IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 99–123

DOI: 10.1214/10-IMSCOLL608

Service times in call centers: Agent heterogeneity and learning with some operational consequences

Noah Gans[1,*], Nan Liu[2,†], Avishai Mandelbaum[3,‡], Haipeng Shen[4,†] and Han Ye[4,†]

University of Pennsylvania, Columbia University, Technion - Israel Institute of Technology, and University of North Carolina at Chapel Hill

Abstract: Telephone call centers are data-rich environments that, until recently, have not received sustained attention from academics. For about a decade now, we have been fortunate to work with our colleague, mentor and friend, Larry Brown, on the collection and analysis of large call-center datasets. This work has provided many fascinating windows into the world of call-center operations, stimulating further research and affecting management practice. Larry's inexhaustible curiosity and creativity, sharp insight and unique technical power, have continuously been an inspiration to us. We look forward to collaborating with and learning from him on many occasions to come.

In this paper, we study operational heterogeneity of call center agents. Our proxy for heterogeneity is agents' service times (call durations), a performance measure that prevalently "enjoys" tight management control. Indeed, managers of large call centers argue that a 1-second increase/decrease in average service time can translate into additional/reduced operating costs on the order of millions of dollars per year.

We are motivated by an empirical analysis of call-center data, which identifies both short-term and long-term factors associated with agent heterogeneity. Operational consequences of such heterogeneity are then illustrated via discrete event simulation. This highlights the potential benefits of analyzing individual agents' operational histories. We are thus naturally led to a detailed analysis of agents' learning-curves, which reveals various learning patterns and opens up new research opportunities.

1. Introduction

A call center is the common name for a service operation over the phone. It consists of groups of people, called agents or customer-service representatives (CSRs), who provide the service while sharing telecommunications and information technology infrastructure. The current paper focuses on "inbound" call centers, which take

*Research supported in part by National Science Foundation (NSF) grant CMMI-0800645.
†Research supported in part by National Science Foundation (NSF) grant CMMI-0800575.
‡Research supported in part by the Israeli Science Foundation (ISF) grant 1357/08, and the U.S.-Israel Binational Science Foundation (BSF) grant 2008480.

[1]Department of Operations and Information Management, The Wharton School, University of Pennsylvania, e-mail: gans@wharton.upenn.edu
[2]Department of Health Policy and Management, Mailman School of Public Health, Columbia University, e-mail: nl2320@columbia.edu
[3]Department of Industrial Engineering and Management, Technion - Israel Institute of Technology, e-mail: avim@tx.technion.ac.il
[4]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, e-mail: haipeng@email.unc.edu; hanye@email.unc.edu

calls that are initiated externally. Examples include phone operations that support hotel, airline, and car-rental reservations and sales, service centers of retail banks and brokerages, as well as various emergency services.

1.1. Measuring and modeling operational performance

Performance of inbound call centers varies predictably and stochastically. *Predictable* variability refers, for example, to variability over time of the mix of types of calls seeking service, the arrival rate of each call type, and the average duration of service (service-time) provided to served calls. *Stochastic* variability captures unpredictable deviations around predictable averages. For example, service times are modeled as random variables, and the number of arriving calls over short periods of time (seconds or minutes or hours) is typically taken to be a Poisson process or one of its relatives, e.g. a Cox process (also called Doubly or Compound Poisson), which is used to accommodate over-dispersion (relative to Poisson) that is often encountered in practice. The significance of predicable vs. stochastic variability is quantified by coefficients of variation (noise-to-signal ratios). For example, over longer periods of time (days, weeks, months), the numbers of arrivals become more predictable, in the sense that means become increasingly accurate representations of reality (formally, coefficients of variation diminish).

Operational performance measures are roughly divided into *efficiency* measures, e.g. occupancy levels of agents, and measures of operational *service-quality*, e.g. abandonment rates of customers and waiting-time measures (average waiting time, fraction waiting more than 20 seconds). These performance measures are, of course, an outcome of the number of agents that is handling calls, which is referred to as *staffing* or capacity level; the latter, in turn, is typically recommended, or determined, by Workforce Management (WFM) systems, which are software packages that help track historical arrivals, forecast future demand, and staff and schedule the workforce accordingly.

1.2. Designing performance

Capacity-planning and workforce scheduling is a process of central importance in call centers. (It is commonly agreed that over two-thirds of the costs of running a call center stems from agent salaries.) This process is complex, and hence it is typically "divided-and-conquered" in four hierarchical steps [15]. First, historical arrival data are used to generate point forecasts of the numbers of calls that are expected to arrive in each time-interval (e.g. 1/2 hour) of the day, over some planning horizon (e.g. one week to a month). Second, the WFM system uses *queueing* models to determine staffing levels required for each time-interval within its planning horizon. Third, work schedules (e.g. shifts) are chosen so that the number of agents on hand, during each interval, does not deviate too much from the staffing requirement. Fourth, individual agents are assigned to schedules through a rostering process. (More complete description of the planning process, as well as of call-center operations and research in general, can be found in [18] and [2].)

Traditionally, the staffing procedure performed in Step 2 (for each time-interval) uses the so-called Erlang-C (or M/M/N) queueing model, which imposes the following three assumptions: (1) the call-arrival process is Poisson, in particular, customers are treated as being homogeneous; (2) service-times are independent and identically distributed (iid) according to an exponential distribution, thus agents

are also assumed to be homogeneous; and (3) arriving callers never abandon the queue, and hence all are ultimately served. [19] have shown that the neglect of customer abandonment is a serious shortcoming, and they propose, as an alternative, the so-called Erlang-A (or M/M/N+M) model. This model acknowledges the abandonment phenomenon by adding to the Erlang-C assumptions that customers are equipped with finite (im)patience, the duration of which is also exponentially-distributed, iid across customers.

The simple Erlang-A model has been proved useful, in some circumstances, for capturing the far more complex reality of call centers (Section 7.3 in [14]). Still, its predictive power is severely limited by its postulates of exponentiality and homogeneity. Such assumptions are imposed for merely mathematical tractability, but they hardly prevail in practice. To wit, the pioneering empirical analysis of [14] reveals that arrival-rates are time-varying (with variations that can hardly be assumed constant over 1/2 or even 1/4 hours), service-times are log-normally distributed (as opposed to being exponential), and patience has a non-monotone hazard rate (e.g. peaking at instances of a voice-message, as opposed to a constant). Now, the accuracy of a model's prediction improves with the validity of its building blocks. It is hence a researcher's challenge to improve the latter, to which we contribute here. We do that in the context of heterogeneity and learning (or forgetting) among agents, as it is manifested through their service times.

1.3. Work on heterogeneous servers

1.3.1. Erlang models and beyond

Both Erlang models assume that service times are independent and identically distributed, within each interval during the planning horizon. But, as already mentioned, this assumption typically fails to hold in call centers. Agents are humans with varying experience and skills, which (along with many other factors) affect their performance (productivity). For example, [24] illustrates a "learning-curve" phenomenon: agents become faster in their jobs as they gain experience. [40] reports a "shift fatigue" effect, in which "operators may initially work faster during periods of overload to work off the customer queue, but may tire and work slower than usual if the heavy load is sustained and if no relief is provided." [14] show that service times at one call center are positively correlated with arrival rates: the higher the arrival rate, the longer the service times. And [35] demonstrate that service times are also positively correlated with individual patience: the longer the service time the longer the customer is willing to wait for service.

1.3.2. Pooling and skill-based routing

Queueing theory has long dealt with heterogeneous servers. Interesting examples arise in the context of pooling servers; for example, combining separate queues, each served by a single server, into a single queue which is served by all these servers. It turns out that server heterogeneity can reduce the operational benefits of pooling to a level that pooling is actually harmful (which explains the existence of express lines in supermarkets). Readers are referred to [30] for details and further references.

A recent line of queueing research has been motivated by Skills-Based Routing (SBR) in call centers, which is the protocol for online-matching of multi-type customers to *multi-skilled* agents, the latter constituting a heterogeneous workforce.

[28] describes SBR in the U.S. Bank that is our data source: the tables and flowchart on pages 13-14 of that document provide a convincing representative example of workforce heterogeneity within today's call centers. (Readers are referred to [18] for details on SBR and its operational complexities.)

The importance of SBR goes beyond its operational benefits (e.g. reducing waiting times). In fact, it has been welcomed in call center circles mostly because it enables the creation of career-paths for agents, hence job-enrichment, which is essential in the monotonous pressured high-turnover environment of call centers. Thus, SBR technology goes hand in hand with agent heterogeneity and learning. [2] surveys this multi-disciplinary perspective, to which readers are referred for SBR-motivated queueing research. More recent relevant work is [42, 6, 22] and [23].

Further recent research by Atar and colleagues [9, 11, 10] models agent heterogeneity by letting service times have an exponential distribution with rates being iid across servers. These rates are fixed at the outset, which provides a random environment with heterogeneous servers. The challenge here is to design and analyze *blind* SBR strategies, blind in the sense that no knowledge of system parameters (most importantly arrival and service rates) is required for application of the protocol. Another example of a blind strategy is the List-Based strategy of [32], which blindly gives preference to fast servers over the slow ones. Yet another example is [39], who propose Value- and Preference-Based Routing: these are protocols in which agents are allowed to express their heterogeneity through preferences for serving specific types of customers, preferences that are then incorporated into the SBR protocols. A final queueing example, from the very different environment of hospitals, is [44]: the subject here is fairness dilemmas that arise from heterogeneity of length-of-stay in hospitals (beds playing the role of servers), in the context of routing patients from emergency departments to hospital wards.

1.3.3. Why exponential service times - QED convenience

Significantly, most SBR-driven queueing research (and all of the above) has confined attention to *exponential* service times. To elaborate, SBR presents a problem that is intractable in closed form, hence one resorts to approximations for tractability and insights. These approximations are based on an asymptotic analysis in a specific many-server operational regime, mathematically founded by [25] for Erlang-C, and then adapted to Erlang-A and call centers in [19]. This regime has been referred to as the QED regime; here QED stands for Quality- and Efficiency-Driven, reflecting the ability to achieve, due to economies of scale (many servers), *both* very high levels of server-efficiency (occupancy levels) and customer-service (short delays).

While the state of servers with exponential service times can be summarized by merely the number of busy servers, say, due to the memoryless property of the exponential distribution, the situation is far more complex in the non-exponential case. Here, the exact state of each individual server must be maintained (e.g. the elapsed service time), which gives rise to a state-space dimension that increases with the number of servers, and hence requires far more complex models (e.g. measure-valued diffusions in the limit). The relaxation of service-time exponentiality is a central challenge for current research. For further leads see [29], which is the general-service time analogue of [19].

1.3.4. A process-view of service: duration and structure

Beyond the above theoretical results, little is known about the nature, magnitude and consequence of server heterogeneity. As suggested by [18], "many sources of systematic variation in service times should be better described and analyzed." For example, what drives the duration of service times? How should one model these driving factors? Why are service-times log-normally distributed (which we demonstrate below)? How should a manager take servers' heterogeneity into account? What if a manager ignores server heterogeneity?

A framework for addressing some of these questions has been proposed by [27]. In that work, the prevalent view of service-time, as a static entity, is altered in favor of viewing the customer-agent interaction as a (dynamic) *process* or, formally, the evolution of a finite-state continuous-time absorbing Markov process. The service process then has two attributes - *duration* and *structure*: its duration, namely service-time, is the time-until-absorption of the corresponding Markov process; its structure is then captured by the sequence of states visited until absorption. This renders the distribution of service-time as Phase-Type [8], where a visit at a state corresponds to a phase in the service process. Interestingly, "Phase-type distributions are the computational vehicle of much of modern applied probability" (page 215 in [8]); but their history in Statistics has been that of recurring, independent rediscoveries of their virtues, with the work of Aalen and colleagues [1] standing out as an outlier.

The process-view of service is advocated in [31]. It describes a course, entitled "Service Engineering", that has been taught at the Technion for over a decade. Section 4.5.2 of that paper, corresponding to a 3-hour lecture in the course, is devoted to the service process, which includes (Figure 20 there) a data-based phase-type model of a telephone call. Such a model constitutes a natural framework for addressing questions that pertain to the fine structure of the service process: for example, identifying phases that have little added-value and quantifying the effects of eliminating or changing them, cost-analyzing the addition of cross-selling phases, and more. Nevertheless, our present paper is restricted to the duration attribute of the service process, namely service-times.

1.3.5. Why study service times

Customer heterogeneity has received ample attention in both practice and theory. Indeed, Skills-Based Routing in call centers, as well as Customer Relationship Management (CRM) (e.g. [47]) and Revenue Management (e.g. [41]), can all be viewed as the study and exploitation of customer heterogeneity. In contrast, server heterogeneity has received relatively scarce attention. And, in particular, research inspired by call centers has not addressed the statistical and practical implications of service-time heterogeneity among agents. This paper aims to take a first, small step towards narrowing this gap, a step that further motivates new research problems.

Are service times worthy of the attention that we advocate? The answer is an emphatic "yes," from both the practical and theoretical view-points. Quoting managers of large call centers (with 1000's of agents), the economic saving of decreasing (or cost of increasing) the average service time by 1 second can be millions of dollars annually. It is thus no wonder that managers of call centers exercise a very tight control over the durations of their agents' service encounters. Furthermore, ser-

vice times trigger wonderful research challenges, as we hope that this paper clearly brings out.

1.4. Our contribution

1.4.1. Overview

We use detailed transaction-level data from a network of call centers to investigate actual heterogeneity among agents. We perform an empirical analysis to characterize and quantify important aspects of agent heterogeneity: learning effects, agent-by-agent differences, shift fatigue, system congestion, etc. The analysis is carried out at the call-by-call scale to identify short-term effects, as well as the day-by-day scale to understand long-term factors. We illustrate the operational effect of agent heterogeneity on system performance using discrete event simulation. We also consider four learning-curve models of long-term agent learning and compare their success in predicting agent service rates.

1.4.2. Covariates

The covariates that we have access to, all of which are operational in nature, can explain only a small part of the variability embodied in individual service times. Having additional covariates (such as agents' professional histories, team memberships, etc.) will certainly improve the situation. However, even then, it is our belief that call-by-call variation (due to call content) will still dominate these covariate effects.

Nevertheless, a number of short-term (operational) covariates are associated with significant differences in mean service times. Specifically, higher abandonment rates, as ancillary measures for system congestion, are associated with increased mean service times. For calls that involve multiple segments (or agents) in the service process, the initial segment has a shorter average service time. Next, average call times are shorter for calls that are terminated by agents. (The desired norm is for calls to be terminated by customers; but see [14], who expose the phenomenon of agents hanging-up on customers.) Finally, most types of transferred calls take longer to serve, on average (with the exception of consumer-loan calls).

1.4.3. Learning

Agent learning is a significant factor that is associated with long-term reduction in average service times. Conversely, after days- or months-long breaks (vacations) in call-handling, average service times systematically increase. This may reflect agents forgetting their skills, thus needing to re-learn. In some cases, it may also reflect a change in the service type handled by a given agent. The analysis also shows that there exists significant heterogeneity across agents, both in initial average service time and in the rate of learning.

A detailed analysis of agents' learning curves reveals that agents exhibit a variety of learning patterns, and among four models tested, a nonparametric spline provides the most robust means of modeling agents' learning behavior across the different patterns. Furthermore, the spline model is most sensitive to local changes in the pattern of service rates. Operationally, discrete event simulation shows that ignoring potential agent heterogeneity may lead to quality of service (QoS) performance that deviates significantly from its designated target.

1.4.4. Paper contents

The rest of the paper is organized as follows. We first describe our call-center data in Section 2. We perform some empirical analysis on a cohort of 21 agents, in Section 3, and report our findings about agent heterogeneity. These results motivate us to perform a simulation study, in Section 4, that reveals some operational effects of ignoring agent heterogeneity. They also prompt us to study more carefully agent learning, in Section 5, which is here based on a sample of 129 agents, all of whom are just starting work as CSRs at the center (and are disjoint from the 21 agents mentioned above). Throughout our study, we identify interesting phenomena that motivate additional research problems - several of these are discussed in Section 6, which concludes the paper.

2. Data background

Our data come from a bank that operated a network of four call centers in the Northeast USA. The data archive calls that were connected to the centers from March 2001 through October 2003. As of April 2001, the four call centers handled 200,000-270,000 calls during weekdays, 120,000-140,000 calls on Saturdays and 60,000-100,000 calls on Sundays. About 80% of these calls completed their service process in Voice Response Units (VRUs), leaving 20% of the calls to be also handled by agents. During the period in which our data were collected, about 5,000 agents worked in these call centers. To sketch a temporal distribution of the workforce, 900-1,200 agents worked on weekdays and 200-500 worked on weekends. Among these agents, the busy ones handled on the order of 50-100 calls each day.

A comprehensive description of the U.S. Bank data-base appears in [16]. The data are what we call *transaction-level operational data* [43]. For each call, the data record its whole event-history, including the arrival, waiting and service (if served) phases. A call served by agents may contain multiple segments, during which single or multiple agents provide the sought-after services. Our data record the IDs that agents used to log into the system, and these IDs typically identify the agents uniquely over the analyzed period. This allows one to extract all the call segments that are served by a particular agent, thus enabling the analysis reported in the sequel.

Our data do not contain accurate agent demographic information, nor do they include human resource records concerning agent hiring/training, previous experience, break/vacation, or agent cohort/supervisor. We can foresee that many of these covariates may be relevant for studying agent heterogeneity. And indeed, as will be clarified later, these data limitations naturally restrict the scope of our analysis and interpretation of the results. We have thus been in constant pursuit of such non-operational agent characteristics, and it is plausible that a call center that has recently partnered with the Technion SEELab will provide us the data we desire.

3. Empirical analysis of agent service times

3.1. Call-by-call analysis

As a pilot study, we focus on a cohort of 21 agents who, to the best of our judgment, all started working on July 1, 2002, at two different sites: "S" and "U." The agents specialized in several different service types, such as retail banking, consumer loans

	Site	Service	Start	ID	Term
1	U	Retail	01-Jul-02	4115	28-Jan-03
2	U	EBO	01-Jul-02	4122	
3	U	Retail	01-Jul-02	4128	06-May-03
4	U	Retail	01-Jul-02	4130	
5	U	On-Line	01-Jul-02	4136	04-Apr-03
6	U	On-Line	01-Jul-02	4151	31-Dec-10
7	U	Loans	01-Jul-02	4235	26-Jul-03
8	U	Loans	01-Jul-02	4243	
9	U	Loans	01-Jul-02	4254	
10	S	Retail	01-Jul-02	6735	
11	S	Retail	01-Jul-02	6737	
12	S	Retail	01-Jul-02	6738	
13	S	Retail	01-Jul-02	6739	
14	S	Retail	01-Jul-02	6740	
15	S	Retail	01-Jul-02	6741	
16	S	Retail	01-Jul-02	6744	
17	S	Retail	01-Jul-02	6747	
18	S	Retail	01-Jul-02	6764	
19	S	Retail	01-Jul-02	6909	
20	S	Retail	01-Jul-02	6911	
21	S	Retail	01-Jul-02	6912	

FIG 1. *Pilot population of 21 agents.*

and online banking, and only a few left the bank's employment ("Term") before the data collection period ended, on October 26, 2003. Figure 1 summarizes this information for our ad hoc cohort.

While traditional capacity planning in call centers assumes that service times within a given interval are independent and identically distributed, most often according to an exponential distribution, the service times in this group are not. As is consistent with [14], these data reflect lognormally-distributed service times.

Figure 2 displays a histogram of $\log_{10}$(service times) of the calls handled by Agent 4115 at site U (referred to as Agent 14115 henceforth), together with a corresponding normal quantile plot. The plots suggest that the service times for calls longer than 15 seconds fit well the lognormal distribution. The relatively heavy

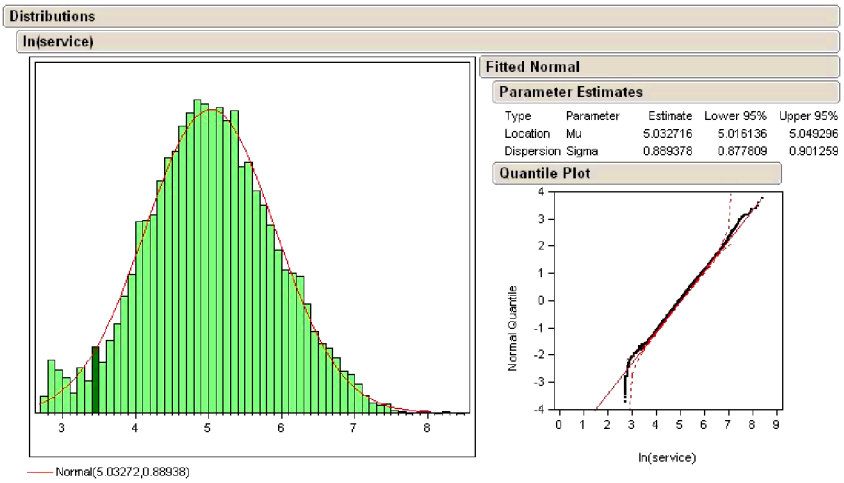

FIG 2. *Agent 14115 - Log(Service Times) for calls longer than 15 seconds.*

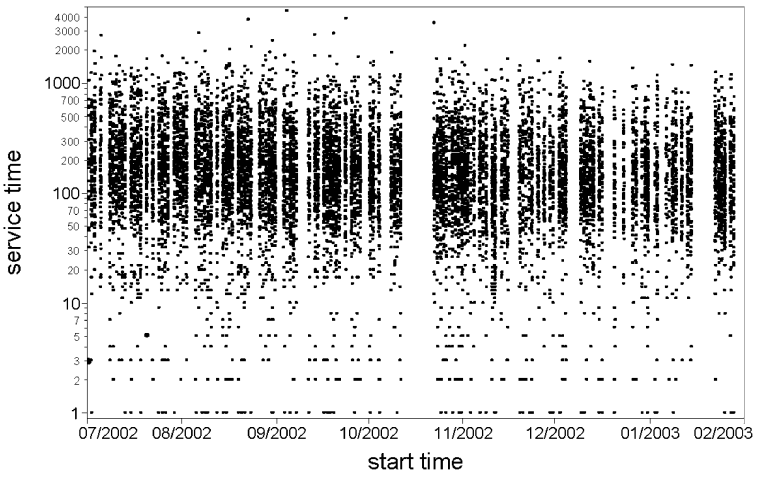

FIG 3. *Agent 14115 - Longitudinal view of service-time evolution.*

presence of short-duration calls (of less than 15 seconds) may be due to system's mis-measurement or, alternatively, peculiar phenomena such as agents hanging up on customers; see [14].

Figure 3 depicts the time series of $\log_{10}$(service times) derived from Agent 14115 (against the dates that the calls started to be served). Note that the data display a very slight but significant (to the eye) downward slope. The fact that their log is *normally* distributed makes the use of lognormally distributed service times appealing for statistical analysis. We are able to develop careful regression analyses of log(service time) on a variety of factors. Below we fit a traditional learning curve model [48] to the data:

$$\log(y_k) = a + b\log(k) + \epsilon_k, \tag{1}$$

where $k \in \{1, 2, \dots, \}$ indexes chronologically each of the calls handled throughout the agent's tenure, y_k denotes the service time of the kth call, and $\epsilon_k \sim N(0, \sigma^2)$. In fitting the above regression model, we have omitted the first 2 weeks of call times (758 out of 11,500 calls).

Not surprisingly, regression results for this model yield $R^2 = 0.005$; the great bulk of call-by-call variability is not explained by learning. Indeed, most of the variation in call times remains unexplained, even when other covariates are included in the regression. We presume that this phenomenon is driven by the fact that call times are driven by the randomly varying work content of various customer questions and requests.

At the same time, the estimate of b is -0.1247 and it is highly significant, with a p-value of less than 0.0001. This coefficient value translates to a systematic decrease of $2^{-0.1247} = 8.3\%$ in expected service time, with each doubling of the number of calls cumulatively handled by the agent. (The underlying calculation is $(e^{a+\sigma^2/2}(2k)^b)/(e^{a+\sigma^2/2}(k)^b) = 2^b$.)

Furthermore, the 8.3% figure is managerially significant. That is, although (after a few weeks of work) the relative number of calls handled in a single day will not be enough to significantly affect the agent's expected service times, cumulative experience does appear to have a significant association with shorter average service times. This is consistent with an earlier, anecdotal report by [24], who claims that the av-

erage call times of AT&T Directory Assistance operators declined systematically over their first 12-18 months' experience. To the best of our knowledge, Gustafson's anecdotal report is the only published empirical evidence of agent learning, prior to the present work.

In addition to learning, there are other factors that might affect service times. [40] reports that agents suffer from "shift fatigue," where they "initially work faster during periods of overload to work off the customer queue, but may tire and work slower than usual if the heavy load is sustained and if no relief is provided." [14] find that service times depend on the time of day as well as the type of service. Another short-term factor is the congestion level at the time of service.

To carefully study the effects of various factors on agent service rate, we consider each of the 21 agents and regress his or her log service times on the following covariates, to which we have access:

- rec_num: n^{th} service the agent has performed over his tenure
- run_length: n^{th} service the agent has performed since the last gap of > 1 hour
- ab_rate: the abandonment rate at the time at which the call is handled
- call_term: call termination type: by the customer, by the agent, or by transfer
- later_seg: 1^{st} or later segment of call for the customer and agent (0/1)

Results from the individual regressions are reported in Figure 4. Observe that, after the table's first two columns, each pair of columns reports the estimate and significance of (the intercept and) a given coefficient. Negative estimates are underlined, and insignificant p-values have shaded backgrounds.

The regression results suggest the following. First, the intercepts that estimate agents' base-line service rates, as well as the "log(rec_num)" that estimate the slopes of the learning curves, are generally significant and vary significantly across agents. Second, the "later segment" coefficient shows that whether or not the agent was the first to work on the call is significant: later segments are associated with increased average log call times. Third, how a call terminates is also significant: "agent term" is associated with decreases in average log service times; and "unknown term" and "transfer term" are associated with increases, except in the case of consumer loans. Here, transfers are associated with decreased average log call times, and we speculate that the more difficult consumer-loan calls are quickly transferred to a specialist group whose call times we do not track. Fourth, higher abandonment rates – a measure of increased system congestion at the time a call is handled – are (usually) associated with statistically significant increases in average log service times, but the effect is not large. Finally, the coefficient for log(run_length), as a proxy for shift fatigue, is mixed in sign and significance.

3.2. Analysis of daily service-time averages

The empirical analysis of call-by-call service times, in Section 3.1, identifies several short-term factors associated with differences in average (log) call times. In this section, we focus on the longer-term evolution of agent service times over days or weeks. For that purpose, we look at daily means of log service times, which average out the effects of short-term factors such as call-by-call variability, time-of-day, shift-fatigue, and system congestion. The analysis reveals longer-term effects, such as learning and forgetting [33].

We now restrict attention to a more homogeneous group of agents, focusing on the 12 retail agents at site S. Figure 5 provides some simple examples of the types of

	agent	intercept	p-value	log (recnum)	p-value	log (run_len)	p-value	ab_rate	p-value
1	14115	5.460	0.000	-0.087	0.000	0.034	0.001	0.019	0.000
2	14122	5.062	0.000	-0.016	0.128	-0.014	0.204	0.018	0.000
3	14128	5.763	0.000	-0.075	0.000	-0.048	0.030	0.019	0.018
4	14130	5.405	0.000	-0.066	0.000	0.028	0.000	0.018	0.000
5	14136	5.251	0.000	-0.059	0.000	0.026	0.002	0.018	0.000
6	14151	5.126	0.000	-0.034	0.001	0.061	0.000	0.017	0.000
7	14235	5.866	0.000	-0.058	0.000	-0.027	0.005	-0.003	0.571
8	14243	5.592	0.000	-0.044	0.000	-0.030	0.000	0.013	0.003
9	14254	5.364	0.000	-0.011	0.142	-0.040	0.000	0.001	0.817
10	26735	4.362	0.000	-0.043	0.000	0.018	0.030	0.007	0.143
11	26737	4.906	0.000	0.002	0.848	0.021	0.015	0.012	0.021
12	26738	5.335	0.000	-0.046	0.000	0.005	0.561	0.017	0.006
13	26739	5.416	0.000	-0.048	0.000	0.004	0.655	0.015	0.000
14	26740	5.203	0.000	-0.046	0.000	-0.006	0.414	0.031	0.000
15	26741	5.147	0.000	-0.052	0.000	0.028	0.000	0.020	0.000
16	26744	5.209	0.000	-0.054	0.000	0.024	0.000	0.015	0.000
17	26747	4.950	0.000	0.023	0.003	0.027	0.001	0.020	0.000
18	26764	5.456	0.000	-0.040	0.000	-0.009	0.400	0.012	0.045
19	26909	5.711	0.000	-0.091	0.000	0.005	0.439	0.019	0.000
20	26911	4.369	0.000	0.087	0.000	-0.022	0.057	0.024	0.000
21	26912	5.703	0.000	-0.078	0.000	0.008	0.212	0.018	0.000

	agent	agent term	p-value	unknown term	p-value	transfer term	p-value	later segment	p-value
1	14115	-0.365	0.000	0.872	0.000	0.273	0.000	0.053	0.650
2	14122	-0.450	0.000	1.069	0.000	0.300	0.000	0.109	0.379
3	14128	-0.538	0.000	0.795	0.000	0.047	0.553	-0.368	0.091
4	14130	-0.775	0.000	0.869	0.000	0.256	0.000	0.302	0.002
5	14136	-1.020	0.000	0.803	0.000	-0.032	0.179	0.126	0.165
6	14151	-1.258	0.000	0.699	0.000	0.091	0.004	0.220	0.068
7	14235	-0.659	0.000	0.365	0.000	-0.190	0.000	0.132	0.000
8	14243	-0.645	0.000	0.460	0.000	-0.110	0.000	0.126	0.000
9	14254	-0.750	0.000	0.376	0.000	-0.131	0.000	0.107	0.000
10	26735	1.087	0.000	1.624	0.000	1.357	0.000	0.562	0.000
11	26737	-1.102	0.000	0.862	0.000	0.144	0.000	0.131	0.114
12	26738	-0.609	0.000	0.952	0.000	0.346	0.000	0.306	0.001
13	26739	-1.227	0.000	0.751	0.000	0.036	0.084	0.023	0.782
14	26740	-0.377	0.000	1.029	0.000	0.213	0.000	0.213	0.010
15	26741	-1.088	0.000	0.855	0.000	0.302	0.000	0.457	0.000
16	26744	-1.197	0.000	0.858	0.000	0.101	0.000	0.376	0.000
17	26747	-0.703	0.000	0.721	0.000	0.167	0.000	0.219	0.015
18	26764	-1.145	0.000	0.774	0.000	0.122	0.000	0.343	0.002
19	26909	-0.449	0.000	0.648	0.000	0.098	0.000	0.286	0.000
20	26911	-0.560	0.000	0.820	0.000	0.241	0.000	0.636	0.000
21	26912	-1.127	0.000	0.666	0.000	0.053	0.006	0.307	0.000

FIG 4. *Regression results for the 21 agents.*

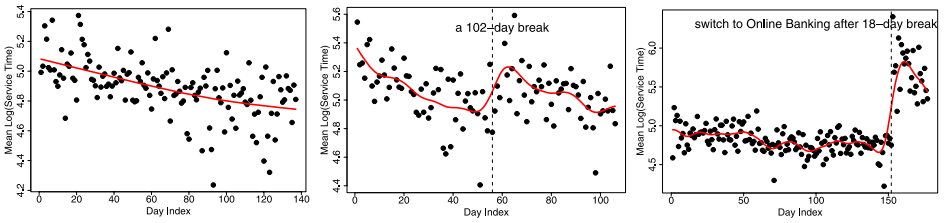

FIG 5. *Long-term trend of daily average service time.*

longitudinal trends that the daily averages of log service-times reveal. Each of the figure's panels shows the day-by-day evolution of service times for a given agent, and each point within a panel plots the daily average of the log(service time) for that agent. We superimpose smooth curves on the panels to highlight underlying patterns.

These plots are representative of a number of phenomena that come out of our preliminary analysis. All three demonstrate systematic, long-term decreases in average (log) service times. This clearly suggests the existence of learning. At the same time, there are differences among the agents' learning rates. The center plot shows an abrupt increase in service times after a long (102 day) absence by that agent, which may reflect a forgetting effect that offsets the gains of learning. The right panel shows an even more pronounced increase in average log service times after a relatively brief (18 day) break. We believe that, in this case, the increase may reflect the agent's transfer to a different type of call.

To study potential differences among the learning behaviors of agents, we omit the call records after these types of breaks. We first use the efficient estimator proposed by [37] to estimate the daily average service time. For each agent, we then fit the following regression model using weighted least squares:

$$\log(m_j) = a + b\log(j) + \epsilon_j, \quad \epsilon_j \sim N(0, \sigma^2/n_j),$$

where m_j is the estimated mean service time for day j, and n_j is the number of calls answered by the agent during that day. The number of calls answered for each day is used as the weight in the regression model, reflecting the fact the response is estimated using that many individual calls.

Figure 6 depicts the daily learning curve for each of the agents, based on the regression results. (Each unit on the x-axis represents one week; hence the time-horizon is 20 weeks.) The learning curves suggest that there are significant variations among the agents and that agent heterogeneity takes two forms: (1) on any given day, there is significant variation among the 12 agents' expected service rates; and (2) the learning-curve parameters themselves, that is, the baseline mean log service time (a) and learning rate (b), are different across agents. The fact that the learning process affects average call-times only gradually, over weeks rather than days, also suggests that one can usefully address short term and longer-term heterogeneity as

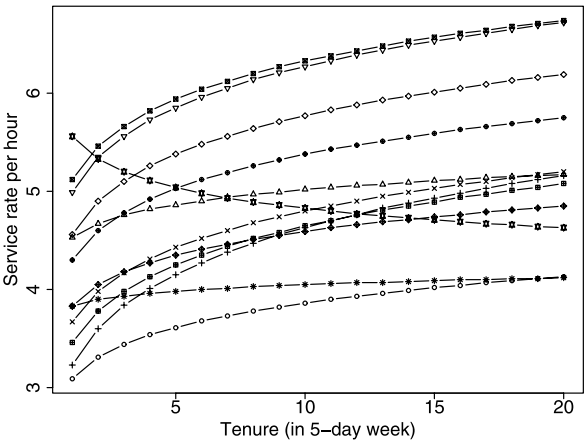

FIG 6. *Daily learning curves of the 12 agents at site S.*

distinct phenomena. More specifically, over a given day, we view an individual agent as having a relatively stable expected service time and treat agent service rates as a set of static parameters. Conversely, over weeks or months, we view agent service rates as evolving according to a heterogeneous set of a's and b's. These learning curves will be used to generate inputs for the discrete-event simulation study in the upcoming section, where we illustrate some short-term operational effects of agent heterogeneity.

4. Effects of agents heterogeneity on operational performance

As outlined in Section 1, call center staffing usually assumes that agents are (statistically) identical, with a common service rate during a staffing interval. In contrast, if we draw a vertical line anywhere along the x-axis of Figure 6, we see that, on any given day, the expected service rate can vary widely among agents. Thus, the misapplication of traditional staffing procedure to agents with heterogeneous service rates can lead to flawed solutions of staffing and scheduling problems.

For example, consider a staffing problem based on the 12 learning curves shown in Figure 6. At the end of week twelve, the estimated service rates per hour are

3.86, 4.05, 4.59, 4.63, 4.65, 4.80, 4.83, 5.02, 5.38, 5.77, 6.27, and 6.33,

with an average of 5.015 services per hour. We shall show that the use of a traditional performance model, which assumes that all agents have a service rate of 5.015, can lead to poor estimates of performance.

Suppose the arrival rate is forecasted to be $\lambda = 21$ calls per hour, and an individual's average time to abandonment is $\theta^{-1} = 30$ minutes. That is, the time at which a customer becomes impatient and hangs up is expected to be $\theta^{-1} = 0.5$ hours, and the average service time is $\mu^{-1} \approx 0.2$ hours. Thus, the average time to abandonment is about 2.5 times that of an average service time (which, perhaps somewhat surprisingly, reflects our practical experience). Then, for statistically identical agents, each with service rate $\mu = 5.015$ calls per hour, results for the M/M/n+M (Erlang-A) model [19] suggest that 6 agents are required to ensure an average delay in queue of 1 minute or less. (In fact, with 6 agents, the estimated delay is 58.8 seconds.) In contrast, 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.

Figure 7 illustrates this effect. The figure's left panel plots the average service rate obtained when 6 of the 12 agents are sampled at random (without replacement). The plot's x-axis marks the sample-average of the 6 sampled service rates, and its y-axis

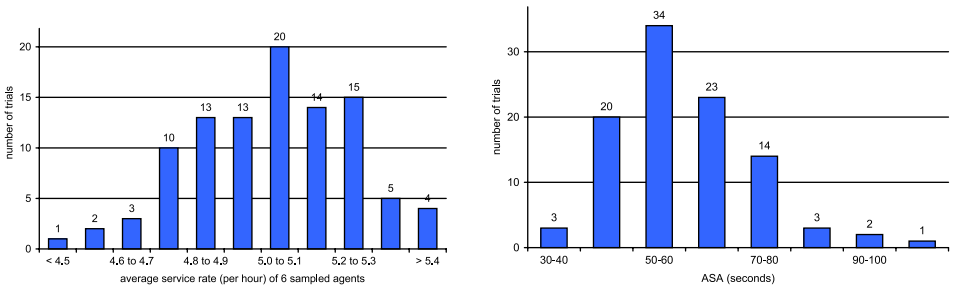

Fig 7. *Average service rate (left) and Approximate ASA (right) obtained from random sampling.*

shows the frequency (in 100 trials) with which various sample results occur. While the mean and median of the trials are close to 5.015, there exists significant variation about this center. The right plot displays an analogous frequency histogram for the average delay in queue, obtained from 100 simulations of the Markovian system that results from $\lambda = 21$, $\theta = 2$, and the six sampled μ's. In each of the 100 trials, we run a discrete-event simulation to estimate the average delay in queue. In the simulations, service is first-come-first-served, and if more than one agent is idle when a call arrives, then the call is routed to the agent that has been idle the longest. (Note that routing to the fastest available agent would systematically improve average delay in queue. See [6].) In 14 of the hundred trials, the actual average delay is more than 20% below the 58.8 second estimate, and in 18 of the cases it is more than 20% above the target.

Now consider a simple scheduling problem with two planning periods, as above, both periods with the same λ, θ and QoS target as before. With our 12 agents, as above, but wrongly assuming that all agents have service rate 5.015, the Erlang-A model would again recommend 6 agents per period. But if we randomly divide the 12 agents into two groups of 6, with each group assigned to one period, then the realized performance may exceed the QoS target in one period and fall short of it in the other. That is, by ignoring service-rate heterogeneity we can encounter both problems, of over- and under-performance, on a single given day.

In actual operations, both the disparity among agents' service rates and the resulting capacity mismatches, may be even worse. Recall that the 12 service rates were sampled from agents who had all started on the same day. In a service operation such as a call center, in which there are many more employees, and in which employee turnover can be both high and highly variable, the disparity among agents' service rates is likely to be even more pronounced. (It is left to discover, however, how the number of agents affects capacity mismatch; after all, 12 agents correspond to an exceptionally small call center.)

Thus, explicit use of individual agents' expected service times should help improve short-term capacity planning. To effectively manage this heterogeneity, one requires two complementary sets of tools: (1) the ability to predict key service-time statistics for each agent available during the planning horizon; and (2) an extension of standard capacity-planning tools to accommodate agent heterogeneity when scheduling. We start to develop the first tool in Section 5 from the learning-curve perspective, and we discuss the development of the second tool in Section 6 as part of our proposed future work.

5. Learning curve modeling

Learning effects have been studied extensively in the past. See [48] for a literature review on this topic. Recent developments include [12, 34, 36, 33], etc. However, this literature focuses on worker learning in the manufacturing/production context, while scarce literature has investigated agent learning in call centers. Our work supplements their work by providing a comprehensive study on the learning curves of a large group of call center agents. To our best knowledge, this work is the first to do so.

As discussed above, we view agents as accumulating experience on a day-to-day basis. In this section, we consider different learning-curve models and compare the performance of their in-sample estimates, as well as the accuracy of their out-of-sample predictions.

5.1. Four learning-curve models

We assume that service times of an agent follow lognormal distributions. For an arbitrary agent, let y_{jk} denote the service time of the kth call during the jth day over this agent's tenure, and n_j be the total number of calls served by this agent during the jth day. Define $z_{jk} = \log(y_{jk})$.

We consider the following three parametric models and one nonparametric model to capture the agent learning effect.

Model 1 $z_{jk} = a + b\log(j) + \epsilon_{jk}$;

Model 2 $z_{jk} = a + b\log(\frac{j}{j+\gamma}) + \epsilon_{jk}, \quad \gamma > -1$;

Model 3 $z_{jk} = a + b\frac{j}{j+\gamma} + \epsilon_{jk}, \quad \gamma > -1$;

Model 4 $z_{jk} = f(j) + \epsilon_{jk}, \quad f(\cdot)$ is a smooth function.

In all four models, the errors are assumed to be normally distributed: $\epsilon_{jk} \sim \mathrm{N}(0, \sigma^2)$, where the homoscedasticity assumption keeps the models parsimonious. The first model adapts the learning curve equation in [48]. Model 2 and Model 3 have a common interesting feature: the mean log service time approaches some limit as the training period approaches infinity. This feature conforms to the conjecture that the service rate eventually stabilizes and will not improve further after the agent has taken calls for a sufficiently long period of time. Unlike the first three models, Model 4 is nonparametric and assumes the least amount of structure on the underlying learning curve. The first three models are estimated via maximum likelihood or constrained maximum likelihood, while the fourth model is estimated via the smoothing spline technique [21].

5.2. Learning patterns of agents

We fit the four learning models to a group of 129 agents, whose records suggest that they are common agents with no previous work experience in the call center. Remarkably, we find that there exists a variety of learning patterns among these agents. They exhibit mainly the following three patterns: (1) always learn, (2) never learn, and (3) learning and forgetting interwoven throughout the whole tenure. The majority of the agents possess the third pattern. For ease of presentation, we name these three learning patterns: the **optimistic** case, the **pessimistic** case and the **common** case, respectively. To illustrate these behaviors, we select two agents from each case and display their estimated learning curves below.

The Optimistic Case: agents always learn

Figure 8 plots the learning curves for two agents: Agent 33156 and Agent 33235, respectively. On both panels of the figure, the x-axis is the duration of the agent's working experience (in days), and the y-axis displays the mean log service time for that day. The dots are the average log service times calculated from the data. More precisely, for day j on the x-axis, the value of the corresponding dot on the y-axis is calculated as

$$\bar{z}_{j\cdot} = \frac{1}{n_j}\sum_{k=1}^{n_j} z_{jk}.$$

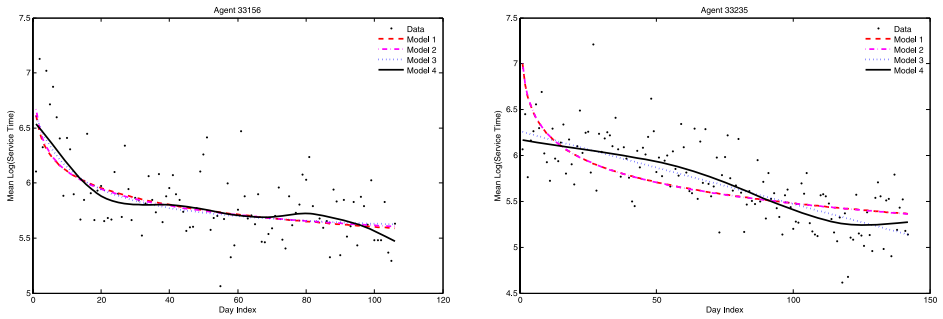

FIG 8. *Learning curves for the optimistic case.*

The four curves show the estimates for the mean log service time, given by the four models that we considered. See the legend within each panel for a detailed description.

From Figure 8, we deduce that, throughout these two agents' tenures, their mean log service times are decreasing. This implies that they are always learning and are getting faster on their jobs as they work longer. However, their learning rates seem to be decreasing, and the learning curve becomes flatter, which suggests that the purely log-log linear learning curve (Model 1) is too simple to capture the underlying behavior.

The Pessimistic Case: agents never learn

Not all agents learn during their working period, and in Figure 9, we show two agents who never learn. As we see, Agent 74527 is getting slower as he works longer, and Agent 76859 seems to maintain a stable service rate throughout her tenure.

From the plot of Agent 74527, one also observes that the nonparametric spline model is more sensitive to the short-term trend of the service rate. In particular, this agent's mean service time has a significant leap around day 130, which is captured nicely by the nonparametric model. The three parametric models are too rigid to fit such a dramatic change.

The Common Case: agents may learn as well as forget

We observe that most agents do not have a monotone learning curve: their log mean service times are "zigzagging" throughout their working period; for such a

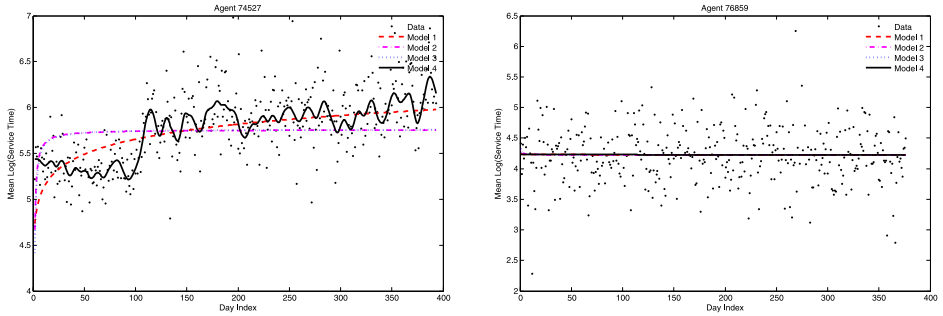

FIG 9. *Learning curves for the pessimistic case.*

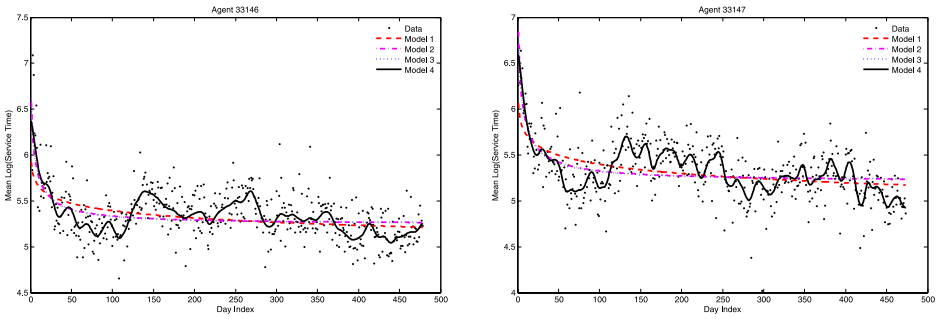

FIG 10. *Learning curves for the common case.*

behavior, the nonparametric model captures much better the trend of the mean log service time. Figure 10 depicts the learning curves of two such agents.

Agent 33146's mean log service time is decreasing during his first 100 working days and afterwards has two significant leaps. The first leap starts at around day 110 and reaches a peak at around day 150. After that, the log service time starts to decrease. The second jump begins at about day 220 and arrives at the apex at about day 270, after which the log service time keeps decreasing until the end.

Agent 33147's learning curve is similar. As we see in the right panel of Figure 10, her mean log service time first drops, then starts to jump at around day 100, reaches its peak around day 130, then begins to decrease slowly until day 240, makes a sharp drop between day 240 and day 280, and finally seems to stabilize, from day 280 onwards.

Based on the above results, we conclude that agents' learning curves can differ significantly. However, we note that the above analysis uses only the service times of the calls; other factors may explain the leaps and bumps observed in the learning curves; however, we do not have access to them.

5.3. Out-of-sample prediction of service rate

As the simulation results in Section 4 suggest, managers need statistical models that can sensitively monitor the service rates of individual agents; otherwise, the call center may end up being overstaffed or understaffed. The analysis in Section 5.2 compares the in-sample performance of four learning models. In addition, we calculate below the out-of-sample prediction errors of the service rate, using the four models.

We incorporate a rolling-window prediction procedure. For a given agent, the schematic algorithm of the out-of-sample prediction exercise is as follows:

Algorithm for Computing Prediction Errors

For $j = 6$ to n, with n being the length of the agent's tenure (in days):

- Fit Model i using data from the 1st day to the $(j-1)$th day, for $i = 1, 2, 3, 4$;
- Predict the mean log service time of the agent on the jth day, denoted as $\hat{z}_{ij}$, using the fitted learning model;
- Predict the mean service rate of the agent on the jth day as $\hat{\mu}_{ij} = e^{-(\hat{z}_{ij} + \hat{\sigma}_i^2/2)}$ where $\hat{\sigma}_i$ is the estimated standard deviation for the measurement error in

Model i;

- Estimate μ_j, the "true" mean service rate of the agent on the jth day, calculating it as the reciprocal of the mean service time of the calls answered on that day;
- Calculate the prediction errors (PE) and relative prediction errors (RPE) of the service rates on the jth day as

$$\mathrm{PE}_{ij} = \hat{\mu}_{ij} - \mu_j, \quad \text{and} \quad \mathrm{RPE}_{ij} = \frac{\hat{\mu}_{ij} - \mu_j}{\mu_j} \times 100\%.$$

End For

After reviewing the out-of-sample prediction performance of the models, we conclude that the nonparametric learning model is more sensitive and effective in monitoring the changes in agent service rate, no matter what the agent's learning pattern. Hence, among the four approaches tested, the nonparametric model is the most robust. To illustrate this observation, below we plot in Figure 11 the prediction errors for Agent 33147. As shown in Section 5.2, this agent has the most intricate learning pattern among the six agents plotted there.

In each panel, the x-axis is the number of historical working days used to fit the learning-curve models for the agent; the y-axis shows either the prediction error or the relative prediction error. The curves for the four models are plotted using different colors, as indicated in the panel legend. From these plots, we observe that, over the full tenure of the agent, the performance of the nonparametric spline model is the best and the most stable. In particular, from Figure 10, we observe that the mean log service time of Agent 33147 has a clear jump from day 100 to day 130. Correspondingly, in Figure 11, we see that the prediction errors of using the spline model are much closer to zero during this time period. In addition, the spline model is also more sensitive to drops in mean log service time (i.e., service rate jumps), during the period between day 240 to day 280 for Agent 33147. These observations imply that the nonparametric model is the most sensitive one in capturing agent service-rate changes.

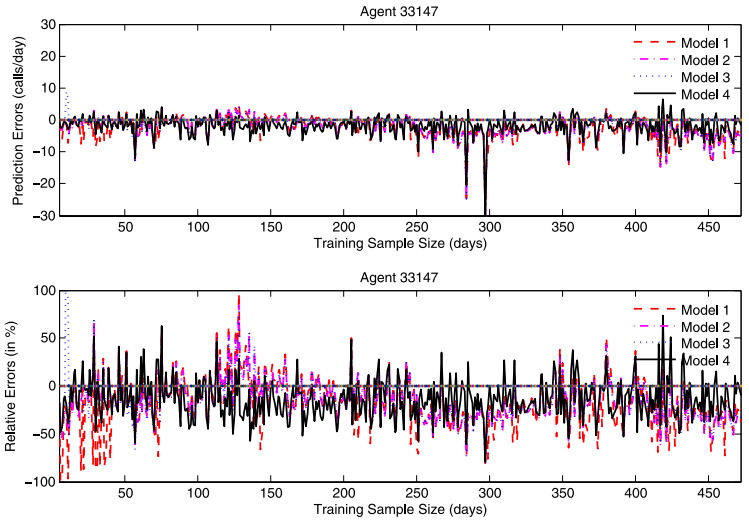

FIG 11. *Prediction errors for Agent 33147.*

6. Conclusion and outlook

In this paper, we studied agent heterogeneity and learning in call centers. We used empirical call-center data to identify several significant factors that affect agent service rate, and we illustrated some operational consequences of ignoring such heterogeneity. But the present paper is only a first step: its empirical analysis should be expanded and the theory it motivates should be addressed. We conclude the paper with an outline of some worthy natural next steps.

Additional data: The data that we have are enough for starting the empirical analysis, which leads to identifying several interesting phenomena, exploring agent learning and generating some insights on agent heterogeneity. The data, however, are not enough for a comprehensive analysis of agent heterogeneity. First, we analyzed only about 140 agents from a single call center - the sample size should increase, and more call centers should be analyzed. But more significantly, we are lacking agents' human-resource records, which are to be linked with our operational data, and which are clearly important for explaining and predicting agent performance. As mentioned, we are hopefully in the process of obtaining such data.

Service-level, efficiency, fairness and learning: The simulation results from Section 4 suggest that, unaccounted for heterogeneity in agent service rates can lead to poor agent staffing and scheduling decisions. Recent work by [5, 42, 9, 6] shows that, given sufficient information, short-term control decisions can actually benefit from agent heterogeneity. For example, routing customers to the fastest idle agents, assuming agent speeds (average service times) are known, could clearly reduce, even optimize, measures that are associated with customer waiting. However, such routing is unfair to the faster servers since they work more than the slower ones - more in the sense of "enjoying" higher occupancy (fraction of busy-time). Thus, a compromise must be struck between fairness towards customers (short delays) and fairness towards agents (low occupancies). This is the subject of [44], who offer tunable protocols that can, for example, ensure that the faster agents work less than slower ones (their occupancy is lower), making faster servers happy, but they also serve more customers (have higher throughput), making management happy. Note, however, that if some slow servers are at the outset of their learning process, higher throughput for them could accelerate their learning and ultimately benefit the organization in the longer-run.

Blind (robust) routing: The above routing schemes are based on knowledge of agent's expected service time, which could be hard to track. To this end, "blind" (robust) strategies are called for, that do not need service-time information yet could compete with the strategies described above. The simplest blind strategy is to choose an idle agent at random. Interestingly, it exposes the *slow-server* phenomenon, where less capacity could actually improve performance: specifically, time-in-system could be reduced if calls are never routed to too-slow servers. (Waiting time, in contrast, will always suffer from capacity reduction; see [44] for further discussion and references.)

Another blind strategy routes customers to the agents who were idle the longest. This was shown in [9] to yield, in the QED regime, equal idleness among heterogeneous agents, who cater to a single queue of homogenous customers. Fairness with shorter waiting times were achieved in [7], but at the cost of knowledge of service rates: the routing policy is a threshold policy that prioritizes faster agents when the number of customers in the system exceeds some threshold level and otherwise prioritizes slower agents. As a summary, a theory is thus required for heterogeneous agents, that trades off service-level, efficiency, fairness (agent preferences)

and learning, and which is based on information that is easily tractable.

Medium-term scheduling: The discussion has been restricted so far to short-term control. Related issues, however, are as interesting and challenging for medium-term scheduling. Specifically, in Step 3 of the traditional capacity-planning hierarchy (Section 1.2), one must schedule enough agents so that their aggregated service rate matches the required capacity, as calculated by the staffing in Step 2. When agents are highly heterogeneous, this scheduling problem can become quite difficult to solve. In the extreme case, in which individual processing rates are accommodated, the problem becomes an assignment problem with many thousands of 0-1 variables, which are required to represent assignments of (100-1000s of) agents to (100s of) possible schedules.

Random effects: Another interesting direction for future research is to incorporate random effects into learning-curve modeling. For example, recall the standard model (1), where a stands for initial speed and b denotes learning rate. Random-effects models would allow either a or b or both to be random variables, the distribution of which is inferred from data. Such models could then incorporate within-agent dependence among the calls handled by the same agent, and enable understanding of a whole agent population, yet based only on a sample (subset) of the agents. In a Bayesian context, the estimated distribution for a or b can serve as the prior for a newly hired or trained agent. Then, as the agent handles calls and gains experience, he moves down his own idiosyncratic learning curve, and his own service records enable improved posterior inferences of his personal a or b. Such information can support the prediction of future performance and the design of career paths, among other things.

Random effects have already been used in closely related work. [26] develop frailty models for customers patience while waiting over the phone (time till they hang up), applying these models to understand how such patience changes with customer redials. Larry Brown and his collaborators have proposed a Bayesian procedure for dynamically updating predictions for intra-day arrival rates, based on continuously updated arrival information [45]. The problem of updating arrival-rate curves was also studied later on by [38] and [20]. We propose to adopt similar ideas in order to update or continue agent learning curves, based on updated information on additional calls handled.

An application of random learning curves is for judging the value of a given agent, say agent i. A manager can then compare her posterior estimates for a_i or b_i against the prior distribution of a or b associated with the population of potential new hires. In some cases, the posteriors may suggest that i is more valuable than a new hire, but in others the posterior may suggest that an untried agent, with the prior distribution, is preferred. In the former case the manager retains the current agent i and in the latter she does not, hiring someone new instead. A fully sequential version of this retention problem has the structure of a multi-armed bandit [4].

Scale: Our simulation model, in Section 4, was that of an unusually small call center (6 agents). It is left to understand the operational effects of heterogeneity and learning in realistic-size call centers (100's of agents): in a large agent population, the scope of heterogeneity is wider but laws of large numbers have their opposite effects - it is plausible that the "winner" will have to be determined from actual call center data. Moreover, a large number of agents renders asymptotic theory natural, for example within the QED and other many-server operational regimes.

Workload instead of Arrivals and Service-Times: Forecasting in call centers, and other service systems, has been traditionally confined to forecasting the *number* of arrivals. As WFM systems presume Poisson arrivals, they actually re-

quire merely arrival rates which, furthermore, are assumed to be piecewise constant over daily time-intervals (e.g. half-hours). But staffing is determined not only by arrival counts but also by the amount of work that each arrival imposes on agents, summarized here in terms of the arrival's service-time. This gives rise to the notion of *offered load*, both stationary and transient, which we now introduce.

For a service system of interest, let S denote its (generic random) service-time, and suppose that the arrival-rate to this system is a constant λ. Then $R = \lambda \times \mathsf{E}[S]$ is called the (stationary) *offered load*, representing the amount of work, measured in units of service-time, that arrives to the system per time-unit (e.g. minutes of work per minute). The driver of staffing is then the offered load R, as opposed to only λ, or λ and μ separately. In particular, WFMs can be fed with R only (though this is not widely acknowledged), and staffing recipes can be formulated in terms of R.

For example, in the context of the Erlang-C [25] and Erlang-A [19] models, but applicable far beyond, a fundamental staffing-recipe is the square-root staffing rule

$$n \approx R + \beta\sqrt{R},$$

where β is a constant that corresponds to grade-of-service; its specific value (which is practically small - less than 1.5 in absolute value) can be determined by economic considerations (for example, the ratio between delay costs and staffing costs, as in [13]). The importance of square-root staffing stems from the fact that it guarantees, uniquely, QED (Quality- and Efficiency-Driven) performance in many-server environments: that is, a natural, delicate balance between high levels of Quality-of-service and Efficiency-of-servers.

Now the call-center environment is time-varying, and there are periods of a day during which arrival-rates change quite abruptly. A time-varying analogue of the offered load is then required, one which will serve as the driver for time-varying staffing. Letting $\lambda(t)$ denote the time-varying arrival rate, the *right* answer is *not* $\lambda(t)/\mu$ (though it is sometimes practically acceptable). Indeed, the answer must be a quantity that, at time t, takes into account work that arrived prior to time t. Formally (assuming that time $t \in (-\infty, \infty)$, for simplicity of notation), the (transient) offered load $R(t)$ turns out to be given by either of the following two representations:

$$R(t) = \mathsf{E}\left[\int_{t-S}^{t} \lambda(u)\, du\right] = \mathsf{E}\int_{-\infty}^{t} \lambda(u)\ P_r\{S > t-u\}\, du\ , \quad -\infty < t < \infty,$$

where, as above, S denotes a generic (random) service time (of *all* customers). (Note that when the arrival rate does not vary with time, $\lambda(t) \equiv \lambda$, then $R(t) = \lambda \times \mathsf{E}[S]$, which is the stationary offered load.) The first representation above tells us that the offered load at time t is determined by the mean number of arrivals during the random time-interval $(t-S, t]$. The second representation is amenable to numerical calculations.

Readers are referred to Section 4.2.6 of [31] (downloadable), for a class-level discussion of the offered load $R(\cdot)$, in particular some practical approximations. Section 6 in [3] discusses the offered load in the context of forecasting towards QED performance, but using its naive representation $\lambda(t)/\mu$.

The time-varying offered load $\{R(t), t \geq 0\}$ is the backbone for time-varying staffing. Indeed, in the context of Erlang-A (and far beyond), setting time-varying staffing $\{n(t), t \geq 0\}$ by

$$n(t) \approx R(t) + \beta \times \sqrt{R(t)},$$

yields a remarkably *time-stable* performance, as discovered in [17]; furthermore, at any time, this stable performance is that of the *stationary* Erlang-A model (M/M/N + M), with the number of servers $n \approx R + \beta\sqrt{R}$ (in which R is a natural time-average of $R(\cdot)$).

The reason for our lengthy discussion of the offered load is that its estimation and prediction requires inference of both arrivals and *services*: the challenge is to model and infer service-times, towards calculation of the time-varying offered load, in call centers where multi-class customers are served by multi-skilled (heterogeneous) agents according to state-dependent protocols. (Interestingly, the offered load must take into consideration also the service-times of the un-served, namely those who abandoned, which [35] are doing for homogeneous customers and servers.) In view of the fact that the offered load is the fundamental primitive for staffing procedures, its modeling and inference poses first-order research challenges.

Tracking and Managing Individual Agents, in Call Centers and Elsewhere: Present technology enables the tracking of call center activities, at any levels of granularity. This has given rise, among other things, to CRM (Customer Relationship Management, or Customer Revenue Management), which we interpret here as managing the relationships between a company and its *individual* customers, taking into account individual event histories and future prospects. (CRM requires data of customer identification, which is lacking from the U.S. Bank data that is used here; but data-bases of other call centers, available from the Technion's SEE-Lab, does have customer IDs.)

Our present work, in analogy to CRM, raises the potential benefits of managing individual agents, again based on their individual event histories and future prospects. The above discussion of learning with random effects is a relevant example. Another example is [46], which requires continuous tracking of individual agents, in order to continuously update staffing needs so that customers eventually hardly wait. Similar ideas were found useful also in [49], who analyze staffing of Emergency Departments (EDs). In EDs, heterogeneity of resources is yet to be understood, and the time-varying view of the offered load is indispensable: this is due to the fact that ED length-of-stay is naturally measured in several hours, hence the offered load at a given time t is affected by events that took place several hours prior to t.

Acknowledgments

The Technion's Laboratory for Service Enterprise Engineering (SEE[1]) has provided access to our U.S. Bank data, and helped at various stages of the research.[2] We are thus grateful to the SEELab staff: Dr. Valery Trofimov, Igor Gavako, Polina Khudyakov, Katya Kutsy, Ella Nadjharov and Pablo Lieberman.

References

[1] Aalen, O. O., Borgan, O. and Gjessing, H. K. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer.

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

[1] Technion SEE Center - http://ie.technion.ac.il/Labs/Serveng.

[2] Researchers interested in getting access to our data, as well as to other SEE databases, should contact A.M. at avim@tx.technion.ac.il.

[3] Aldor-Noiman, S., Feigin, P. D. and Mandelbaum, A. (2010). Workload forcasting for a call center: Methodology and a case study. *Ann. Appl. Statist.* To appear.
[4] Arlotto, A., Chick, S. and Gans, N. (2010). Hiring and retention of heterogeneous workers. In preparation.
[5] Armony, M. (2005). Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* **51** 287–329.
[6] Armony, M. and Mandelbaum, A. (2008). Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Operations Research*. To appear.
[7] Armony, M. and Ward, A. R. (2010). Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems. *Operations Research*. To appear.
[8] Asmussen, S. (2000). *Ruin Probablities*. World Scientific.
[9] Atar, R. (2008). Central limit theorem for a many-server queue with random service rates. *Ann. Appl. Probab.* **18** 1548–1568.
[10] Atar, R., Shaki, Y. Y. and Shwartz, A. (2010). A blind policy for equalizing cumulative idleness. Preprint.
[11] Atar, R. and Shwartz, A. (2008). Efficient routing in heavy traffic under partial sampling of service times. *Mathematics of Operations Research* **33** 899-909.
[12] Bailey, C. (1989). Forgetting and the learning curve: A laboratory study. *Management Science* **35** 340–352.
[13] Borst, S., Mandelbaum, A. and Reiman, M. (2004). Dimensioning large call centers. *Operations Research* **52** 17–34.
[14] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005). Statistical analysis of a telephone call center. *J. Amer. Statist. Assoc.* **100** 36–50.
[15] Buffa, E., Cosgrove, M. and Luce, B. (1976). An integrated work shift scheduling system. *Decision Sciences* **7** 620–630.
[16] Donin, O., Trofimov, V., Zeltyn, S. and Mandelbaum, A. (2006). *The Call Center of US Bank*. DataMOCCA (Data MOdels for Call Centers Analysis), The Technion SEE Lab. Available at http://ie.technion.ac.il/Labs/Serveng/files/The_Call_Center_of_US_Bank.pdf.
[17] Feldman, Z., Mandelbaum, A., Massey, W. A. and Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54** 324–338.
[18] Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5** 79–141.
[19] Garnett, O., Mandelbaum, A. and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4** 208–227.
[20] Goldberg, Y., Ritov, Y. and Mandelbaum, A. (2010). The best linear unbiased estimator for continuation of a function. In preparation.
[21] Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall/CRC.
[22] Gurvich, I. and Whitt, W. (2009). Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research* **34** 363–396.
[23] Gurvich, I. and Whitt, W. (2009). Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Operations Management* **11** 237–253.

[24] Gustafson, H. (1982). Force-loss cost analysis. *WH Mobley, Employee Turnover: Causes, Consequences, and Control* **7**. Addison-Wesley, Reading, MA.
[25] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29** 567–588.
[26] Khudiakov, P. and Gorfine, M. (2010). Frailty models for patience of customrs who redial for a telephone service. In prepration.
[27] Khudyakov, P., Gorfine, M. and Mandelbaum, A. (2010). Phase-type models of service times. In preparation.
[28] Lieberman, P., Trofimov, V. and Mandelbaum, A. (2008). *Empirical Analysis of SBR in US Bank*. DataMOCCA (Data MOdels for Call Centers Analysis), The Technion SEE Lab. Available at http://ie.technion.ac.il/Labs/Serveng/files/Skills-Based-Routing_USBank.pdf.
[29] Mandelbaum, A. and Momcilovic, P. (2010). Queues with many servers and impatient customers. *Mathematics of Operations Research*. To appear.
[30] Mandelbaum, A. and Reiman, M. (1998). On pooling in queueing networks. *Management Science* **44** 971–981.
[31] Mandelbaum, A. and Zeltyn, S. (2009). Service engineering: Data-based course development and teaching; full version. To be published in *"IFORMS Transaction on Education", special issue on "Teaching Services and Retail Operations Management"*. Available at http://ie.technion.ac.il/serveng/References/teaching_paper.pdf.
[32] Momcilovic, P. and Mandelbaum, A. (2010). List-based routing. In preparation.
[33] Nembhard, D. and Osothsilp, N. (2002). Task complexity effects on between-individual learning/forgetting variability. *International Journal of Industrial Ergonomics* **29** 297–306.
[34] Nembhard, D. and Uzumeri, M. (2000). Experiential learning and forgetting for manual and cognitive tasks. *International Journal of Industrial Ergonomics* **25** 315–326.
[35] Reich, M., Mandelbaum, A. and Ritov, Y. (2010). The workload process: Modelling, inference and applications. In preparation.
[36] Shafer, S., Nembhard, D. and Uzumeri, M. (2001). The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Management Science* **47** 1639–1653.
[37] Shen, H., Brown, L. and Zhi, H. (2006). Efficient estimation of log-normal means with application to pharmacokinetic data. *Stat. Med.* **25** 3023–3038.
[38] Shen, H. and Huang, J. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management* **10** 391–410.
[39] Sisselman, M. E. and Whitt, W. (2007). Value-based routing and preference-based routing in customer contact centers. *Production and Operations Management* **16** 277-291.
[40] Sze, D. (1984). A queueing model for telephone operator staffing. *Operations Research* **32** 229–249.
[41] Talluri, K. T. and Ryzin, G. V. (2005). *The Theory and Practice of Revenue Management*. Springer.
[42] Tezcan, T. (2008). Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Mathematics of Operations Research* **33** 51–90.

[43] Trofimov, V., Feigin, P., Mandelbaum, A., Ishay, E. and Nadjharov, E. (2006). DATA-MOCCA: Data model for call center analysis. Technical Report, Technion – Israel Institute of Technology.

[44] Tseytlin, Y., Mandelbaum, A. and Momcilovic, P. (2010). Queueing systems with heterogeneous servers: On fair routing of patients in emergency departments. In preparation.

[45] Weinberg, J., Brown, L. and Stroud, J. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *J. Amer. Statist. Assoc.* **102** 1185–1198.

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

[47] Winer, R. S. (2001). A framework for customer relationship management. *California Management Review.*

[48] Yelle, L. (1979). The learning curve: Historical review and comprehensive survey. *Decision Sciences* **10** 302–328.

[49] Zeltyn, S., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Marmor, Y. N., Mandelbaum, A., Shtub, A., Lauterman, T., Schwartz, D., Moskovitch, K., Tzafrir, S. and Basis, F. (2009). Simulation-Based Models of Emergency Departments: Operational, Tactical and Strategic Staffing. Submitted to *ACM Transactions on Modeling and Computer Simulation (TOMACS).*