Les Cahiers du GERAD ISSN: 0711–2440

Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers

R. Ibrahim, P. L'Ecuyer,
H. Shen, M. Thiongane

G–2015–72

August 2015

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers

Rouba Ibrahim [a]

Pierre L'Ecuyer [b]

Haipeng Shen [c]

Mamadou Thiongane [b]

[a] *Department of Management Science and Innovation, University College London, London, UK, WC1E 6BT*

[b] *GERAD & Department of Computer Science and Operations Research, Université de Montréal, Montréal (Québec) Canada, H3C 3J7*

[c] *Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 27599*

rouba.ibrahim@ucl.ac.uk
lecuyer@iro.umontreal.ca
haipeng@email.unc.edu
mdthiongane@yahoo.fr

August 2015

Les Cahiers du GERAD

G–2015–72

Abstract: Traditionally, both researchers and practitioners rely on standard Erlang queueing models to analyze call center operations. In those models, service times are assumed to be independent and identically distributed exponential random variables with a constant mean. Going beyond such unrealistic assumptions has strong implications, as is evidenced by theoretical advances in the recent literature. However, there is very little empirical research, analyzing the statistical properties of service times in practice, to support that body of theoretical work. In this paper, we carry out a large-scale data-based investigation of service times in a call center with many heterogeneous agents and multiple call types. We observe that, for a given call type: (a) the service-time distribution depends strongly on the individual agent, (b) that it changes with time, and (c) that average service times are correlated across successive days or weeks. We develop stochastic models that account for these facts. In our proposed models, the service-time distribution is assumed to be lognormal with a mean that obeys a linear mixed-effects model with a weekly Gaussian random effect, and these successive weekly effects obey an autoregressive process of order one. We compare our models to simpler ones, e.g., where the mean service time depends only on the agent and call type, or only on the call type, and we find that our proposed models have a better goodness-of-fit, both in-sample and out-of-sample. We also perform simulation experiments to show that the choice of model can have a significant impact on the estimates of common measures of quality of service in the call center. Our study provides empirical support to the theoretical research that goes beyond standard modelling assumptions in service systems.

Key Words: Applied probability, call centers, service times, agent heterogeneity, correlation.

1 Introduction

The effective management of call centers is a challenging task mainly because managers are consistently facing considerable uncertainty; see Gans et al. (2003) and Aksin et al. (2007) for background on call centers. Among important sources of uncertainty are call arrival rates which are typically both time-varying and stochastic, service times which are random and whose distribution may depend on the call type and the agent who handles it, and agents who may not show up or may not follow their planned schedules; see Bhulai and Koole (2003), Avramidis et al. (2004), Avramidis and L'Ecuyer (2005), Aldor-Noiman et al. (2009), Gans et al. (2010), Ibrahim et al. (2012), Oreshkin et al. (2015), and references therein.

In this paper, we focus on the effective modelling of service times in call centers. In particular, we carry out a large-scale, in-depth, empirical investigation of service times in call centers. We analyze data gathered at the call center of Hydro-Québec (HQ), which is a government-owned public utility overseeing the generation, transmission, and distribution of electricity for the province of Quebec, in Canada. This real call center setting is complex, consisting of many heterogeneous agents and multiple distinct call types. Our data show that service times differ greatly across such agents, vary in time, and exhibit strong serial and cross correlations. We propose new service-time models which account for those features, and which are a good fit to real-life data, both in-sample and out-of-sample. We are interested in out-of-sample predictions because it is important to verify that our models are reliable tools to predict the future mean service times of agents, which have a considerable impact on future system performance.

Proposing new and more realistic service-time models, as we do in this paper, is important for the effective simulation of call centers. Simulation is an important tool that can be used to evaluate performance measures such as service levels and average waiting times, and to construct work schedules for agents and call routing rules by stochastic optimization algorithms (Avramidis et al., 2010; Chan et al., 2014). We use simulation to show that the choice of service-time model can have a significant impact on the performance measures in call centers, and formulate valuable insights on the practical usefulness of those findings.

1.1 Background, positioning in the literature, and contributions

Traditionally, both researchers and practitioners relied on standard Erlang queueing models to analyze call center operations. In Erlang queueing models, agent service times are modelled as independent and identically distributed exponential random variables with a constant mean. Going beyond this standard modelling assumption has important operational consequences, as is evidenced by multiple advances in the recent literature.

Agent heterogeneity. There is a stream of papers which studies queueing models with heterogeneous servers, with applications to call center management. One central question which arises in this context is how to route incoming calls to heterogeneous agents so as to minimize a given performance measure, such as the mean waiting time. Given the complexity of this problem, most papers resort to finding optimal routing policies in large-scale systems under heavy-traffic conditions; e.g., see Armony (2005), Gurvich and Whitt (2009), Armony and Ward (2010), Armony and Mandelbaum (2010), and references therein. Mehrotra et al. (2012) resort to a numerical study to characterize overall performance in terms of customer waiting time and overall resolution rate. In general, these papers show that control decisions can actually benefit from agent heterogeneity, e.g., routing incoming calls to the fastest idle agents reduces customer waiting.

There is very little empirical research supporting that body of theoretical work. To the best of our knowledge, the only exception is Gans et al. (2010) who analyzed call-center data and identified both short-term and long-term factors associated with agent heterogeneity in practice. They also described results from a small simulation study illustrating the operational consequences of ignoring such heterogeneity. (We revisit their example and extend some of their conclusions in § 6.) Gans et al. indicated that an interesting extension of their paper is to incorporate random effects in service-time models so as to "capture within-agent dependence among the calls handled by the same agent, and enable understanding of a whole agent population" (p. 118). We consider such random-effects models in this paper. In general, random effects represent additional, unobservable and uncontrollable, variability which causes systematic deviations from the

average performance of the agent, and due to which successive service times may be dependent. Dependencies between service times are often observed in data, and are therefore important to model, as we do in this paper.

Dependencies among the service times. Service times in practice are often dependent. For one example, an agent may be overworked in given periods (e.g., in weeks where congestion is higher than usual) and this could affect his/ her performance in all services that s/he performs during such periods, typically resulting in that agent either slowing down or speeding-up; see Delasay et al. (2015), Dong et al. (2015), Feldman et al. (2015), and references therein. In this case, agents (servers) may be viewed as *strategic decision makers* that influence they own service rates. As a result of such strategic behavior, successive service times are dependent. For a second example, in a technical call center, there may be a product defect due to which there are multiple related calls, whose durations are all longer than average. In this example too, service times (call durations) are dependent.

There is a well-developed theory studying the performance impact of dependence among service times in single-server queues; e.g., see Chapter 9 of Whitt (2002) for a detailed treatment. However, Pang and Whitt (2012) are among the first to consider the multi-server case, which is more reasonable from a practical perspective. They considered a weakly dependent stationary sequence of service times and demonstrated that, in the heavy-traffic limit, the impact of those dependencies is determined by the bivariate cumulative distribution function of service times. In their numerical study, they considered an EARMA sequence of service times, which is stationary with exponential marginal distributions and the correlation structure of an autoregressive-moving average process. Pang and Whitt demonstrated, via theoretical analysis and computer simulation, how dependencies between service times can significantly alter large system performance. In particular, they showed that those correlations strongly impact the distribution of the number of customers in queue which, in turn, affects staffing decisions. Pang and Whitt concluded their paper by calling for "empirical studies to estimate the strength of dependence among service times in applications" (p. 278). We conduct such a study in this paper.

Time dependence. There are relatively few papers which consider queueing models with time-varying service rates, since this feature substantially complicates the analysis. Some exceptions include Mandelbaum et al. (1999), Liu and Whitt (2011), and references therein. These papers demonstrate the operational impact of including time-varying service rates; their results apply generally and do not assume a specific form for time dependence in the service rate. Aldor-Noiman et al. (2009) used predictions of future arrival counts and mean service times to estimate future loads in call centers. Aldor-Noiman et al. allowed for mean service times to be time-dependent, and showed how errors in predicting future loads can impact staffing decisions. Their paper assumed homogeneous agents and a single call type. Our service-time models account for time dependence as well, albeit in a much more complex setting, with multiple call types and many heterogeneous agents.

Lognormal distribution. In their seminal paper, Brown et al. (2005) performed a detailed statistical analysis of call center data and showed that service times are not exponentially distributed, as was traditionally assumed, and that the lognormal distribution is a remarkably good fit for the service-time distribution instead. Deslauriers (2003) had also observed the same thing. Motivated by this, Shen and Brown (2006) proposed a new method for inference about non-parametric regression curves when the errors are lognormally distributed, and illustrated their method with both a simulation study and the analysis of real-life call center data. Mandelbaum and Zeltyn (2010) advocated a process-view of service times which are modeled as the evolution of a finite-state continuous-time absorbing Markov process (phase-type distribution). Here, even though we use additional information when modelling service times, such as the time when the call is answered, we continue to assume the lognormality of the individual service times.

In this paper, we supplement the body of theoretical research above with supporting empirical work. As such, we take a step towards filling that gap in the literature. In addition to proposing new service-time models that are a good fit to data, we quantify the performance impact of our alternative service-time models through a simulation study.

1.2 Organization

Here is how the rest of this paper is organized. In § 2, we describe and do a preliminary analysis of the data set that motivated this research. In § 3, we describe our candidate models. In § 4, we compare the in-sample goodness of fit of our models. In § 5, we compare the out-of-sample predictive accuracy of our models for a large pool of agents. In § 6, we present the results of simulation experiments which quantify the performance impact of our different models. In § 7, we make concluding remarks. In the online supplement, we describe additional models which we considered, and present additional details and numerical results.

2 Preliminary data analysis

The present data were gathered at the call center of HQ. The call center is virtual with over 15 locations across Quebec. They were collected over the span of one year, ranging from January 3, 2011 to December 31, 2011. The call center is open on weekdays and closed on weekends (Saturday and Sunday). The data consist of daily averages of service times for alternative agents and different call types. Even though it is desirable to study call-by-call data, many call centers still routinely collect aggregate summary data instead; see Pinedo et al. (1999) and Oreshkin et al. (2014), for example. Therefore, it is important to develop service-time models whose parameters can be estimated with such aggregated data, as we do here. In addition to daily averages of service times, the data contain information about the daily number of calls handled by each agent, per call type. Call types are distinguished by both the nature of the service request and the language, either French or English, in which the call is handled.

A service time often consists of a first part handled by an interactive voice response (IVR) system, and a second part where the call is handled by an agent. Since we are interested in modelling service times from the viewpoint of agents, we do not consider the IVR part because agents are not involved for that part. The time spent by customers in the IVR is studied by Salcedo-Sanz et al. (2010) and Colladon et al. (2013), for example. From the viewpoint of an agent (our viewpoint), an individual service time is the sum of: (i) the time spent actually talking to the customer (call time), and (ii) the post-call time spent by the agent to wrap up issues related to the call, during which s/he remains unavailable.

2.1 Overview

In our data set, there are 148 call types handled by a group of 1,655 agents. Alternative agents have different skills and they handle different call types depending on those skills. In total, there are 16,328 distinct agent/call type combinations, where each combination corresponds to an agent handling a particular call type. Many call types have very few corresponding calls, and are not interesting for us to study. We remove from our data set all call types that have less than 10 calls in total, across all agents, and are left with 86 call types handled by a total of 1,562 agents.

To sketch a temporal distribution of the workforce, we plot in Figure 1 the average number of agents answering calls per weekday, with 95% confidence bands that correspond to the 2.5% and 97.5% empirical quantiles, based on agents who have handled at least 10 calls in the data. We see that that the number of agents is highly variable on Mondays, and that Fridays have the least number of agents, on average. In Figure 2, we plot the total average call volume per weekday, including all call types. Consistent with Figure 1, Figure 2 shows that call volumes on Mondays exhibit the highest variance, and that call volumes on Fridays are lowest on average.

Agents typically handle more than one call type on any given day; also, a single call type is typically handled by more than one agent. For example, roughly 400 agents handle from 1 to 3 distinct call types, and about 25 call types are handled by roughly 65 agents each. The median of the total number of call types handled per agent (over the one year period) is 13, and the median of the number of agents handling a given call type is 33.

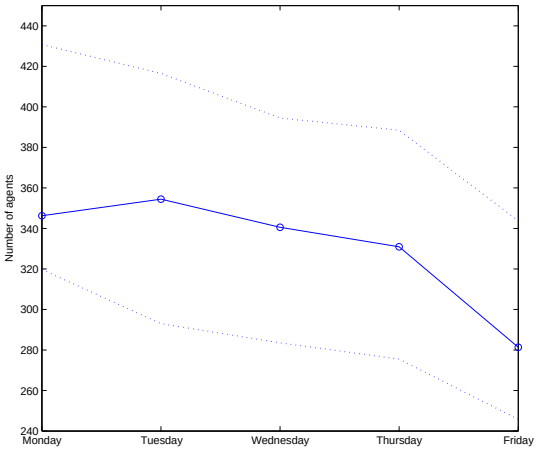

Figure 1: Average number of agents per weekday and corresponding 95% confidence bands.

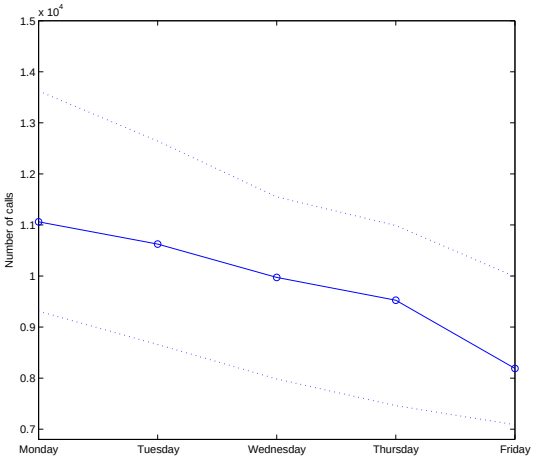

Figure 2: Average number of answered calls and corresponding 95% confidence bands.

2.2 Statistics on service times

We report several important empirical observations from our data. Our stochastic models for service times are developed to incorporate such features.

Variation across call types. Figure 3 gives a scatter plot of the empirical means versus variances of service times for different call types in our data. Each point in the plot corresponds to a given (mean, variance) pair, corresponding to a given call type. Figure 3 shows that there are significant differences in means and variances across different call types. As expected, Figure 3 shows that call types with longer durations generally exhibit higher variances. We take this variation between call types into account in § 3.

Agent heterogeneity. Service time distributions for the same call type vary considerably depending on the agent. In Figures 4 and 5, we illustrate this agent heterogeneity. We plot average service times for two call types: A, which is handled by 991 agents, and B, which is handled by 997 agents, as a function of the total number of calls answered (over the one-year period covered by our data) by each agent.

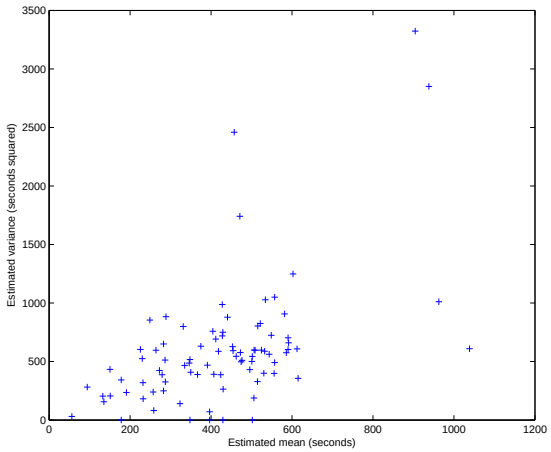

Figure 3: Each point corresponds to a (mean, variance) pair for a given call type.

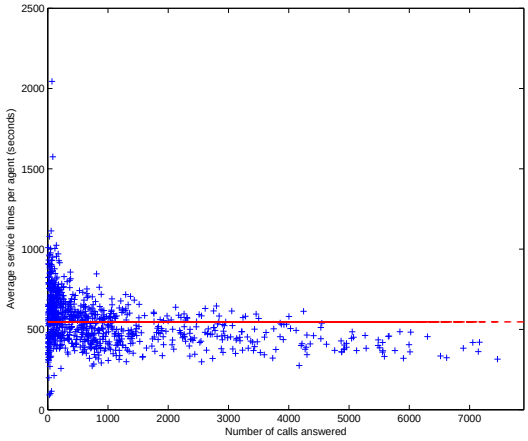

Figure 4: Average service times for different agents handling type A calls as a function of the total number of calls answered per year. The horizontal line is the overall average across all agents.

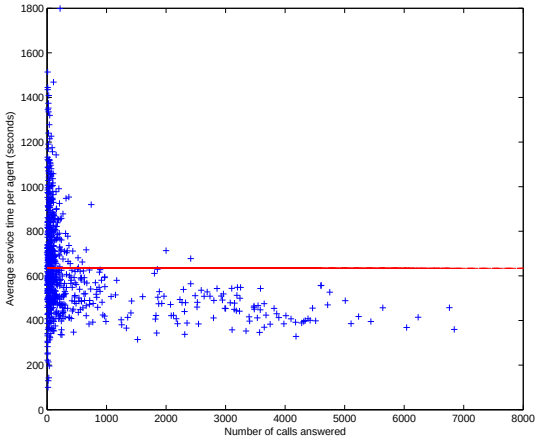

Figure 5: Average service times for different agents handling type B calls as a function of the total number of calls answered per year. The horizontal line is the overall average across all agents.

The horizontal line in each figure indicates the overall average service time across all agents, for each call type. Figures 4 and 5 show that there is significant variability in service times across all agents. Figures 4 and 5 also show that there are clearly clusters of agents who seem to perform in a roughly similar manner (having either shorter or longer than average service times). In general, agents who have handled many calls during the year are much faster on average than those who have handled few calls. The latter are either agents who have handled very few calls in general, or ones who have mostly handled calls of other types. In general, it appears that agents who have handled more calls tend to exhibit less variance in their service times. In other words, the larger dispersion is mainly exhibited by less experienced agents (those answering fewer calls). In Figures 6 and 7, we plot estimates of the variances of service times for all agents handling Type A and Type B calls, respectively, as a function of the total number of calls of that type answered by the agent. Figures 6 and 7 confirm that there are clear differences in variance of service times across agents.

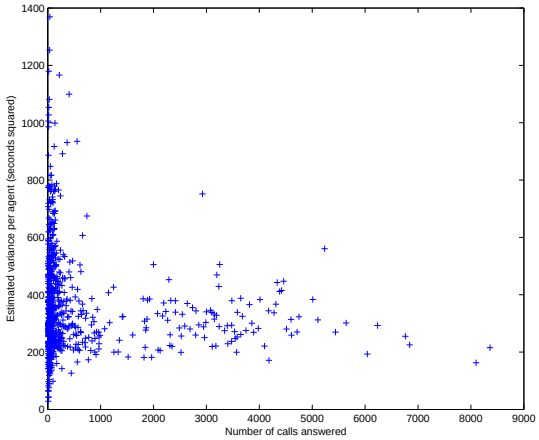

Figure 6: Estimated variances of service times for agents handling type A calls as a function of the total number of calls answered per year.

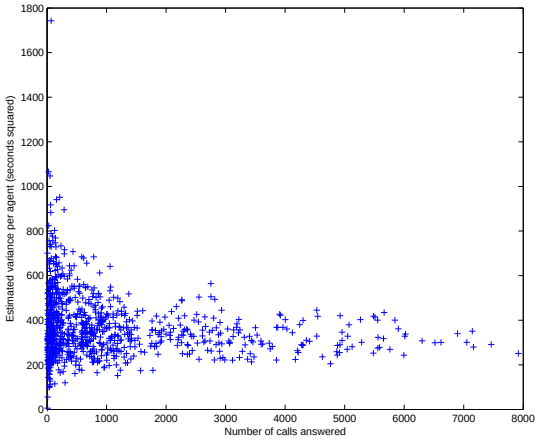

Figure 7: Estimated variances of service times for agents handling type B calls as a function of the total number of calls answered per year.

In Figure 8, we plot the average service times for four agents handling calls of type B, as a function of time (index of day). Additionally, we include horizontal lines corresponding to the overall average service times for those agents. Figure 8 illustrates that different agents exhibit different behavior. Indeed, the top two agents are evidently slower than the bottom two agents, and their service times also have higher variances.

Time dependence. In addition to variability across different agents, our data show that the average service time for a given agent and a given call type varies significantly over time.

In Figure 9, we plot the average daily service times for an agent handling four different call types, as a function of time. These daily averages clearly vary with time. Figure 9 illustrates a phenomenon which could be important from an operational perspective: The agent seems to be slowing down as he/she handles more call types. Indeed, this is apparent starting day 208 when the agent begins handling calls of type 4. Thereafter, Figure 9 shows that the average service times for call types 1 and 2 increase. Based on such observations, we experimented with including the number of call types handled by an agent as a covariate in our service-time models. We did not include such models in this paper because they lead to less accurate out-of-sample predictions of future expected service times on average over all agents in our data; see the online supplement for more details. Figure 8 also illustrates that average service times fluctuate across successive days.

In Figure 10, we illustrate time dependence by plotting the evolution, over time, of the daily average service times for an agent a_1 handling type A calls. In Figure 10, we also include the best linear fit for the data. This plotted line clearly shows an upward trend in the average service times for this agent. In our data, we observed both upwards and downward trends, depending on the agent. One explanation for downward trends is that agents are learning with time; see Gans et al. (2010) for further empirical support. There may be many other explanations for such trends. For example, with upward trends, it may be that agents are getting bored and less motivated to answer calls quickly.

2.3 Cohort C of 200 agents

The total number of calls answered per agent varies widely across agents in our data. The maximum is 14,715 calls handled by one agent over the one year period, but hundreds of agents have answered very few calls. For these agents, it is difficult to fit service-time models and do reliable predictions. Moreover, with insufficient data it is hard to reach meaningful results. For the remainder of this paper, we restrict our attention to agents who answered a relatively large number of calls; specifically the 200 agents who answered the most calls during the year. These 200 agents answered a total of 1,175,178 calls, which corresponds to roughly

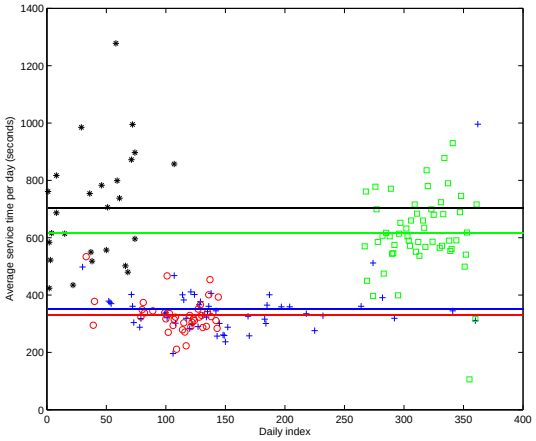

Figure 8: Average service times for 4 agents handling type B calls versus the index of day.

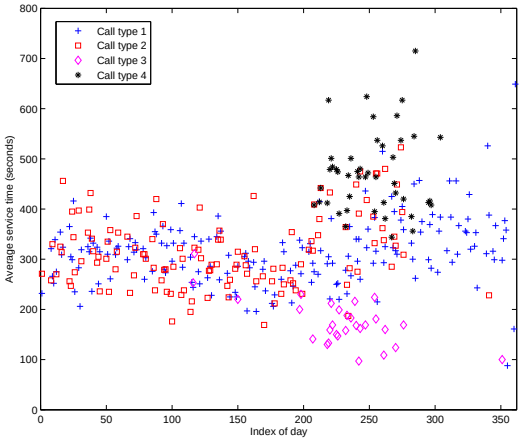

Figure 9: Average daily service times for an agent whose skill set increases on day 208.

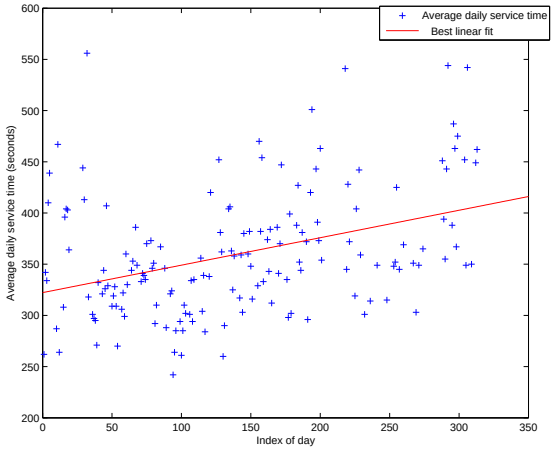

Figure 10: Evolution of the average service time of agent a_1 for type A calls, and best linear fit.

half of the total number of calls incoming to the center during the year. For each one of these 200 agents, we removed (agent, call type) pairs where the agent handles the call type for less than a total of 10 days in our data set. We do this to avoid considering (agent, call type) pairs with too few observations.

There is a total of 550 different (agent, call type) pairs which remain in our cohort. In Tables 1–3 of the supplement, we present detailed out-of-sample prediction results for each agent in our cohort. The results are also summarized in § 5. In Tables 1–3, we also report the total number of out-of-sample predictions that we made for each agent, across all call types that the agent handles. We provide additional detail in the supplement.

Hereafter, we refer to our cohort of agents as cohort C. It is important to note that these are not the 200 rightmost agents in Figures 4–7. In total, these 200 agents handle 30 different call types, and the number of skills per agent ranges from 1 to 8. The average number of skills per agent, in this subset of the data, is 3.9.

3 Models for service times

In this section, we propose alternative models for the process of service times. We begin by describing two benchmark models which mimic standard practice.

3.1 Benchmark models: Model $B1$ and Model $B2$

The preliminary analysis of § 2 suggests that service times depend strongly on both the agent and the call type considered; see Figures 4–10. Let $S_{i,j}$ denote the service time of a call of type j handled by agent i, where $j = 1, 2, ..., J$ and $i = 1, 2, ..., I$.

In our first benchmark model, Model $B1$, we assume that $S_{i,j}$ are i.i.d. lognormal random variables with expected value m_j and variance v_j, for every i and j, where m_j and v_j depend solely on the call type j. In our second benchmark model, Model $B2$, we assume that the expected value $m_{i,j}$ and the variance $v_{i,j}$ depend on both the call type j and the agent i.

Since our data only consist of aggregated daily averages of service times, instead of detailed call-by-call data, it is not immediately clear how to compute point estimates for those expected values and variances. To do so, we adopt here the method of moments as in Deslauriers (2003). Alternatively, for a review of estimation methods (for the mean) with more detailed data, see Shen and Brown (2006).

Method of moments. We provide additional details for this method by focusing on estimation for Model $B2$. For Model $B1$, we do the same but do not distinguish between alternative agents handling the same call

type. Let $n_{i,j}^{(k)}$ be the number of calls of type j handled by agent i on day k, where $k = 1, 2, ..., K_{i,j}$, and $K_{i,j}$ is the total number of days where agent i handles calls of type j. Let $\hat{m}_{i,j}^{(k)}$ be the average service time for a call of type j handled by agent i on day k, based on a sample of $n_{i,j}^{(k)}$ answered calls. Our data set contains day-by-day values for both $n_{i,j}^{(k)}$ and $\hat{m}_{i,j}^{(k)}$. We define $\hat{m}_{i,j}$ and $\hat{v}_{i,j}$ as follows:

$$\hat{m}_{i,j} = \frac{\sum_{k=1}^{K_{i,j}} n_{i,j}^{(k)} \hat{m}_{i,j}^{(k)}}{\sum_{k=1}^{K_{i,j}} n_{i,j}^{(k)}} , \tag{1}$$

and

$$\hat{v}_{i,j} = \frac{1}{K_{i,j} - 1} \sum_{k=1}^{K_{i,j}} n_{i,j}^{(k)} (\hat{m}_{i,j}^{(k)} - \hat{m}_{i,j})^2 . \tag{2}$$

These $\hat{m}_{ij}$ and $\hat{v}_{i,j}$ are unbiased estimators of $m_{i,j}$ and $v_{i,j}$ for each agent i and call type j; see Deslauriers (2003) for additional details.

3.2 Model $A1$: Fixed-effects model

The preliminary analysis of § 2 suggests that the average service time for a given agent and call type is not constant over time; see Figure 10. Let $M_{i,j}^{(k)}$ be a random variable representing the average service time for a call of type j handled by agent i on day k. This is what we observe in our data. In Model $A1$, we assume that $M_{i,j}^{(k)}$ follows a Gaussian process which is a linear additive fixed-effects model incorporating an intercept and a linear trend. That is, we assume for each pair (i, j) that:

$$M_{i,j}^{(k)} = \alpha_{i,j} \cdot k + \beta_{i,j} + \epsilon_{i,j}^{(k)} . \tag{3}$$

The coefficients $\alpha_{i,j}$ and $\beta_{i,j}$ are real-valued constants that need to be estimated from data, and $\epsilon_{i,j}^{(k)}$ are i.i.d. normal random variables with mean 0 and variance $\sigma_{\epsilon_{i,j}}^2 / n_{i,j}^{(k)}$, where $n_{i,j}^{(k)}$ is the number of calls of type j answered on day k by agent i. That is, the number of calls answered in a given day is used as a weight in our regression model. We estimate the model in (3) using weighted least squares. Of course, modeling the mean service time as a linear function of time can only make sense as a rough approximation over a limited time interval. For example, a time-decreasing mean is typically due to a learning effect, but this effect eventually saturates and the slope of the decrease should become closer to 0 as time goes on. In fact, we will find that this model with a linear trend is outperformed by our next two models, which do not include such a linear trend. In addition to a linear time trend, we also considered quadratic and logarithmic trends. However, since models with such trends performed consistently worse, we only present results for a linear trend in this paper.

We found that the normality assumption of the mean service times is reasonable in our data. That is to be expected since our data consist of daily averages where each average is typically calculated based on tens of service times per day. For example, in Figures 11 and 12, we present normal Q-Q plots for the residuals of Model $A1$ for two agents, a_1 and a_2, along with pointwise 95% confidence bands. Agents a_1 and a_2 handled many calls of type A: 8360 and 8098 calls, respectively. We also got consistent results for agents who answer a relatively small number of calls of a given call type, e.g., 200-300 calls during the year. For all agent-call type pairs, we conducted the Lilliefors test for normality on the residuals of model $A1$; across all such pairs, the first three empirical quartiles of the distribution of p-values for this test are 0.005, 0.08, and 0.3, respectively. Overall, we found that there was typically not enough statistical evidence to reject the null hypothesis that $\alpha_{i,j} = 0$. Specifically, the empirical estimates of the first three quartiles of the distribution of corresponding p-values are given by: 0.007, 0.2, and 0.5, respectively; in particular, we could not reject the null hypothesis in more than 60% of the agent-call type pairs. We also conducted Ljung-Box tests on the autocorrelations of residuals for Model $A1$, and the quartiles of the empirical distribution of p-values were 0.04, 0.3, and 0.6. For at least 25% of the agent-call type pairs, autocorrelations are significant at the 5% level.

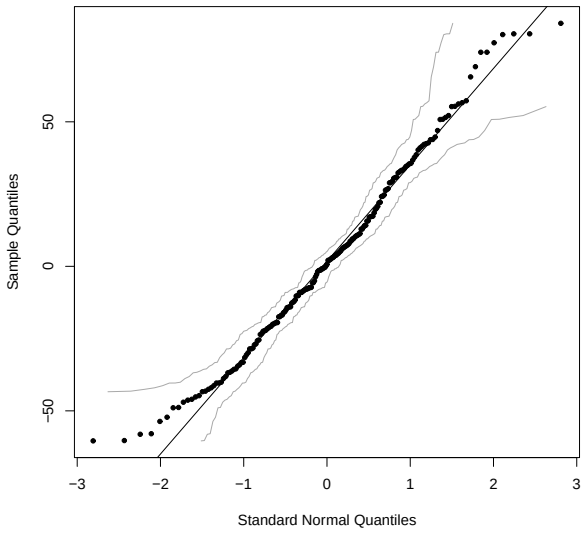

Figure 11: Q-Q plot for the residuals of Model $A1$ for agent a_1 and 95% confidence bands .

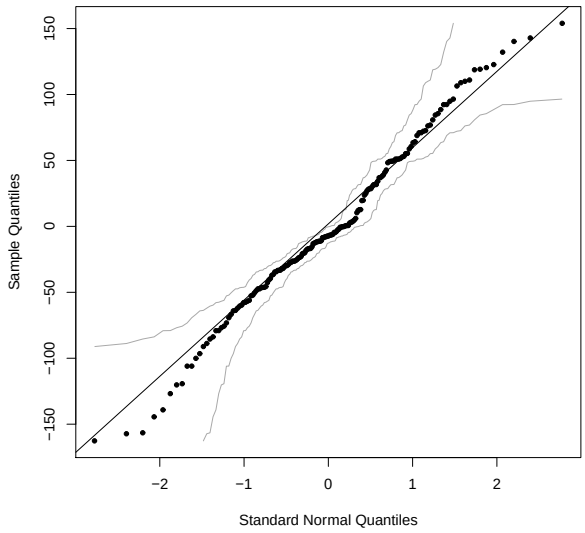

Figure 12: Q-Q plot for the residuals of Model $A1$ for agent a_2 and 95% confidence bands .

3.3 Model A2: Serial correlations

Capturing dependencies between successive service times amounts to capturing dependencies between (approximately) normal mean service times. Mixed-effects models are ideal to capture such dependencies with roughly normally-distributed data; we now propose one such model. We consider a Gaussian linear mixed-effects model for $M_{i,j}^{(k)}$:

$$M_{i,j}^{(k)} = \beta_{i,j} + \gamma_{i,j}^{(w_k)} + \nu_{i,j}^{(k)} \ , \tag{4}$$

where $\gamma_{i,j}^{(w_k)}$ is a random effect specific to the week w_k of day k, and $\nu_{i,j}^{(k)}$ is a normally distributed residual error. We assume that these residuals $\nu_{i,j}^{(k)}$ are independent normal with mean 0 and variance $\sigma^2_{\nu_{i,j}}/n_{i,j}^{(k)}$. The residual variance of $\nu_{i,j}^{(k)}$ is specific to each (i,j) pair; as such, we can capture differences in variance across different agent/skill pairs. The random effects $\gamma_{i,j}^{(w_k)}$ are normally distributed weekly deviations which we use to capture correlations in the average service times, for the same agent and call type, across successive weeks and across successive weekdays within the same week. Because of the aggregated nature of the available data, we do not consider a daily random effect in (4), but rather a weekly one, and we do not impose a covariance structure on the residuals $\nu_{i,j}^{(k)}$. Indeed, both would lead to identification issues since we do not have data for individual calls during a given day.

Thereafter, we omit the subscript of a random variable when the specific index is not important. In a Gaussian mixed-effects framework, $\gamma_{i,j}^{(w_k)}$ and $\nu_{i,j}^{(k)}$ are assumed to be normally distributed and independent. Here, we assume that the random effects, $\gamma_{i,j}^{(w_k)}$, are identically normally distributed with expected value $E[\gamma_{i,j}^{(w_k)}] = 0$ and variance $\mathrm{Var}[\gamma_{i,j}^{(w_k)}] = \sigma^2_{\gamma_{i,j}}$, and that $\gamma_{i,j}^{(w_k)}$ follows a first-order autoregressive covariance structure, AR(1). That is,

$$\gamma_{i,j}^{(u)} = \rho_{i,j}\gamma_{i,j}^{(u-1)} + \psi_{i,j}^{(u)} \ , \tag{5}$$

where $\rho_{i,j}$ is the autocorrelation parameter, and $\psi_{i,j}^{(u)}$ are i.i.d. normally distributed random variables with $E[\psi_{i,j}^{(u)}] = 0$ and $\mathrm{Var}[\psi_{i,j}^{(u)}] = \sigma^2_{\gamma_{i,j}}(1-\rho^2_{i,j})$. The covariance between $\gamma_{i,j}^{(u_1)}$ and $\gamma_{i,j}^{(u_2)}$ is given by

$$\mathrm{Cov}(\gamma_{i,j}^{(u_1)}, \gamma_{i,j}^{(u_2)}) = \sigma^2_{\gamma_{i,j}}\rho_{i,j}^{|u_2-u_1|} \ . \tag{6}$$

Assuming an AR(1) covariance structure for $\gamma_{i,j}^{(w_k)}$ is both useful and computationally efficient, because it requires the estimation of only two parameters, $\sigma_{\gamma_{i,j}}$ and $\rho_{i,j}$.

Here, the weekly random effect that follows an AR(1) process replaces the linear trend that we had in Model $A1$. It allows for a situation where for a given agent, for example, the mean decreases for some period of time due to learning, then remains stable, then increases later because the agent loses interest or has other problems, etc. This AR(1) process is very simple and yet sufficiently flexible to model such mid-term variations in the mean.

We also tried Model $A2$ with a linear trend as in $A1$ in addition to the AR(1) term, and found that the version without the linear trend provided a better fit to the data out of sample. For this reason, we omitted the linear trend. In Table 1, we present point estimates for the different parameters of Model $A2$, based on our data, for 3 agent/call type combinations. The p-values in the table are computed automatically in SAS® as follows: Assuming the normality of the random effects and residuals, we can construct a statistic depending on the fixed effects (β) and the random effects (γ) which can be shown to have approximately a $t-$distribution for which we can estimate the degrees of freedom. Based on this statistic we can make inference about whether the random and fixed effects (the linear trend) are equal to 0. The weekly random effect and the autocorrelation parameter are generally found to be statistically significant. When testing the linear trend, the quartiles of the empirical distribution of p-values were 0.06, 0.3, and 0.6. The trend is statistically significant at the 95% level for 125 pairs out of 550. For the model without the linear trend, autocorrelation is statistically significant at the 95% level for 246 pairs out of 550, and the quartiles of the distribution of p-values are 0, 0.002 and 0.3.

Table 1: Results for Model $A2$ for 3 different agent/call type combinations. Point estimates of model coefficients are shown with corresponding standard errors and p-values of t-tests for statistical significance.

(Agent, call type)	Category	Value	Std. error	p-value
(i_0, j_0)	$\sigma^2_{\gamma_{i_0,j_0}}$	1270	1004	0.1035
	ρ_{i_0,j_0}	0.705	0.269	0.00870
	$\sigma^2_{\epsilon_{i_0,j_0}}$	146000	20800	$< .0001$
	β_{i_0,j_0}	602	45.7	$< .0001$
	α_{i_0,j_0}	-0.634	0.204	0.00250
(i_1, j_1)	$\sigma^2_{\gamma_{i_1,j_1}}$	608	356	0.0439
	ρ_{i_1,j_1}	0.870	0.0846	$< .0001$
	$\sigma^2_{\epsilon_{i_1,j_1}}$	93867	10300	$< .0001$
	β_{i_1,j_1}	295	21.9	$< .0001$
	α_{i_1,j_1}	0.0885	0.101	0.383
(i_2, j_2)	$\sigma^2_{\gamma_{i_2,j_2}}$	1320	684	0.0267
	ρ_{i_2,j_2}	0.652	0.244	0.00760
	$\sigma^2_{\epsilon_{i_2,j_2}}$	51000	8030	$< .0001$
	β_{i_2,j_2}	283	25.4	$< .0001$
	α_{i_2,j_2}	0.243	0.156	0.124

3.4 Model $A3$: Serial and cross correlations

Dependencies in the time series of service times may be due to factors linked to the agents themselves, such as stress, fatigue, demotivation, etc. Such short-term effects may influence agent performance during a given period of time and cause dependencies between the service times of all calls handled by that same agent. Considering models with cross correlations is therefore important to capture similar effects.

In Model $A3$, we jointly model the service times of different call types handled by the same agent. We consider a mixed-effects model for the mean service time (just as in Model $A2$) where we merge alternative call types together and have the same weekly random effect common to all types handled by the same agent. This gives:

$$M_{i,j}^{(k)} = \beta_{i,j} + \gamma_i^{(w_k)} + \nu_{i,j}^{(k)} \ . \tag{7}$$

The intercept $\beta_{i,j}$ is specific to call type j handled by agent i. We continue to assume an AR(1) covariance structure for $\gamma_i^{(w_k)}$. We let $\gamma_i^{(w_k)}$ depend on the agent i and the week w_k, but not on the call type j. We also continue to assume that model residuals are i.i.d. normal with expected value 0 and variance $\sigma^2_{\nu_{i,j}}/n^{(k)}_{i,j}$. The random effect $\gamma_i^{(w_k)}$, which is common for all call types handled by agent i, exploits both serial correlations across successive weeks, and cross correlations across different call types. The residual variance of $\nu_{i,j}^{(k)}$ is specific to each (i,j) pair; as such, we capture differences in variance across different agent/skill pairs.

As an illustration, Table 2, gives parameter estimates of Model $A3$ for the agent i_0 considered in Table 1. Here, the p-value for the weekly random effect is 0.0858. Some other p-values are quite small. We also tested Model $A3$ with a linear trend, for our cohort of 200 agents, and the quartiles of the distribution of p-values for the test on the linear trend were 0.04, 0.3, and 0.6. That is, for most agents the trend is not significant. In out-of-sample goodness of fit tests and predictions based on Model $A3$ both with and without this linear trend, the version without a trend fit the data better. Therefore, we omit this linear trend from consideration in § 4 and 5. For the model without a trend, the autocorrelation parameter is usually found to be statistically significant: the quartiles of the distribution of p-values were (approximately) 0, 0.005, and 0.2.

Table 2: Results for Model $A3$ for agent i_0, featured in Table 1, answering 3 different call types, numbered 1-3. Point estimates of model coefficients are shown with corresponding standard errors and p-values for statistical significance.

Category	Value	Std. error	p-value
$\sigma^2_{\gamma_{i_0}}$	1240	901	0.0858
$\rho_{\gamma_{i_0}}$	0.687	0.270	0.0108
$\sigma^2_{\epsilon_{i_0,1}}$	146000	47200	0.001
$\sigma^2_{\epsilon_{i_0,2}}$	149000	32300	< 0.0001
$\sigma^2_{\epsilon_{i_0,3}}$	145000	19800	< 0.0001
$\beta_{(i_0,1)}$	562.1	107	< 0.0001
$\beta_{(i_0,2)}$	454	43.2	< 0.0001
$\beta_{(i_0,3)}$	624	42.7	< 0.0001
$\alpha_{(i_0,1)}$	-0.628	0.686	0.361
$\alpha_{(i_0,2)}$	-0.371	0.193	0.0564
$\alpha_{(i_0,3)}$	-0.727	0.192	0.0002

4 Goodness of fit of the models

In this section, we assess the goodness of fit to data of our candidate models.

4.1 Model residuals

We begin by analyzing the residuals of each model, where model residuals are defined to be equal to the differences between the observed daily average service times and the corresponding fitted values. In Table 3, we present summary statistics for the square of the residuals for our cohort C of agents; see § 2. Table 3 shows that Models $A2$ and $A3$ are better fits to data than Model $A1$, $B1$ and $B2$, and that Model $A2$ is a slightly better fit than Model $A3$. Models $B1$ and $B2$ lag behind, and Model $B1$ clearly yields the worst fit.

We also compute estimates of the *root mean squared errors* (RMSE) for the cohort C under the different models. In Figure 13, we present box plots for the RMSE's across all models. Figure 13 shows that Models $A2$ and $A3$ are a better fit to data than the rest of the models. In Figure 14, we plot the empirical cumulative distribution functions (ECDF) of the RMSE's for all models. Once more, Figure 14 shows that Models $A2$ and $A3$ provide superior fits to data compared to other models. For the RMSE's, we ran matched-pair t-tests, at the 95% confidence level, for all model pairs and found that the differences in RMSE's were all significantly different from 0 (the corresponding p-values of all tests were very close to 0).

Table 3: Summary statistics for the square of residuals, under each model, across the cohort C of agents.

Statistic	Model $B1$	Model $B2$	Model $A1$	Model $A2$	Model $A3$
Mean	15,243	10,131	9,214	7,272	7,428
Median	4,238	2,296	2,099	1,574	1,624
First quartile	894	454	415	308	322
Third quartile	13,396	7,844	7,217	5,581	5,690

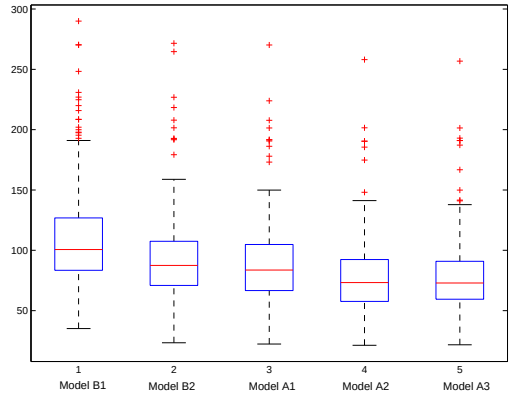

Figure 13: Box plots of the RMSE's of model residuals when fitting all models to the data from the cohort C of agents.

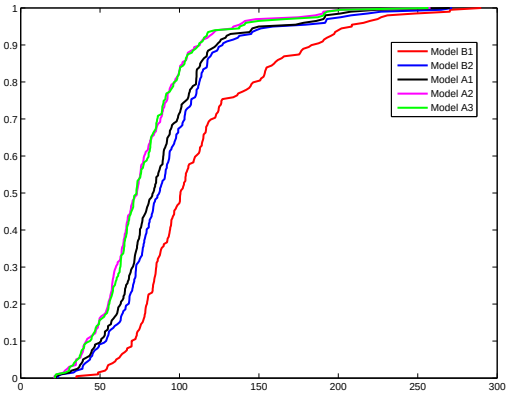

Figure 14: ECDF's for the RMSE's of model residuals when fitting models to the data from the cohort C of agents.

4.2 Distributional fits

We can use our candidate models to obtain full distributional fits for the service times, beyond expected values. However, given the aggregated nature of our data, it is not possible to investigate how well our models fit the distributions of the individual service times. Instead, we can only test how well they reproduce the distribution of daily averages observed in our data sample. In order to do so, we simulate independent replications of the mean service times under each model, and use the probability integral transform (PIT) (as in Rosenblatt (1952)) which is defined as follows for day k, agent i, and call type j:

$$\mathrm{PIT}_{i,j}^{(k)} \equiv \frac{1}{N}\sum_{l=1}^{N} \mathbb{I}(m_{i,j}^{(k)} < s_{i,j}^{(k)}(l)) \ , \tag{8}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, N is the number of simulation replications, $m_{i,j}^{(k)}$ is the observed average service time on day k for agent i and type j, and $s_{i,j}^{(k)}(l)$ is the corresponding simulated value, on replication l, assuming the same number of calls on day k as was observed in the data set. This measure has been used extensively in econometrics; see Weinberg et al. (2007).

Assume that $M_{i,j}^{(k)}$ is the random variable representing the mean service time for each i, j, and k. If our model is correct, then the distribution of $S_{i,j}^{(k)}$ (according to which we simulate $s_{i,j}^{(k)}$) and the distribution of $M_{i,j}^{(k)}$ should be identical. In particular, the random variable $F_{S_{i,j}^{(k)}}(M_{i,j}^{(k)})$ should then be $U[0,1]$, where F_X denotes the cumulative distribution function of a random variable X. It is not hard to see that $\mathrm{PIT}_{i,j}^{(k)}$ is, for large N, distributed as $F_{S_{i,j}^{(k)}}(M_{i,j}^{(k)})$. Therefore, if our model is correct then the random sample $\mathrm{PIT}_{i,j}^{(1)}, \mathrm{PIT}_{i,j}^{(2)}, \mathrm{PIT}_{i,j}^{(3)}, ..., \mathrm{PIT}_{i,j}^{(K)}$ should behave like i.i.d uniform $U[0,1]$ random variables for all i, j. To test whether the PIT's under a given model are a good fit to the uniform distribution, we use chi-square goodness

of fit tests under the 95% confidence criterion. Under these tests, the null hypothesis is that the PIT comes from a $U[0,1]$ distribution.

In Table 4, we summarize the results of those chi-square tests for the cohort C of agents that we considered. Recall that there are 550 different agent/skill combinations in the subset of the data that we consider. For each combination, we consider whether the uniform distribution is a good fit for the corresponding PIT's. Then, for each model, we report the proportions of the chi-square tests that reject the null hypothesis at the 95% level. Table 4 shows that Models $A2$ and $A3$ yield better distributional fits to the data than Model $A1$ and the benchmark models, and that Models $A2$ and $A3$ fit the data roughly similarly. Indeed, Table 4 shows that, with Model 1, we fail to reject the null hypothesis of uniform PIT's for only 59% of the different agent/skill combinations. In contrast, with Model $A2$, we fail to reject this null hypothesis for roughly 73% of the different agent/skill combinations, and with Model 3 we do so in roughly 70% of the agent/skill combinations. While these numbers remain below the desired 95% level, the improvement in fit over the remaining models, particularly the benchmark models commonly used in practice, is significant.

Table 4: Summary of results of chi-square goodness of fit tests to the uniform distribution for the PIT of all models, across the 550 agent/call type combinations considered. We report the proportions of tests for which the null hypothesis that the PIT's fit the uniform distribution is rejected at the 95% level.

	Model $B1$	Model $B2$	Model $A1$	Model $A2$	Model $A3$
Reject	0.812	0.533	0.413	0.267	0.300

5 Predictions of average service times

We now compare the statistical models of § 3 based on their out-of-sample forecasting performance, for our cohort C of agents. For each agent and call type, each model, and each day i, we estimated the model based on all the observations up to day $i-\delta$ only (the learning period), where δ is a selected prediction lag or *lead time*, and from that we computed a prediction $\hat{m}_i$ of the average service time m_i for day i. We considered only days i for which $i-\delta \geq 60$. Each $\hat{m}_i$ is an out-of-sample forecast (based only on past information). We consider three different prediction lead times δ, namely 2 weeks, 1 week, and 1 day, to mimic real-life challenges faced by call center managers. We roll the learning period forward so as to preserve the length of the lead time. We re-estimate all model parameters after each prediction. We use the Mixed Procedure in SAS® to compute maximum likelihood estimates of the parameters for Model $A2$ and Model $A3$, and to generate the corresponding forecasts; see §2 of the supplement for a detailed description of how we compute each prediction.

5.1 Performance measures

We quantify the accuracy of a point prediction by computing the *root mean squared error* (RMSE) per day, defined by:

$$\mathrm{RMSE} \equiv \sqrt{\frac{1}{K}\sum_{i=1}^{K}(m_i - \hat{m}_i)^2}\ , \tag{9}$$

where m_i is the observed average service times for a given day i, $\hat{m}_i$ is the predicted value of m_i, and K is the total number of predictions made. The RMSE is in units of seconds. We also compute the *mean absolute percentage error* (MAPE), which gives a relative measure of accuracy defined by:

$$\mathrm{MAPE} \equiv 100 \cdot \frac{1}{K}\sum_{i=1}^{K}\frac{|m_i - \hat{m}_i|}{m_i}\ . \tag{10}$$

5.2 Predictive performance

In Tables 5 and 6, we report aggregate results for the out-of-sample forecasts over all 200 agents, call types, and days. Detailed numerical results for each agent are given in Tables 3-5 of the supplement. In Table 5, we include estimates of the average, the median, and the first and third quartiles of the MAPE's and RMSE's obtained across all agents. Recall that each MAPE and RMSE is over all call types and days, for each agent. We highlight in bold the minimum RMSE and MAPE in each row. Clearly, Model $A3$ is superior throughout. It clearly outperforms our benchmark models, commonly used in practice, particularly with a short forecasting lead time (one day). We now briefly discuss the results for lead times of 2 weeks and 1 day, respectively.

Two-weeks-ahead predictions. Over this long lead time, Models $A2$ and $A3$ perform nearly the same. Model $B2$ is competitive as well, and yields smaller prediction errors than Model $A1$, for both the MAPE and the RMSE. The average MAPE for Model $A3$ is roughly 12% lower than for $A1$ and 3% lower than for $B2$. On the other hand, Model $B1$ clearly lags behind, with a MAPE 26% larger than for Model $A3$. Similar results hold when comparing the average RMSE, which is roughly 18% larger for Model $A1$ than for $A3$, and 3% larger for $B2$ than for $A3$. We ran matched-pair t-tests, for all possible model pairs, to test if the differences in the RMSE's and MAPE's are significantly different from 0. The p-values of those tests were all very small (the difference is clearly significant, with $p < 0.0001$) except in two cases: When comparing the RMSEs for Models $B1$ and $A1$, we obtain $p = 0.48$ and when comparing the MAPEs for Model $A2$ and $A3$, we obtain $p = 0.12$.

One-day-ahead predictions. With a forecasting lead time of 1 day, the advantage of exploiting correlations between weekly averages increases. Models $A1$, $A2$, and $A3$ yield more accurate predictions than both benchmark models, with Model $A3$ taking the lead. For example, the average MAPE is roughly 6% smaller for Model $A3$ than for $B2$, and roughly 5% smaller for $A2$ than for $A1$. The average RMSE is roughly 10% smaller for Model $A3$ than for $A1$, and roughly 9% smaller for $A2$ than for $A1$. Model $B1$ is clearly outperformed by all other models. In matched-pair t-tests, all p-values were very small ($p < 0.0001$).

Table 5: Predictive accuracy for Models $A1$, $A2$, and $A3$, averaged across our cohort C of agents.

Forecast lead time of two weeks										
	Model $B1$		Model $B2$		Model $A1$		Model $A2$		Model $A3$	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Average	107	26.2	91.5	20.1	109	22.0	89.6	19.6	**88.7**	**19.4**
Median	95.2	21.7	88.1	18.6	97.4	20.7	86.7	18.0	**86.0**	**17.9**
First quartile	77.3	17.3	70.0	15.3	74.7	16.5	69.1	15.0	**68.2**	**14.9**
Third quartile	122.0	28.9	109	23.3	136	25.7	108	22.8	**106**	**22.8**
Forecast lead time of one week										
	Model $B1$		Model $B2$		Model $A1$		Model $A2$		Model $A3$	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Average	107	26.2	90.6	19.9	100.0	20.4	88.3	19.2	**87.2**	**19.0**
Median	94.7	21.8	87.6	18.5	96.1	19.4	85.0	17.8	**84.2**	**17.6**
First quartile	77.3	17.4	68.5	15.3	73.5	15.9	67.9	15.0	**67.4**	**14.7**
Third quartile	122	28.8	109	23.3	124	23.5	107	**22.3**	**105**	22.7
Forecast lead time of one day										
	Model $B1$		Model $B2$		Model $A1$		Model $A2$		Model $A3$	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Average	107	26.2	89.9	19.6	95.2	19.4	86.6	18.5	**85.4**	**18.4**
Median	95.1	21.7	86.1	18.3	91.6	18.4	82.3	17.2	**82.3**	**17.1**
First quartile	77.6	17.4	67.6	15.2	70.4	15.3	66.8	14.6	**65.9**	**14.2**
Third quartile	122	28.7	108	23.2	117	22.0	104	**21.6**	**102**	21.7

Table 6: Proportions where a given model is the winner, i.e., yields the smallest performance measure, across our cohort C of agents.

	Model $B1$	Model $B2$	Model $A1$	Model $A2$	Model $A3$
Forecast lead time of two weeks					
MAPE	0.202	0.227	0.202	0.157	**0.253**
RMSE	0.167	0.182	0.106	0.212	**0.364**
Forecast lead time of one week					
MAPE	0.172	0.187	0.227	0.202	**0.278**
RMSE	0.141	0.152	0.131	0.232	**0.389**
Forecast lead time of one day					
MAPE	0.141	0.146	0.207	0.232	**0.338**
RMSE	0.116	0.116	0.136	0.278	**0.424**

Proportion of wins for each model. Table 6 compares the models from a different viewpoint: it reports the proportions of agents (in our cohort C) where each model yields the smallest performance measure, across all models. For example, the first row in Table 6 indicates that, across the 200 agents considered, the smallest MAPE was achieved by Model $A1$ for 20.2% of the agents, by Model $A2$ for 15.7% of the agents, and by Model $A3$ for 25.3% of the agents. With a forecasting lead time of two weeks, Model $B2$ is competitive, but it is still outperformed by Model $A3$. The proportions in Table 6 do not sum up exactly to unity because there may be multiple minimizers of the MAPE (or RMSE) for a given agent. In Table 6, Model $A3$ generally performs better than all remaining models. This is especially true with a short forecasting lead time. For example, with a lead time of 1 day, Model $A3$ yields the smallest RMSE for 42.4% of the agents, compared with 11.6% for $B2$.

In Figures 15 and 16, we plot the empirical cumulative distribution functions for the RMSE's and MAPE's, respectively, for all models, with a lead time of one day. These figures illustrate the improvement in forecasting accuracy that is summarized in Table 5.

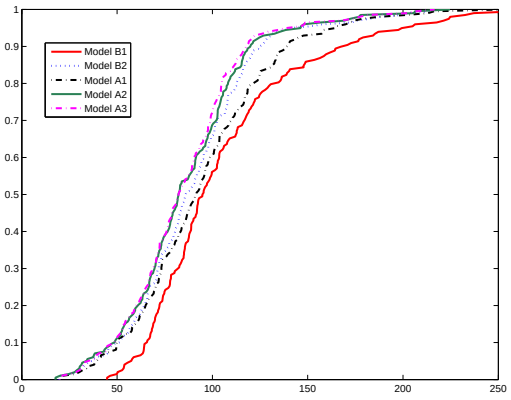

Figure 15: ECDF's for the RMSE's for a forecasting lead time of 1 day, across all agents in cohort C.

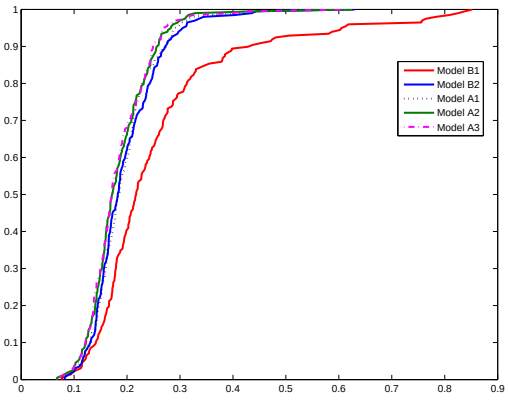

Figure 16: ECDF's for the MAPE's for a forecasting lead time of 1 day, across all agents in cohort C.

6 Simulation experiments

In the previous sections, we illustrated the improvement in goodness-of-fit to data, both in-sample and out-of-sample, which results from considering more realistic service-time models. We now discuss the results of simulation studies which assess the impact of considering different service-time models on performance measures of interest in a call center. Specifically, we consider both the average waiting time (AWT) of calls and the service level (SL(s)), defined as the percentage of calls whose waiting times are less than s seconds (for fixed s).

6.1 Performance impact of the different service-time models

Our objective in this subsection is to quantify the differences in *average* system performance across our alternative service-time models. As such, we show that effectively modelling service times also matters from an operational point of view. To do so, we consider a model of a small subset of the real HQ call center from which our data was taken.

We consider two call types and two agent groups, and an N-queueing design for call routing (1 flexible and 1 dedicated agent pools). The two call types are for the same type of service (related to billing), but one is in French (F) while the other is in English (E). The first group (F, with 10 agents, numbered from 1 to 10) can only handle the first call type and the second group (EF, with 2 agents, numbered 11 and 12) can handle both. That is, F agents only answer calls in French and EF agents are bilingual. Those 12 agents are the ones that worked on each day of week number 45 in our data set, and handled only those two call types. The center is open from 8h to 18h. The arrival process for each of the two call types is piecewise-Poisson, with a gamma random arrival rate in each 15-minute time interval, and a normal copula to model the dependence between those rates. This model is explained in Oreshkin et al. (2015), where it was also shown to provide a good fit to the arrival data for this HQ call center. The arrival rates were scaled down to fit our smaller number of agents. Abandonments are modeled with exponential patience times. Calls are served in first-come-first-served order within each group. All other details, such as the parameters of the arrival-rate model, the abandonment rates, and the detailed staffing and routing rules are given in the supplement (§3).

The target day that we simulate is Friday of week 45. For each selected agent, and each skill for the EF agents, we estimate the parameters of models $B1$, $B2$, $A2$, and $A3$ based on all data collected until Thursday of week 45. (We omit $A1$ because it is outperformed by $A2$ and $A3$.) Using those parameter estimates, we generate service times for each agent on that Friday. For $B1$ and $B2$, this is straightforward. For $A2$, as in (4), we can write, omitting the specific indices for simplicity, that $M = \beta + \gamma + \nu$ where (conditional on past information) β is a constant, γ is normally-distributed with conditional mean $\hat{\gamma}$ and conditional variance v, and ν is normal with mean 0 and variance σ_ν^2/N, where N is the number of calls of that type handled by the agent on that day. To simulate one day, we first generate γ from the appropriate normal distribution and, conditional on γ, we generate independent service times that are lognormal with mean $\beta + \gamma$ and variance σ_ν^2. For $A3$, we proceed in a similar way. In Table 1 of the online supplement, we present the RMSE and MAPE (comparing mean actual and predicted service times) for call types E and F, corresponding to each service-time model. And, in Table 2 of the online supplement, we present the observed average service times for our simulated Friday, the forecasts from each service-time model, and estimated values for σ_ν^2 and v for each agent and skill.

Our model was simulated for $r = 10,000$ independent days under each service-time model with all remaining system parameters held fixed. We purposely choose a large number of simulation replications so as to ensure that the differences observed between our alternative service-time models are statistically significant. We used inversion with common random numbers across the four models to generate the lognormal service times. Thus, the only differences between the four models (in the simulations) are the means and variances of the lognormal service times. We computed the AWT W and the service level SL(120) (the fraction of calls answered within 120 seconds), for each simulated day. In Table 7, we report 95% confidence intervals for $\mathbb{E}[W]$ and $\mathbb{E}[SL]$ based on those r simulations, for each model and call type. Table 7 illustrates potential differences in the SL and AWT across the different models. Although such differences may appear minimal at first glance, they could lead to significant cost savings in practice. For example, ACS Wireless found

Table 7: Performance estimates and confidence intervals for our N model.

	SL (%)		AWT (s)	
Model	F	E	F	E
$B1$	82.68 ± 0.31	56.17 ± 0.41	68.72 ± 1.40	261.80 ± 3.9
$B2$	78.28 ± 0.25	55.31 ± 0.40	73.76 ± 1.30	166.68 ± 2.00
$A2$	79.58 ± 0.23	55.91 ± 0.40	68.20 ± 1.23	160.06 ± 1.98
$A3$	79.26 ± 0.24	54.95 ± 0.40	69.75 ± 1.27	162.42 ± 1.99

that decreasing the AWT by a mere 0.6 second lead to $8 million of savings annually (Hanks 2014). Also, small percent differences in the SL could mean the difference between abiding and violating service-level agreements, which may involve hefty penalties for the call center. *Thus, our numerical results illustrate that different service-time models may lead to different average system performances. Such differences could lead to significant cost reductions in practice.*

Our service-time models could be used in practice to enable a more accurate assessment of the performance of different agents. This, in turn, allows for better agent classification into pools handling different call types. Next, we investigate such selection effects.

6.2 Extending the small example in Gans et al. (2010)

It is intuitively clear that changing the expected service times (or service rates), all else remaining unchanged, can have a strong impact on the performance measures in the system. In simple models, e.g., Erlang-A for a single call type, such performance impact can be computed or approximated easily by a formula. It has also been observed that if agents are heterogeneous in terms of expected service times and a subset of those agents is chosen at random for a given day, then the mean expected service time for the selected agent group has more variability than if all agents were identical (Gans et al. 2010), and the AWT over the day should also exhibit a larger variance and (intuitively) a larger mean.

Gans et al. (2010) showed that agent heterogeneity affects the average waiting times. In this subsection, we go beyond their results and look in more detail on the impact of that heterogeneity on the *distribution of the AWT*, going beyond just the means. In particular, suppose that we have a fixed pool of agents, each with a given skill set and a given service-time distribution for each skill. Suppose that we select, at random, a subset S from this pool, say with a fixed number of agents for each skill set, on a given day. Letting W denote the AWT value for that day, we have that:

$$\mathrm{Var}[W] = \mathrm{Var}[\mathbb{E}[W \mid S]] + \mathbb{E}[\mathrm{Var}[W \mid S]]; \tag{11}$$

in this variance decomposition, the first term is the variability due to the randomness of S, and the second term is the residual variability of W once the set S is known.

Gans et al. (2010) reported an experiment in which they selected 12 agents with exponential service times and service *rates* ranging from 3.86 to 6.33, with an average of 5.015. They simulated 100 independent days as follows. On each day, they picked a set S of 6 agents at random from the original 12, and then simulated a small Erlang-A call center model with Poisson arrivals at a rate of 21 calls per hour, abandonment rate of 2 per hour, and the 6 selected agents. With a service rate of 5.015 for all agents, the Erlang-A formula gives an AWT of 58.8 seconds over an infinite-time horizon. Gans et al. plotted a histogram of the 100 values of W obtained from their simulation and observed a large variability: about 1/3 of the values are more than 12 seconds (20%) away from 58.8. As such, their example illustrates that "a random draw of 6 from the 12 service rates described above will most typically yield results that do not match the intended QoS target" (p. 111).

We now go further: the variability of W has two parts, as shown in (11), and only the first part is due to the random draw of service rates. To estimate the two parts, for each of the $C_6^{12} = 924$ subsets of 6 agents that can be selected, we simulated 1000 days with their model parameters and computed the average

and variance of the resulting 1000 values of W, say M_i and V_i for subset S_i, for $i = 1, \ldots, 924$. We have $\mathbb{E}[M_i] = \mathbb{E}[W \mid S_i]$ and $\mathbb{E}[V_i] = \mathrm{Var}[W \mid S_i]$, so we can view the distribution of the M_i's as an estimate of the distribution of $\mathbb{E}[W \mid S]$. We can also estimate $\mathrm{Var}[\mathbb{E}[W \mid S]]$ by the empirical variance of the M_i's, and $\mathbb{E}[\mathrm{Var}[W \mid S]]$ by the average of the V_i's. We computed those estimates and obtained $\mathbb{E}[W] \approx 64.84$, $\mathrm{Var}[W] \approx 1530.29$, $\mathrm{Var}[\mathbb{E}[W \mid S]] \approx 165.25$, and $\mathbb{E}[\mathrm{Var}[W \mid S]] \approx 1365.04$. Thus, the choice of S only accounts for 10.8% of the variance of W. Nevertheless, there are some choices of S for which $\mathbb{E}[W \mid S]$ differs considerably from $\mathbb{E}[W]$. In particular, if S contains the six fastest agents, then we obtain $\mathbb{E}[W \mid S] \approx 0.21$, whereas if it contains the six slowest agents, then we have $\mathbb{E}[W \mid S] \approx 365.69$. This is indeed a very large spread. *That is, selecting the agents at random does not have a large impact on the variance of the AWT. Nevertheless, selecting specific agent subsets can lead to significantly different values for the AWT.*

6.3 Impact of agent selection

To illustrate the potential impact of agent selection for a given day, we now consider additional simulations for models $B2$, $A2$, and $A3$ (we omit model $B1$ because it assumes that all agents are identical). Here, we summarize our results in Tables 8 and 9; for corresponding histograms of the results, see Figures 7-9 in the supplement. In the previous example, the F agents numbered 1, 2 and 8 (in Table 2 of the supplement) are slowest and agents 3, 7, and 10 are the fastest (according to actual mean service times). We replace the three slowest agents by clones of the three fastest agents, so that we now have two copies of agents 3, 7, and 10. With this new staffing, we simulated $r = 10,000$ independent days and computed W and SL for each day, as in the previous example. In Table 8, we report 95% confidence intervals for $\mathbb{E}[W]$ and $\mathbb{E}[SL]$ based on these r simulations, for each model and each call type. We find that, in comparing with Table 7, performance has improved significantly in the system, across all service-time models. We also note in passing that the differences between the performances according to the alternative models is greater under this choice of agent pool.

Suppose now that we replace EF agent 11 by a clone of (the slower) agent 12 in our original example, and we repeat the same experiment. In Table 9, we present the results for this case. Comparing Table 9 with Table 7 shows that system performance is significantly degraded, despite changing only one agent. *Thus, our numerical results show that selecting specific agent groups with different processing speeds, based on our service-time models, can lead to significant differences in system performance.*

Table 8: Performance estimates and confidence intervals for our N model, with faster agents.

	SL (%)		AWT (s)	
Model	F	E	F	E
$B2$	83.25 ± 0.28	59.87 ± 0.37	57.57 ± 1.03	147.17 ± 1.80
$A2$	81.94 ± 0.31	58.38 ± 0.40	60.65 ± 1.11	151.50 ± 1.8
$A3$	85.38 ± 0.26	60.46 ± 0.37	48.89 ± 0.95	143.80 ± 1.69

Table 9: Performance estimates and confidence intervals for our N model with two slow EF agents.

	SL (%)		AWT (s)	
Model	F	E	F	E
$B2$	76.54 ± 0.36	52.68 ± 0.40	78.89 ± 1.38	175.86 ± 2.02
A2	76.88 ± 0.36	52.68 ± 0.40	78.89 ± 1.38	175.86 ± 2.02
$A3$	76.13 ± 0.26	50.78 ± 0.36	76.48 ± 0.9	182.28 ± 1.65

7 Concluding remarks

In this paper, we took a data-based approach to modelling service times in call centers. We evaluated the goodness-of-fit to data, both in-sample and out-of-sample, of several service-time models. Our models incorporate several properties commonly observed in practice, such as: (1) agent/call type heterogeneity, (2) a time-dependent performance of agents, (3) the existence of cross/ serial correlations in the data. In general, we found that models which exploit those properties are superior to models which do not. To demonstrate the added benefit of that improved goodness-of-fit, we presented and discussed results of simulation experiments which showed that: (1) selecting different service-time models may have a significant impact on average system performance, potentially leading to significant cost cuts, and (2) our service-time models may be used to aid in agent classification into different pools, and system performances under different pools can be drastically different.

Given the promising results that we obtained using Models $A2$ and $A3$, one possible direction for future research is to consider alternative similar models which incorporate daily, or intra-daily, random effects when modeling individual service times. For these models, we may also experiment with nonparametric functions for the trends. To do so requires access to a detailed call-by-call data set. Given the results of § 5, we anticipate that such models will lead to increasingly accurate predictions of future mean service times in the system. With a detailed call-by-call data set, it would also be possible to test for the goodness of fit and predictive accuracy of those models beyond the mean service time. That is, we could test how well those models fit the entire distributions of individual service times in the system. Indeed, complex operational decisions in call centers rely on having models that accurately fit and predict those distributions. We leave developing such models, testing them with data, and assessing their operational performance via simulation, to future research.

References

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

Aldor-Noiman, S., Feigin, P.D. and A. Mandelbaum. 2009. Workload forecasting for a call center: Methodology and a case study. The Annals of Applied Statistics, 3(4), 1403–1447.

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

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

Armony, M. and A. Ward. 2010. Fair dynamic routing in large-scale heterogeneous server systems. Operations Research, 58(3), 624–637.

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

Avramidis, A.N. and P. L'Ecuyer. 2005. Modeling and simulation of call centers. Proceedings of the 2005 Winter Simulation Conference, IEEE Press, 144–152.

Avramidis, A.N., Chan, W., Gendreau, M., L'Ecuyer, P., Pisacane, O., 2010. Optimizing daily agent scheduling in a multiskill call centers. European Journal of Operational Research, 200(3), 822–832.

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

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

Chan, W., Koole, G. and P. L'Ecuyer. 2014. Dynamic call center routing policies using call waiting and agent idle times. Manufacturing & Service Operations Management, 16(4), 544–560.

Colladon, A.F., Naldi, M. and M.M. Schiraldi. 2013. Quality management in the design of TLC call centres. International Journal of Engineering and Business Management, 5(48), 1–9.

Deslauriers, A. 2003. Modélisation et simulation d'un centre d'appels téléphoniques dans un environement mixte. Master's thesis, Dept. Computer Science and Operations Research, Université de Montréal, Montréal, Canada.

Delasay, M., Ingolfsson, A., and B. Kolfal. 2015. Modeling load and overwork effects in queueing systems with adaptive service rates. Working paper.

De Véricourt, F. and Y.P. Zhou. 2006. On the incomplete results for the heterogeneous server problem. Queuing Systems, 52(3), 189–191.

Dong, J., P. Feldman, and G. Yom-Tov. 2015. Service systems with slowdowns: Potential failures and proposed solutions. 2015. Operations Research, 63(2), 305–324.

Feldman, P., J. Li, Yom-Tov, G. and E. Yom-Tov. 2015. Service time sensitivity to load: Who is to "blame"? Working paper.

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

Gans, N., Liu, N., Mandelbaum, A., Shen, H. and H. Ye. 2010. Service times in call centers: Agent heterogeneity and learning with some operational consequences. A Festschrift for Lawrence D. Brown, IMS Collections, 6, 99–123.

Gurvich, I. and W. Whitt. 2009. Queue-and-idlness-ratio controls in many-server service systems. Mathematics of Operations Research, 34(2), 363–396.

Hanks, J. 2014. How the right headset affects call center productivity and the bottom line. Available online at http://telecom.hellodirect.com/docs/Tutorials/Productivity.1.080701.asp (Accessed July 2015).

Ibrahim, R., L'Ecuyer, P., Régnard, N. and H. Shen. 2012. On the modeling and forecasting of call center arrivals. Proceedings of the 2012 Winter Simulation Conference, IEEE Press, 1–12.

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

Liu, Y. and W. Whitt. 2011. Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment. Queueing Systems, 67, 145–182.

Mandelbaum, A. and S. Zeltyn. 2010. Service engineering: Data-based course development and teaching. INFORMS Transactions on Education, 11(1), 3–19.

Mandelbaum A., Massey W.A., Reiman M.I., and A. Stolyar. 1999. Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. Proceedings of the 37th Allerton Conference, Monticello, IL, 1095–1104.

Mehrotra, V. and J. Fama. 2003. Call center simulation modeling: Methods, challenges, and opportunities. Proceedings of the 2003 Winter Simulation Conference, IEEE Press, 135–143.

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

Oreshkin, B., L'Ecuyer, P. and N. Régnard. 2015. Rate based arrival process models for modeling and simulation of call centers. Manuscript, DIRO, Université de Montréal.

Pang, G. and W. Whitt. 2012. The impact of dependent service times on large-scale service systems. Manufacturing and Service Operations Management, 14(2), 262–278.

Pichitlamken, J., Deslauriers, A., L'Ecuyer, P. and A.N. Avramidis. 2003. Modeling and simulation of a telephone call center. Proceedings of the 2003 Winter Simulation Conference, IEEE Press, 1805–1812.

Pinedo, M., Seshadri, S. and J.G. Shanthikumar. 1999. Call centers in ?nancial services: Strategies, technologies, and operations. In E.L. Melnick, P. Nayyar, M.L. Pinedo, and S. Seshadri, editors, Creating Value in Financial Services: Strategies, Operations and Technologies, Kluwer.

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

Rosenblatt, M. 1952. Remarks on a multivariate transformation. Annals of Mathematical Statistics, 23(3), 470–472.

Salcedo-Sanz, S., Naldi, M., Pérez-Bellido, A. M., Portilla-Figueras, J.A. and E.G. Ortíz-García. 2010. Evolutionary optimization of service times in interactive voice response systems. IEEE Transactions on Evolutionary Computation 14(4), 602–617.

Shen, H. and L. Brown. 2006. Nonparametric modelling of time-varying customer service times at a bank call center. Applied Stochastic Models in Business and Industry, 22(3), 297–311.

Weinberg, J., Brown, L.D. and J.R. Stroud. 2007. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. Journal of American Statistical Association, 102(480), 1185–1199.

Whitt, W. 2002. Stochastic-Process Limits. Springer-Verlag, New York.

Zhan, D. and A.R. Ward. 2013. Routing to minimize waiting and callbacks in large call centers, Manufacturing and Service Operations Management, forthcoming.