at the individual call level demonstrates a learning effect that is strongly significant statistically, but explains only a small portion of the variation in service times in individual calls, the learning curve effect being dominated by the stochastic nature of service requests.

Robbins (2007) analyzes data from a call center outsourcer that provides IT support services. This analysis also finds a highly significant learning effect, but the analysis focuses on average monthly talk times per agent. At this higher level of granularity much of the stochastic variability of talk times is averaged out and the learning effect explains a greater portion of data variability. The learning curve effects in this analysis are summarized by Robbins and Harrison (2011) in the context of a new service launch problem.

### 2.5. Dynamic Routing

Several papers have examined scenarios similar to ours and investigated the impact of dynamic routing policies when servers/agents are heterogeneous. Armony (2005) examines service systems in which servers are separated into pools that operate at different speeds. They consider systems operating in the quality-efficiency driven (QED) regime, a mode of operation where jobs face a probability of queuing that is greater than zero, but less than one. In this scenario they prove that fastest server first (FSF) routing is asymptotically optimal. Armony and Ward (2010) investigate a similar scenario where agents are again separated into different pools that operate with different processing speed. They develop a model that attempts to minimize wait time given a fairness constraint; fairness defined as some equality between the workload of agents. While routing to the fastest agent will minimize wait time, it means that the fastest agents will be the busiest. Whereas fairness is often considered in the treatment of customers, see for example the review in Avi-Itzhak et al. (2004), Armony and Ward are among the first to consider fairness from the perspective of the servers. Their model implements a threshold routing policy that alternates between routing to the FSF when the number of customers in the system is large, and the slowest server first (SSF) when the number of customers in the system is small. Armony and Mandelbaum (2011) examine large-scale service systems where customers are homogeneous but servers are heterogeneous. Again, their model assumes that agents are separated into pools where service time is identical in each pool. They solve the staffing and routing problem jointly to minimize costs. Mandelbaum et al. (2012) examine alternate routing policies in the context of a hospital emergency department. Mehrotra et al. (2012) evaluate routing calls based on average agent handling time and first call resolution rate.

Although these papers address problems similar to the problem in this paper, there are several key distinctions. These papers all assume agents are assigned to pools, and that agents within the pools are homogeneous. Our model treats each agent individually effectively creating pools of size one. More importantly these previous papers all recognize that learning effects may lead to agent heterogeneity, none of them allow for active learning to change the relative levels of productivity. Furthermore, none of these papers allow for agent turnover. In effect, these papers are short-term focused, attempting to optimize performance over a narrow time frame. Our analysis is long term. We wish to see if alternative routing can balance the short term effects of faster processing with the long-term effect of faster learning to improve long-term performance.

## 3. Empirical Analysis

The models in this paper are motivated by observations that real-world call centers differ from call center models in some significant ways. Our observations are based in large part on empirical analysis of data from a call center outsourcing firm that effectively gives us a view of more than 20 IT help desk operations. These data were first analyzed by Robbins (2007) and later summarized by Robbins and Harrison (2011), and some of the graphs from those papers are reproduced here. For comparison we also consider empirical results reported in Gans et al. (2010) and Brown et al. (2005).

Although many deviations from standard call center models are noted in these papers, the issues most relevant for this analysis are the following:

- Agents are not homogeneous. Productivity varies from agent to agent, and an agent's productivity varies over time as the agent learns.
- Agent turnover is significant. Agents quit frequently and the call center is in a constant state of renewal, training new agents to replace departed agents.
- The workload often changes faster than the level of staffing. The call center workload varies considerably from day to day and hour to hour. The agents' desire to work long shifts on a regular basis means that call centers alternate between periods of over- and under capacity.

### 3.1. Agent Heterogeneity

Gans et al. (2010) analyze agent learning effect at a detailed call by call level. They fit a basic learning curve model as in (1) to a portion of their data. They find that learning curve effects are highly significant with a $p$-value less than 0.0001. The model has a coefficient of determination of $R^2 = 0.005$, indicating that other factors account for much of the call to call variability in service time. The model yielded a progress rate of 8.3%. Analysis of individual agent learning effects demonstrated variability in learning rate from agent to agent.

Data from our call center outsourcer support the notion that agents are heterogeneous and learn over time. The data here are more aggregate, consisting of the monthly talk time averages reported on each agent's monthly evaluation scorecards over a 23-month period. We analyzed the average monthly talk times for 53 agents all providing technical support to the same corporate customer. Employment durations ranged from 3 to 19 months. Figure 2 shows the reported monthly average talk time for each agent as a function of his or her month of active service.

Each point on this graph (Figure 2) represents an individual agent employed for one month. From this data we can make two key observations. First, agents exhibit a wide variation in average talk time. Month 1 average talk times range from about six minutes to more than 20 minutes. Second, individual agent talk time improves with experience. Talk time declines from an average of 12.7 minutes in the first month, to 8.3 minutes in the sixth month. Comparing the average talk time for all agents in their $n$th month of service to the average of the $n-1$th month of service and using a standard $T$-test found the month-to-month reduction is statistically significant at the 0.1 level through the first five months of service. The data indicate that agents experience productivity improvements that can be fit to a standard learning curve model. Figure 3 shows data on a log–log scale with a linear trend line fit to the data.

As in the previously discussed example (Gans et al. 2010), the learning effect is highly significant with a $p$-value of less than $2 \times 10^{-31}$. The $R^2$ value of 27% is much higher than the results discussed , presumably because this analysis deals with aggregated data. The progress rate for this data set is estimated at 14.0%, also much larger than the 8.3% reported previously. Although it is uncertain why the rate of progress is more substantial, one can hypothesize that the nature of the work, solving user IT problems, may lend itself to a more significant learning effect.

### 3.2. Agent Turnover

Call centers have a reputation, well supported by the data, for high level of employee turnover. This claim is supported by the empirical data we have available. Figure 4 shows the month–by-month annualized turnover rate over an approximately 28-month period for our call center outsourcing firm.

**Figure 2.** Monthly Average Talk Time

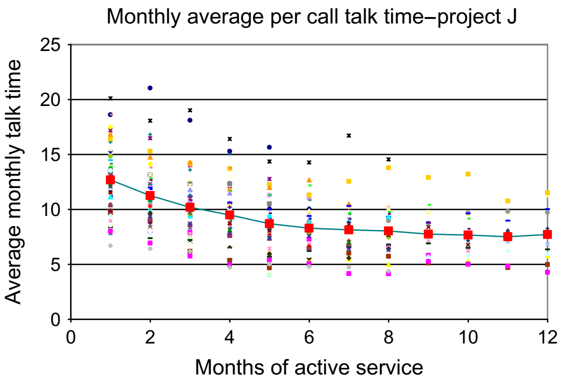


**Figure 3.** Talk Time by Length of Service

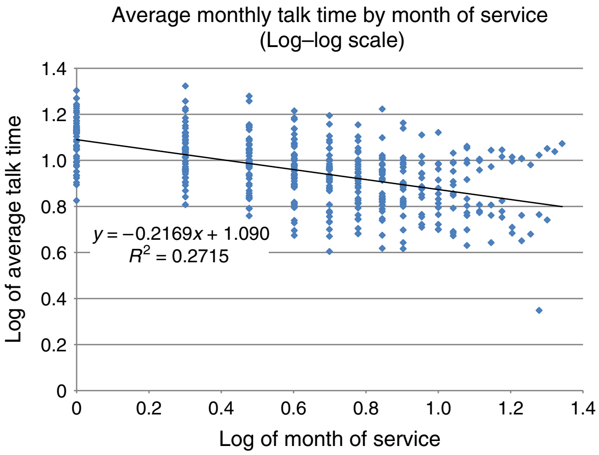
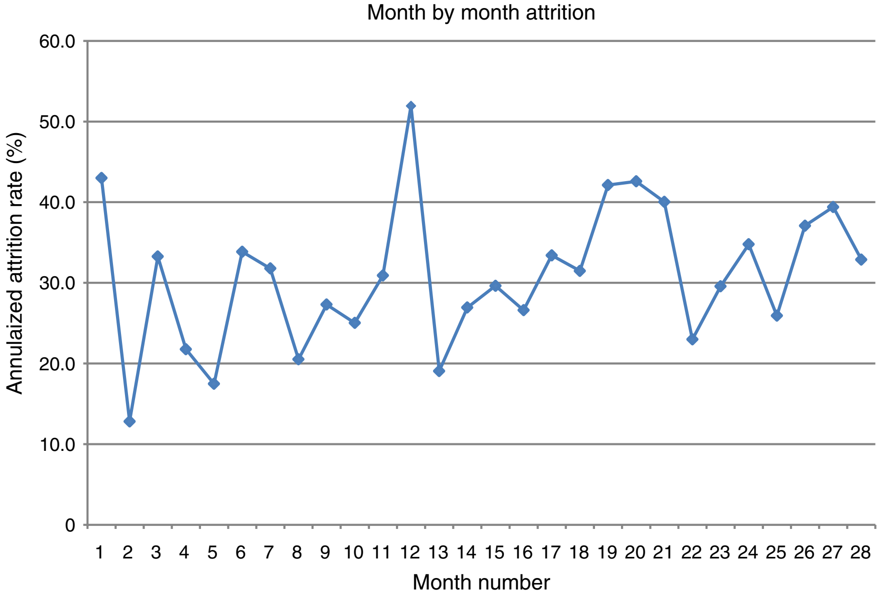


**Figure 4.** Annualized Attrition Rates by Month



We see that turnover is high and that it varies significantly from month to month, ranging between 13% and 52%, averaging about 30%. While we need to be cautious and not overgeneralize the results from this single company, it does suggest that attrition is a significant management issue.

We performed a more detailed analysis of the turnover reviewing termination data on the 1,400 terminations that occurred over a period just over five years. Fitting the data to a Weibull distribution yields a shape parameter equal to 0.918 and a scale parameter equal to 0.0309 that results in a declining hazard rate. Whereas about 15% of the new hires wash out, quitting within the first three months, those who remain employed become on average less likely to quit with each month of service. The result of this turnover is an agent base dominated by relatively inexperienced employees. Figure 5 shows a histogram of agent tenure at this call center outsourcing firm. The high level of turnover has created an environment heavily skewed toward inexperienced agents.

### 3.3. Workload Variability

The workload in a typical call center is quite variable exhibiting strong day-of-week and time-of-day seasonality. Daily call volume often declines over the course of a week. Over the course of a day, call centers often experience wide swings in volume with a dual peak pattern being common; a large peak of volume in the late morning, a drop off for lunch, and a second peak in the early afternoon. This phenomenon is illustrated clearly in Figure 6, which shows the average number of calls per half hour for a corporate help desk. Similar patterns are seen for other call centers in papers by Robbins (2007), Brown et al. (2005), and Gans et al. (2003).

**Figure 5.** Histogram of Agent Tenure

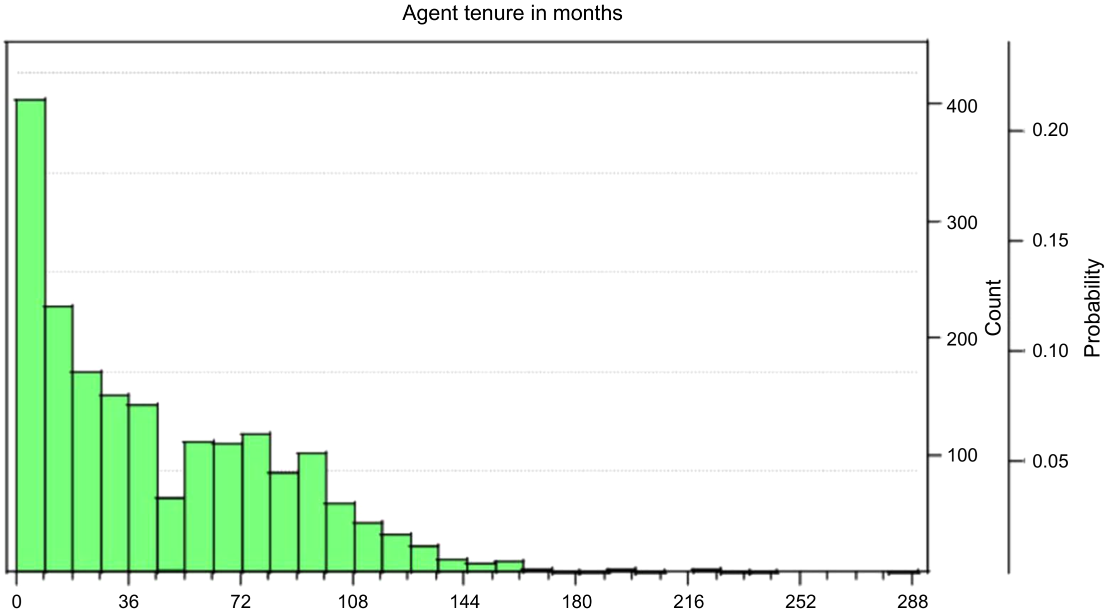
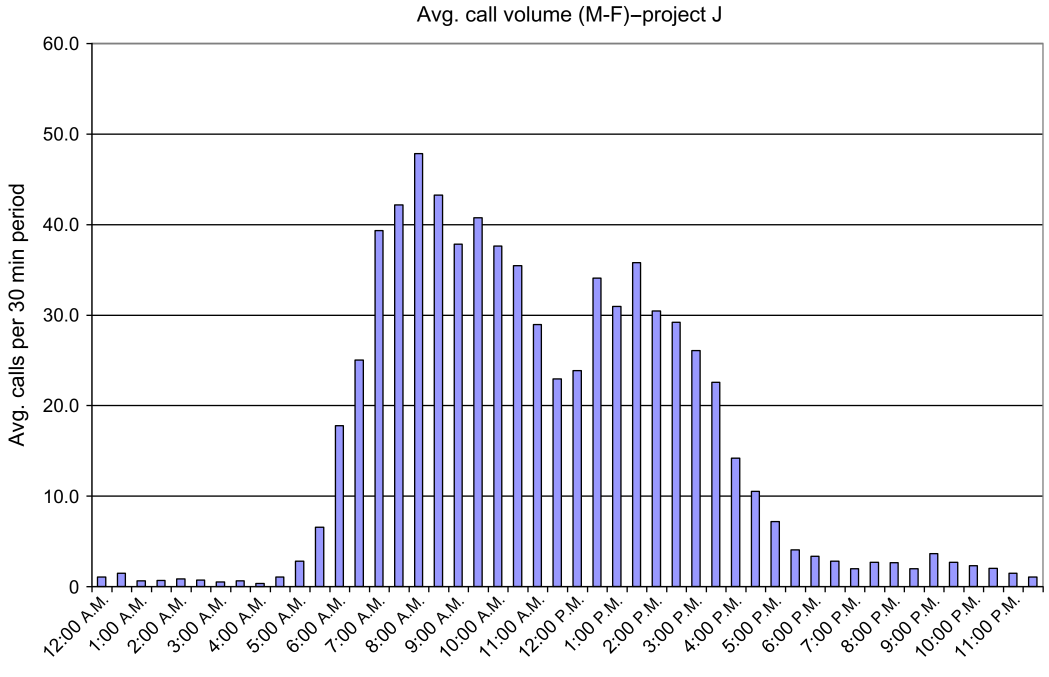


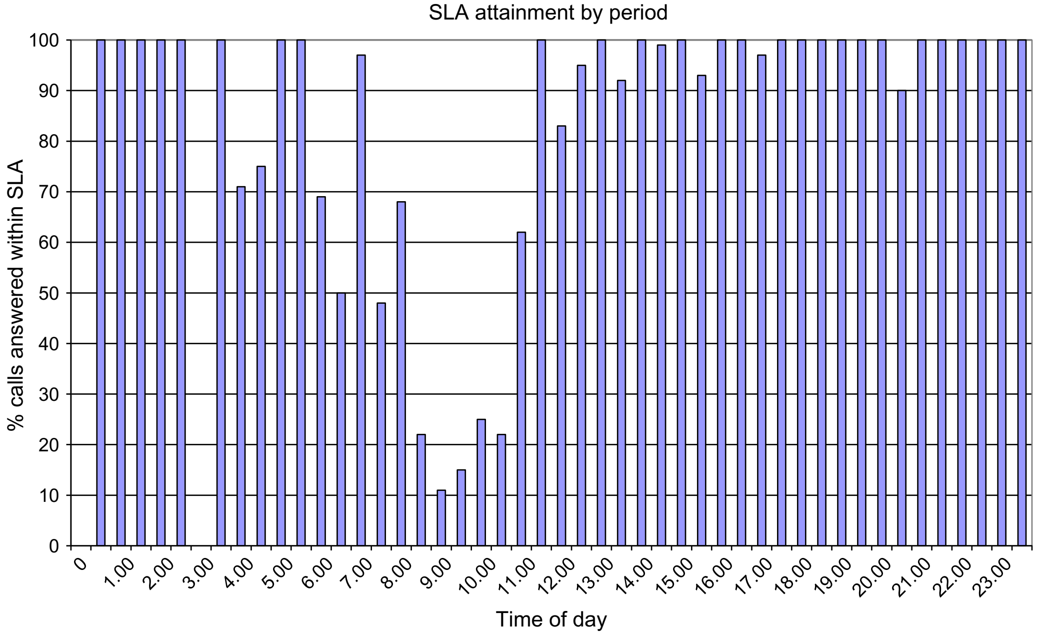
**Figure 6.** Average Arrivals per Half Hour



One of the key challenges associated with scheduling a call center is matching supply with demand. Given the desire of most agents to work shifts that are long relative to demand peaks, hourly mismatches are difficult to avoid. Given the desire of agents to work multiple days, daily mismatches are also difficult to avoid. Much of the scheduling literature deals with finding the best way to minimize this mismatch, but even in the best of cases the call center often switches between periods of over- and under capacity. This effect is illustrated in Figure 7, which shows the achievement of the specified service level agreement (SLA), referred to as SLA attainment, for the same call center illustrated in Figure 6.

The graph shows that for much of the day this call center is overstaffed; utilization is low and 100% of the calls are answered within the SLA. During busy periods the call center was understaffed; utilization is high and service level attainment was very poor. In this particular environment, the call center is responsible for meeting

**Figure 7.** SLA Attainment by 30 Minute Period



the service level over an extended period of time, a global service level agreement. Call center managers balance these periods of over- and understaffing to meet the service level at the end of the measurement period, typically one month.

Note that even in cases where the call center seeks to achieve the service level target in every 30 minute period, it is most often the case that call volume will change more rapidly than the staffing and the call center will alternate between under- and overstaffing. Robbins (2007) examines this effect and measures the excess staffing percentage, showing how this metric decreases with increased scheduling flexibility. We can therefore reasonably assume that in many real-world applications the call center shifts between periods of high and low utilization.

# 4. Call Center Simulation

## 4.1. Call Center Model

In this section we present a revised model of a call center. We are interested in analyzing long-term average performance of the call center. Because call volume typically changes faster than staffing levels, we model call arrivals according to a nonhomogeneous Poisson process with arrival rate $\lambda(t)$. The instantaneous arrival rate $\lambda(t)$ varies in a bimodal pattern over the course of an eight hour shift as implied by Figure 7. This pattern is illustrated in Figure 8. The specific pattern of arrivals is determined by three parameters: offered utilization in the high traffic mode, offered utilization in the low traffic mode, and the proportion high traffic parameter.

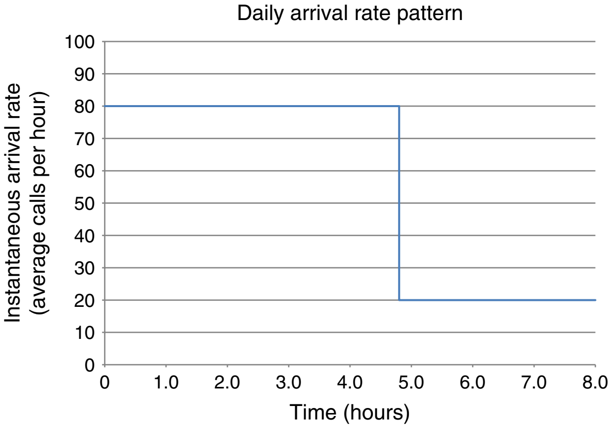
The nominal time required to service a call is an exponentially distributed random variable with mean $\mu^{-1}$; the realized service time will depend on the agent that services the call and that agent's level of experience. Arriving calls are routed to the agent identified by the routing algorithm. If all agents are busy, the call is place in a FCFS queue. Callers who join the queue have a patience time that follows an exponential distribution. If wait time exceeds their patience time, the caller will abandon the queue, i.e., hang up.

Agent efficiency is a function of experience, with efficiency improving based on a standard learning curve model as per Equation (1). For example, assume that $i$th call received has a nominal service time of $S_i$. Further assume that this call is ultimately routed to agent $j$ and that it represents the $k$th call handled by that agent where the agent has a learning rate of $b$. The realized service time for this call will be

$$\widehat{S}_i^{jk} = S_i k^{\log b/\log 2}. \tag{5}$$

Routing calls to experienced agents will result in a shorter service time, with the difference depending on the experience of the agent and the rate of learning. (Our model assumes that all agents learn with the same rate.)

**Figure 8.** Simulated Arrival Rate



Agents, however, do not remain employed forever, rather they quit after some length of service time. The time an agent works before quitting is assumed to follow a Weibull distribution.

## 4.2. Routing Policy

In this analysis we evaluate several different routing policies; namely:

- *Longest Idle Routing* (*LIR*): find the agent who has been idle for the longest time. This is a standard, default routing scheme.
- *Fastest Server First* (*FSF*): find the idle agent with the largest number of calls completed, equivalent to Most Experienced Routing.
- *Slower Server First* (*SSF*): find the idle agent with the smallest number of calls completed, equivalent to Least Experienced Routing.
- *Experienced Based Routing*: our modified routing policy switches between Fastest Server First (Most Experienced) and Slowest Server First (Least Experienced) routing based on call center conditions.

In relative terms, least experienced routing is used when the call center is in low volume mode, and most experienced routing when the call center is in high volume mode.

## 4.3. Experimental Design

To evaluate the performance of the experience-based switch routing approach we conduct a series of designed experiments. Our objective is to test experience-based routing over a wide range of parameter values. Based on the assumptions for our call center, we define the set of nine experimental factors listed in Table 1. We set the low and high values of the parameters so that our experimental design covers a wide range of reasonable call center scenarios.

The average arrival rate in the simulation is a quantity derived from other experimental factors by

$$\bar{\lambda} = \rho N \mu. \tag{6}$$

Given the relatively large number of experimental factors, a well-designed experimental approach is required to efficiently evaluate the experimental region. A common approach to designing computer simulation experiments

**Table 1.** Experimental Factors

| Factor | Symbol | Low | High |
|---|---|---|---|
| 1 Number of agents | $N$ | 10 | 100 |
| 2 Learning rate | $b$ | 70% | 100% |
| 3 Offered utilization—High | $\rho_H$ | 80% | 110% |
| 4 Offered utilization—Low | $\rho_L$ | 20% | 50% |
| 5 Proportion high utilization | $p_H$ | 30% | 70% |
| 6 Nominal talk time (mins) | $S$ | 2 | 20 |
| 7 Average time to abandon (secs) | $a$ | 100 | 1,000 |
| 8 Quit distribution scale (days) | $\alpha$ | 90 | 400 |
| 9 Quit distribution shape | $\beta$ | 0.8 | 1.1 |

is to employ either a full or fractional factorial design (Law 2007). However, the factorial model only evaluates corner points of the experimental region and implicitly assumes that responses are linear in the design space. Given the nonlinear relationship of performance metrics, we chose instead to implement a space filling design based on Latin hypercube sampling as discussed in Santner et al. (2003). Given a set of $d$ experimental factors and a desired sample of $n$ points, the experimental region is divided into $n^d$ cells. A sample of $n$ cells is selected in such a way that the centers of these cells are uniformly spread when projected onto each of the $d$ axes of the design space. We chose our design point as the center of each selected cell. This experimental design allows us to select an arbitrary number of points for any experiment.

### 4.4. Simulation Model

Our call center model is evaluated using a straightforward discrete event simulation model. The purpose of the model is to predict the long term, steady state behavior of the queuing system. The model generates random numbers using a combined multiple recursive generator (CMRG) based on the MRG32k3a generator described in L'Ecuyer (1999). Common random numbers are used across design points to reduce output variance. To reduce any start up bias, we use an extensive warm-up period of 1,000,000 calls, after which all statistics are reset. This extensive warm-up period allows the call center, as well as the employee experience base, to reach steady state. The model is then run for an evaluation period of 100,000 calls and summary statistics are collected. For each design point we repeat this process for 15 replications and report the average value across replications. Our primary analysis is based on an experiment with 250 design points.

Our simulation model creates the preselected number of design points then runs each point three times, once with the switch routing policy, once with the baseline longest idle policy, and once with the most experienced policy. Data are collected for each run for a number of performance metrics. The model was coded in Visual Basic and the experiment was run on a standard desktop computer. To complete the entire experiment the model ran for approximately 90 hours.

The specific process for each replication is as follows. The input factors are chosen based on the experimental design. An agent pool of size $n$ is initialized by randomly setting a current length of service and time to quit for each. Once the agent pool is initialized the call center operation simulation begins. The initial arrival rate is calculated based on the specified nominal talk time, number of agents, and offered utilization rate according to Equation (6). We begin the simulation in the high utilization mode and switch to low utilization after the high utilization percentage of an eight hour shift. We switch back and forth between high and low utilization in each eight hour period for the duration of the simulation.

Because realized talk times will vary based on dynamic agent experience levels, we must calculate an appropriate mean for the exponentially distributed nominal talk time. To perform this calculation we first estimate the total number of calls an agent will handle over their lifetime by calculating the average number of calls received during the average tenure of an agent, divided by the number of agents. We then calculate the productivity scaling factor for an agent with that level of experience using Equation (5). The average of the exponential distribution used to generate nominal talk times is then set based on that scaling factor. Also generated for each call is a random time to abandon. If the time to abandon passes while the call is still in queue the call is abandoned and removed from the queue. Otherwise the abandonment event is cancelled.

When a call arrives and agents are available it is routed based on the current routing policy. If all agents are busy the call is placed into the queue. Once the call has been assigned to an agent, the realized talk time is calculated based on the average talk time and the agent's experience. The agent is committed for the realized talk time.

When the call completes, the agent processes the next call from the queue, or if no calls are queued, becomes idle. Over the course of the simulation we collect a number of performance statistics. We collect the proportion of calls waiting, the average speed to answer, the abandonment rate, the TSF (based on a 30-second wait period), and the total time in system.

## 5. Evaluating the EBR Policy

### 5.1. Simulation Experiment Overview

To evaluate the general implications of this policy we run an experiment evaluating the experience-based routing (EBR) policy against the standard baseline of longest idle routing (LIR). In this experiment we test 250 design points, evaluating each design point independently under EBR and LIR policies. At each design point we use 1,000,000 warm-up calls and 100,000 process calls to calculate performance metrics that we average across 15 replications.

**Table 2.** Summary Performance Metrics (EBR vs. LIR)

| | EBR average | LIR average | Average change | Proportion EBR better (%) |
|---|---|---|---|---|
| | | Experience based vs. longest idle routing | | |
| Probability of wait (%) | 37.9 | 40.5 | −2.6 | 95.6 |
| Abandonment rate (%) | 7.7 | 8.3 | −0.6 | 96.4 |
| Average speed to answer | 39.0 | 42.2 | −3.1 | 96.4 |
| TSF (%) | 68.9 | 66.5 | 2.3 | 94.8 |
| Time in system | 13.55 | 13.55 | 0.00 | 46.0 |

| | EBR average | LIR average | Average change | DPs where EBR outperforms LIR (%) |
|---|---|---|---|---|
| | | Experience based vs. longest idle routing | | |
| Probability of wait (%) | 37.9 | 40.5 | −2.6 | 95.6 |
| Abandonment rate (%) | 7.7 | 8.3 | −0.6 | 96.4 |
| Average speed to answer | 39.0 | 42.2 | −3.1 | 96.4 |
| TSF (%) | 68.9 | 66.5 | 2.3 | 94.8 |
| Time in system | 13.55 | 13.55 | 0.00 | 46.0 |

### 5.2. Simulation Results

Our preliminary results indicate that EBR outperforms LIR across most performance measures under most conditions. Summary results are presented in Table 2. Table 2 lists the average value of key metrics under the EBR and LIR policies along with the average change. It also identifies the proportion of design points where EBR outperforms LIR.

These summary results indicate that EBR outperforms standard routing significantly on speed of answer and abandonment metrics in the vast majority of our test points. Total time in system (TIS) is effectively unchanged even though on average a higher percentage of calls are serviced under EBR.

Several scatter plots of the results are shown in Figure 9. In each graph, a 45-degree line from the lower left to the upper right represents equivalent outcomes. These graphs confirm that EBR improves speed of answer and abandonment metrics. The graph also confirms that TIS is virtually unchanged due to our modified routing policies.

EBR effectively allows us to improve the performance of our call center, servicing a higher percentage of calls, answering them faster but servicing them slower, those last two effects balancing out so that TIS is effectively unchanged. The speed of answer improvement can be significant and this effect is more clearly seen in Figure 10, which presents a histogram of the time- and percentage-based change in the average speed to answer (ASA) metric. On average, calls are answered in 3.1 seconds, or 11.7% faster. The improvement in ASA exceeds 10% in 49.2% of our test points.

Our initial analysis reveals that EBR can be a useful tool to improve system performance. We now focus on identifying the conditions where EBR has the most potential.

**5.2.1. Drivers of Routing Impact.** To get a basic idea of which of our test factors have the greatest impact on the improvement of call center performance we analyze the relationship between our experimental factors and run a regression analysis based on first order effects. The dependent variable in our model is the percentage improvement in ASA; the independent variables are the experimental factors normalized onto a scale of $-1$ to 1. Normalizing the variables make it clear which variables have the largest impact. Table 3 shows the results of the regression analysis. The $p$-values shown in bold indicate the variables significant at the .05 level.

The resulting model is highly significant with a very small $p$-value on the order of $10^{-22}$. The $R$-squared value of 39.5% is also reasonably high. The most influential variables, based on the $p$-values and coefficient magnitudes, are the three utilization parameters. As an alternative way to analyze this data, we divide it into quartiles and calculate the average value of the improvement and the average value of each factor in each quartile. These results are shown in Table 4.

Although EBR improves speed of answer performance in almost all cases, the level of improvement does vary considerably. Based on this data, a picture begins to emerge of the conditions under which EBR has the most effect. Not surprisingly the improvement is the most significant when call center conditions allow for the most modification of call routing, particularly to lower experienced agents. Under conditions with lower levels

**Figure 9.** Comparison of Key System Metrics (EBR vs. LIR)

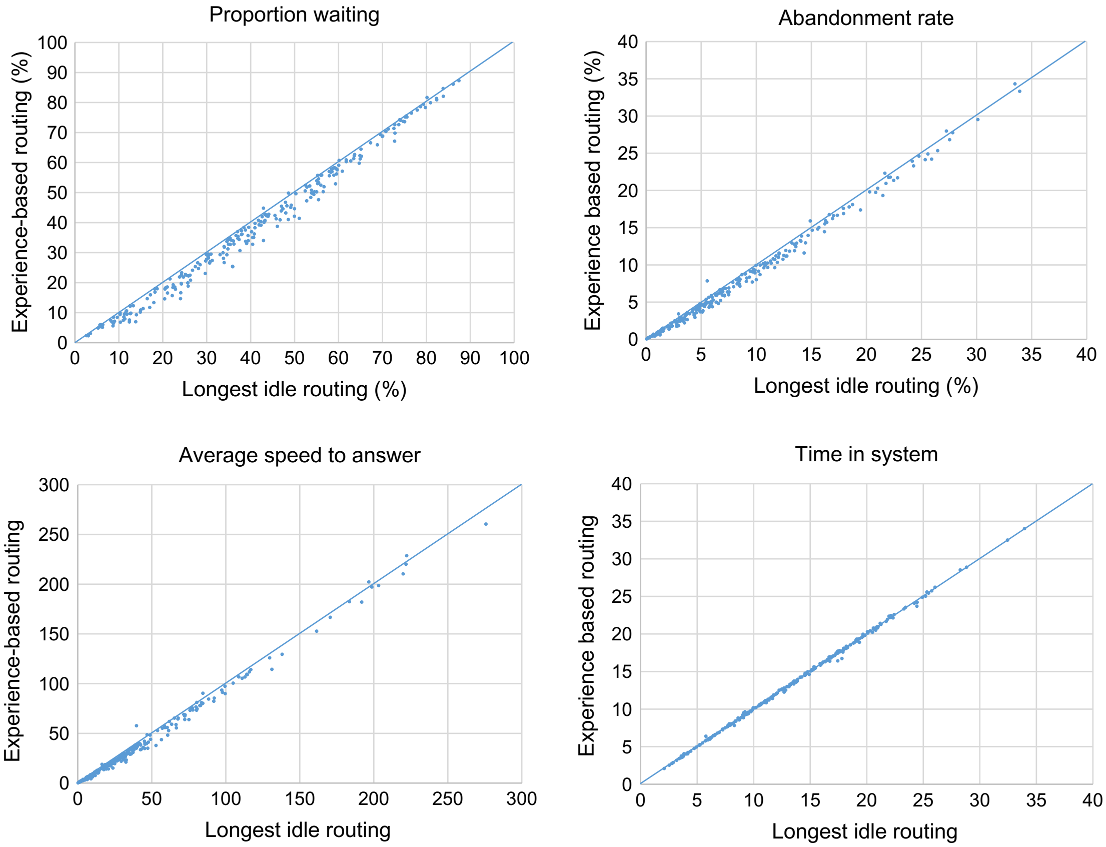
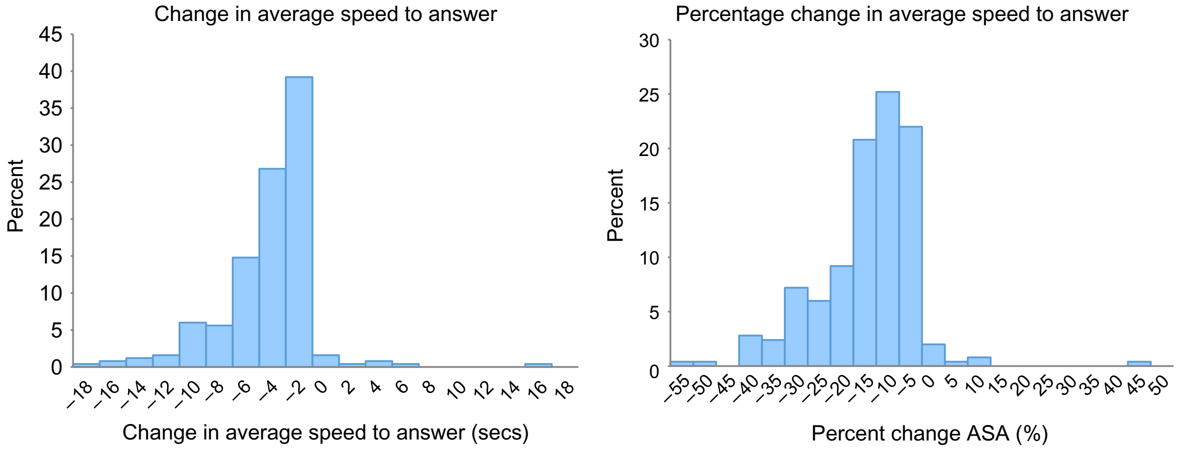


**Figure 10.** Histogram of Change and Percentage Change of Average Speed to Answer



of slow period utilization that last longer and significant learning effects (low learning rates) agent learning can be accelerated more significantly since more opportunities exist to route calls to lower experienced agents. A low average utilization means more agents will be available to pick from when a new call arrives. Also, in environments with higher turnover there will be more inexperienced agents who can benefit from learning during slow periods. Furthermore, when busy period utilization is lower there are more opportunities to route calls to more experienced agents. Stated another way, in very high utilization environments the number of agents available when a call arrives will be small so the opportunity to make a significant improvement by picking the fastest agent is reduced.

**Table 3.** Regression Analysis Results

Regression analysis

| | | | |
|---|---|---|---|
| $R^2$ | 0.395 | | |
| Adjusted $R^2$ | 0.373 | $n$ | 250 |
| $R$ | 0.629 | $k$ | 9 |
| Std. error | 0.127 | Dep. var. | **% Chg ASA** |

ANOVA table

| Source | SS | df | MS | $F$ | $p$-value |
|---|---|---|---|---|---|
| Regression | 2.5345 | 9 | 0.2816 | 17.44 | **4.24E−22** |
| Residual | 3.8760 | 240 | 0.0161 | | |
| Total | 6.4105 | 249 | | | |

Regression output

| | | | | | Confidence interval | |
|---|---|---|---|---|---|---|
| Variables | Coefficients | Std. error | $t(\text{df} = 240)$ | $p$-value | 95% lower | 95% upper |
| Intercept | −0.1510 | 0.0080 | −18.784 | **4.97E−49** | −0.1668 | −0.1351 |
| No. of Agents | −0.0387 | 0.0142 | −2.725 | **0.0069** | −0.0667 | −0.0107 |
| Low Utilization Level | 0.1089 | 0.0144 | 7.590 | **7.04E−13** | 0.0807 | 0.1372 |
| Talk Time | −0.0627 | 0.0141 | −4.445 | **1.34E−05** | −0.0905 | −0.0349 |
| Time to Abandon | −0.0090 | 0.0140 | −0.644 | 0.5200 | −0.0366 | 0.0185 |
| Learning Rate | 0.0545 | 0.0140 | 3.882 | **0.0001** | 0.0268 | 0.0821 |
| Average Days to Quit | 0.0395 | 0.0141 | 2.799 | **0.0055** | 0.0117 | 0.0672 |
| Quit Distribution Shape Parameter | 0.0301 | 0.0140 | 2.153 | 0.0324 | 0.0026 | 0.0576 |
| High Utilization Level | 0.0751 | 0.0141 | 5.339 | **2.16E−07** | 0.0474 | 0.1028 |
| High Utilization Proportion | 0.0831 | 0.0142 | 5.842 | **1.67E−08** | 0.0551 | 0.1111 |

**Table 4.** Quartile Averages

| Quartile | No. of agents | Low utilization level (%) | High utilization level (%) | High utilization proportion (%) | Learning rate (%) | Talk time | Time to abandon | Average days to quit | Quit distribution shape parameter | Average change in ASA | Average % change in ASA (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $Q1$ | 61.48 | 29.9 | 90.7 | 46.4 | 82.4 | 11.57 | 580.09 | 213.73 | 0.94 | −4.79 | −26.2 |
| $Q2$ | 57.77 | 33.7 | 95.0 | 49.4 | 85.3 | 11.54 | 491.01 | 250.68 | 0.96 | −3.68 | −12.7 |
| $Q3$ | 51.71 | 37.2 | 96.4 | 52.6 | 87.0 | 10.66 | 589.89 | 260.56 | 0.94 | −3.11 | −7.0 |
| $Q4$ | 49.02 | 38.9 | 98.0 | 51.6 | 85.2 | 10.22 | 538.71 | 255.36 | 0.96 | −0.94 | −1.0 |

### 5.3. Fastest Server First Routing

The data seems to indicate that improvement comes from both accelerating learning during slow periods and decreasing talk time during busy periods. Thus far it is unclear from our analysis if one effect is more important than the other. To evaluate this issue we run a second experiment that compares experience-based routing against fastest server first (FSF) routing.

To conduct this test we rerun the experiment with identical conditions, using the same 250 design points. The only change we make is implementing FSF routing instead of EBR. Summary results of this experiment are provided in Table 5 and Figure 11. Table 6 provides a comparison of all three routing methods.

Somewhat counterintuitively, EBR outperforms FSF in speed of answer metrics by a wider margin than it outperformed LIR. FSF has the worst speed of answer metrics and the highest abandonment rate. It does however have the best TIS metrics.

The implication is that the routing of calls to the least experienced agent during the slow periods has a very significant impact on call center performance under our test conditions. Under FSF, low experienced agents will handle very few if any calls during slow periods, so their learning will be delayed. During busy periods routing to the fastest agent will optimize performance given the current level of experience in the agent pool likely leading to the time in system improvement.

**Table 5.** Summary Performance Metrics (FSF vs. EBR)

| | EBR average | FSF average | Average change | Proportion EBR better (%) |
|---|---|---|---|---|
| | Experience based vs. fastest server first routing | | | |
| Probability of wait (%) | 37.9 | 43.2 | −5.3 | 92.0 |
| Abandonment rate (%) | 7.7 | 9.9 | −2.2 | 98.8 |
| Average speed to answer | 39.0 | 50.6 | −11.5 | 98.8 |
| TSF (%) | 68.9 | 62.9 | 5.9 | 92.8 |
| Time in system | 13.55 | 13.11 | 0.44 | 12.8 |
| | EBR average | FSF average | Average change | DPs where EBR outperforms LIR (%) |
| | Experience based vs. fastest server first routing | | | |
| Probability of wait (%) | 37.9 | 43.2 | −5.3 | 92.0 |
| Abandonment rate (%) | 7.7 | 9.9 | −2.2 | 98.8 |
| Average speed to answer | 39.0 | 50.6 | −11.5 | 98.8 |
| TSF (%) | 68.9 | 62.9 | 5.9 | 92.8 |
| Time in system | 13.55 | 13.11 | 0.44 | 12.8 |

**Figure 11.** Comparison of Key System Metrics (FSF vs. EBR)

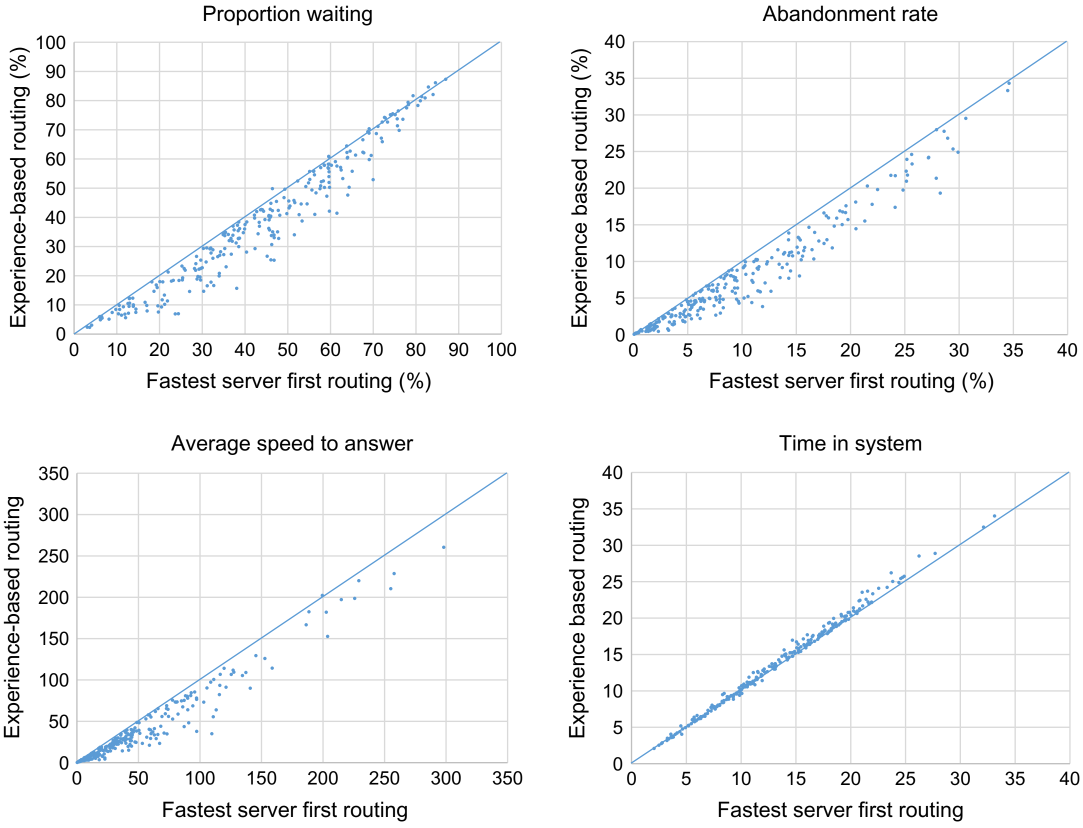


**Table 6.** Comparison of Three Routing Methods

| | EBR average | LIR average | FSF average |
|---|---|---|---|
| Probability of wait (%) | 37.9 | 40.5 | 43.2 |
| Abandonment rate (%) | 7.7 | 8.3 | 9.9 |
| Average speed to answer | 39.0 | 42.2 | 50.6 |
| TSF (%) | 68.9 | 66.5 | 62.9 |
| Time in system | 13.55 | 13.55 | 13.11 |

## 6. Conclusions and Future Research

In this paper we have examined several conditions that deviate from the assumptions used in standard call center models. Our modeling assumptions are based on empirical observations from real-world call centers. Most notably our model recognizes that call center services are delivered by human agents that are inherently heterogeneous. We postulate that much of that heterogeneity is driven by improvements that comes with experience; improvements can be modeled using standard learning curve models. It is the combination of learning curve effects and agent turnover that creates the heterogeneity in our agent base. Furthermore we assume, again based on empirical observations, that call center traffic varies more rapidly than call center staffing leading to an environment where the call center switches between periods of high and low relative intensity. It is this bimodal intensity that enables our routing policy—a policy that routes calls to agents, not based on which agent has been idle the longest, but on the relative experience level of agents. When call volumes are relatively low, we route the least experienced agent to facilitate learning. When call volumes are relatively high, we route the most experienced agent to facilitate efficiency.

Our analysis indicates that this routing policy improves system performance and customer service over a very wide range of conditions. More specifically, we find that when agent heterogeneity becomes large, and volumes are low enough to allow for a significant level of discretion on how calls are routed, system performance can improve quite significantly. Our approach accelerates agent learning and therefore improves the quality of service delivered to customers. Our analysis shows that it is primarily the routing to low experience agents during slow periods that drives performance improvements as our EBR policy significantly outperforms FSF routing. The EBR policy is also a fairer policy than FSF as the workload is more evenly balanced between less and more experienced agents.

Our model has significant implications for call center managers. In environments where learning is significant, managers should seek to maximize new agent utilization during low intensity periods, routing to the least experienced agent first. This will accelerate their learning in low-risk environments where virtually all calls are answered with no wait. During high intensity periods calls should be handled as efficiently as possible by routing to the fastest agent. Our analysis indicates that this combination can significantly reduce abandonment and decrease wait time with little change in total wait and service time.

In the future this analysis can be expanded on in several ways. Our analysis assumes that agents learn at the same rate. Follow-up analysis could examine the impact of differential learning rates. Also, we route calls based on experience as a proxy for productivity. If we were to allow for differential learning rates, we might choose to route calls based on an agent's actual average talk time performance rather than the number of calls that agent has handled. Further analysis could also consider the gaps that might occur between when an agent quits and when they are replaced. Shortening this gap would clearly have a positive effect on system performance. Finally, our model assumes that the routing policy has no impact on turnover. It is, however, conceivable that faster learning could have a positive impact on job tenure. It could be the case that agents that feel more productive are less likely to quit. This provides for an interesting line of empirical as well as model-based research.

### References

Abelson MA, Baysinger BD (1984) Optimal and dysfunctional turnover: Toward an organizational level model. *Acad. Management Rev.* 9(2):331–341.

Argote L (1999) *Organizational Learning: Creating, Retaining and Transferring Knowledge* (Springer, New York).

Argote L, Epple D (1990) Learning curves in manufacturing. *Science* 247(4945):920–924.

Argote L, Beckman SL, Epple D (1990) The persistence and transfer of learning in industrial settings. *Management Sci.* 36(2):140–154.

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

Armony M, Mandelbaum A (2011) Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Oper. Res.* 59(1):50–65.

Armony M, Ward AR (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Oper. Res.* 58(3):624–637.

Asher H (1956) *Cost-Quantity Relationships in the Airframe Industry* (Rand Corporation, Santa Monica, CA).

Avi-Itzhak B, Levy H, Raz D (2004) Quantifying fairness in queueing systems: Principles and applications. Technical report, Rutgers Center for Operations Research, Rutgers University, Piscataway, NJ.

Bartholomew DJ (1971) The statistical approach to manpower planning. *The Statistician* 20(1):3–26.

Bartholomew DJ, Forbes AF (1979) *Statistical Techniques for Manpower Planning* (Wiley, Chichester, UK).

Brown L, Gans N, Mandelbaum A, Sakov A, Halpeng S, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100(469):36–50.

Charland PJ, Robbins T, Rodriguez E, Nifong WL, Chitwood RW Jr (2011) Learning curve analysis of mitral valve repair using telemanipulative technology. *J. Thoracic Cardiovascular Surgery* 142(2):404–410.

Cordes CL, Dougherty TM (1993) A review and integration of research on job burnout. *Acad. Management Rev.* 18(4):621–656.

Cotton JL, Tuttle JM (1986) Employee turnover: A meta analysis and review with implications for research. *Acad. Management Rev.* 11(1):55–70.

Ebbinghaus H (1964) *Memory: A Contribution to Experimental Psychology* (Dover Publications, New York).

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

Gans N, Liu N, Mandelbaum A, Shen H, Ye H (2010) Service times in call centers: Agent heterogeneity and learning with some operational consequences. Berger JO, Cai TT, Johnstone IM, eds. *A Festschrift for Lawrence D. Brown*, Vol. 6 (IMS Collections, Beachwood, OH), 99–123.

Green L, Kolesar P (1991) The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* 37(1):84–97.

Green LV, Kolesar PJ, Whitt W (2005) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–39.

Green LV, Kolesar PJ (1997) The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates. *Management Sci.* 43(1):80–87.

Green LV, Kolesar PJ, Soares J (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. *Oper. Res.* 49(4):549–564.

Holman D (2002) Employee wellbeing in call centers. *Human Resource Management J.* 12(4):35–50.

Jennings OB, Mandelbaum A, Massey WA, Whitt W(1996) Server staffing to meet time-varying demand. *Management Sci.* 42(10): 1383–1394.

Kaul S, Shah NL, Menon M (2006) Learning curve using robotic surgery. *Current Urology Reports* 7(2):125–129.

Kim Y, Krishnan R, Argote L (2012) The learning curve of IT knowledge workers in a computing call center. *Inform. Systems Res.* 23(3, Pt. 2):887–902.

L'Ecuyer P (1999) Good parameters and implementations for combined multiple recursive random number generators. *Oper. Res.* 47(1): 159–164.

Law AM (2007) *Simulation Modeling and Analysis* (McGraw-Hill, Boston).

Lawless JF (2003) *Statistical Models and Methods for Lifetime Data* (Wiley-Interscience, Hoboken, NJ).

Liu Y, Whitt W (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6):1551–1564.

Mandelbaum A, Zeltyn S (2007) Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. Spath D, Fähnrich K-P, eds. *Advances in Services Innovations*, Part I (Springer-Verlag, Berlin), 17–45.

Mandelbaum A, Momčilović P, Tseytlin Y (2012) On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Sci.* 58(7):1273–1291.

Mehrotra V, Ross K, Ryder G, Zhou Y-P (2012) Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing & Service Oper. Management* 14(1):66–81.

Passerotti CC, Passerotti AM, Dall'Oglio MF, Leite KR, Nunes RL, Srougi M, Retik AB, Nguyen HT (2009) Comparing the quality of the suture anastomosis and the learning curves associated with performing open, freehand, and robotic-assisted laparoscopic pyeloplasty in a swine animal model. *J. Amer. College Surgeons* 208(4):576–86.

Pisano GP, Bohmer RMJ, Edmondson AC (2001) Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac surgery. *Management Sci.* 47(6):752–768.

Robbins TR (2007) Managing service capacity under uncertainty. Unpublished PhD thesis, Pennsylvania State University, Smeal College of Business, University Park, PA.

Robbins TR, Harrison TP (2011) New project staffing for outsourced call centers with global service level agreements. *Service Sci.* 3(1): 41–66.

Santner TJ, Williams BJ, Notz W (2003) *The Design and Analysis of Computer Experiments* (Springer, New York).

Singh J (2000) Performance productivity and quality of frontline employees in service organizations. *J. Marketing* 64(2):15–34.

Singh J, Goolsby JR, Rhoad GK (1994) Behavioral and psychological consequences of boundary spanning burnout for customer service representatives. *J. Marketing Res.* 31(4):558–569.

Witt LA, Andrews MC, Carlson DS (2004) When conscientiousness isn't enough: Emotional exhaustion and performance among call center customer service representatives. *J. Management* 30(1):149–160.

Wright TP (1936) Factors affecting the cost of airplanes. *J. Aeronautical Sci.* 3(4):122–128.

Yelle LE (1979) The learning curve: Historical review and comprehensive survey. *Decision Sci.* 10(2):302–328.