STATISTICAL ANALYSIS OF CALL-CENTER OPERATIONAL DATA: FORECASTING CALL ARRIVALS, AND ANALYZING CUSTOMER PATIENCE AND AGENT SERVICE

HAIPENG SHEN
Department of Statistics and Operations Research,
University of North Carolina at Chapel Hill, Chapel Hill,
North Carolina

Call centers are modern service networks in which agents provide services to customers via telephones. They have become a primary contact point between service providers and their customers. *Inbound* call centers take calls that are initiated by customers. Examples include phone operations that support reservation and sales for hotels, airlines, and car rental companies, as well as service centers of retail banks and brokerage firms. *Outbound* call centers initiate calls to outside parties, for example, those in organizations such as telemarketers, bill collection agencies, and marketing research and polling companies.

The increasing importance and complexity of call-center operations have rendered call centers into a fertile ground for academic research, both empirical and theoretical. Two survey papers [1,2] and an on-line bibliography [3] have well documented the current state of relevant research on call-center operations and directions for future research. This article focuses on statistical research about inbound call centers, in the areas of call arrival forecasting and analysis of customer patience and agent service times. There has been empirical research on call centers that is done outside the operations community and covers behavioral and psychological aspects [2] and human resource issues [4].

Brown *et al*. [5] described a simplified path that each call follows through a typical inbound call center. A customer calls the telephone number associated with the center. Once connected, the customer usually first enters an interactive voice response unit (IVR) and identifies himself or herself. In the IVR, the customer can perform some self-service transactions, and sometimes conclude the service there. In case the customer still wants to speak with an agent, if an agent is available and is capable of performing the desired service, then the customer is connected to the agent to receive service immediately; otherwise, the customer joins an invisible queue of waiting callers and starts to wait. It is often the case that informational or noninformational announcements such as music, commercials, or updates about any progress in the queue are provide to the customer while waiting. Eventually, the customer is either served by an agent or becomes impatient and hangs up (i.e., abandons) before an agent is available to serve him or her. Sometimes, the customer may call again after getting a busy signal, or abandoning the system, or even finishing a service.

The above simple description suggests that call centers can be modeled as queueing systems. Call-center managers then use queueing models to manage their center operations, seeking to appropriately balance agent utilization and service quality according to quality of service measures such as *average speed of answer* (ASA) or *fraction of abandoning customers*.

Existing queueing models usually make specific theoretical assumptions about various system primitives, such as the call arrival process, service durations, or customer abandonment behavior. The classical Erlang-C queueing model, for example, makes three assumptions: (i) the call arrival process is Poisson; (ii) service durations are exponentially distributed; and (iii) customers never abandon the queue before getting served. The third assumption turns out to be a serious shortcoming for modeling call centers, because of the prevalence of customer abandonment in call centers. The more recent so-called Erlang-A model is a more realistic

Wiley Encyclopedia of Operations Research and Management Science, edited by James J. Cochran

model that explicitly takes into account abandonments but assumes the *time to abandonment* distribution is exponential [6].

Gans *et al.* [1] assert that "the modeling and control of call centers must necessarily start with careful data analysis." The queueing models all involve certain parameters for the system primitives, for example, the Erlang-A model needs one to specify call arrival rates, service rates, and customer abandonment rates. To apply these models in practice, the associated parameters need to be estimated or forecasted through statistical analysis of historical call-center operational data. Furthermore, Brown *et al.* [5] showed that statistical analysis could be used to validate and calibrate queueing theoretical models.

Despite the importance of statistics, however, the empirical research on call centers remained very scarce until recently. The recent emerging growth, especially in the area of call forecasting, is primarily driven by increased access to call-center data and by widened recognition of the importance of empirical research.

This article aims to provide an introductory review of some empirical research about inbound call centers. It starts with a brief description of call-center operational data and an analysis tool, continues in the sections titled "Call Forecasting" and "Customer Patience and Agent Service Times" with general overviews of statistical research in call forecasting, and analysis of customer patience and agent service times respectively, and ends with a discussion of several directions for future research in the section titled "Future Challenges."

CALL-CENTER OPERATIONAL DATA

Call-center workforce management systems (WFM) routinely collect call-center operational data. The data record the physical process by which calls are handled. For example, among many other things, they consist of the time stamps of the various stages that a call goes through and the associated operational variables for each stage, which provide necessary starting points for most of the empirical research reported below.

A collaborative team of operations management researchers and statisticians performed the first thorough statistical analysis of a (then) unique call-by-call operational database collected at a call center of a small Israeli bank [5]. The analysis finds that several common queueing model assumptions are rejected empirically; in addition, some queueing-theoretic results are shown to be robust against such violations while a few others are not. The data and an earlier empirical analysis are publicly accessible from the Service Enterprise Engineering (SEE) Center at the Technion.

One key lesson the team learned is that, for successful and comprehensive statistical analysis, one needs call-by-call (or transaction-by-transaction) data that record the detailed event-history of the individual calls. For example, the team used the call-by-call data to develop statistical models for the distributions of call arrivals, agent service times and customer patience, which, in turn, allows one to make rigorous and valid statistical inference, or identify interesting empirical observations. Unfortunately, in practice call centers typically only use the call-by-call data to calculate simple summary statistics for the calls that arrive within fixed intervals of time, often 15 or 30 min, and then discard the transaction-level data. As demonstrated in Ref. 5, the aggregated summaries are too simple to provide enough details for meaningful statistical modeling, several examples of which are discussed later.

In addition, the WFMs are designed to collect the operational data in an "engineering-efficient" way; hence the collected data in most cases are not directly amenable for statistical analysis. The experience of analyzing the pilot study helped the team to define and develop a *DATA MOdel for Call-Center Analysis* (DATA-MOCCA) [7]. The mission of the DATA-MOCCA project is to collect, preprocess, organize, analyze, and distribute transactional data from call centers. The core of DATA-MOCCA is a database system that is useful for the analysis of the call-by-call data that are gleaned from call centers' telephone switches and information systems. The latest distributable implementation includes two large databases, one of a medium sized US bank and the other of an Israeli telecom company, along with a user-friendly query engine, SEESTAT, that facilitates empirical exploration and statistical analysis of the data. A public version of SEESTAT can be downloaded from the Technion SEE Center.

CALL FORECASTING

Seventy percent of call-center operating expenses are salary costs for human resource [1]. Effective management of a call center requires its manager to match up the center resource with future workload during any planning period. The workload or *offered load* is defined to be the product of the call arrival rate and the mean service time of the arrivals. One can interpret workload as the expected time units of work that arrives to the system per unit of time. For example, consider a stationary 30-min interval during which the arrival rate is 4 and the mean service time is 15 min, then the workload to the system would be $4 \times 15/30 = 2$. The workload serves as a lower bound for the staffing level for an Erlang-C system, which does not allow blocking or abandonment. Ideally, for staffing purposes, a manager wants to forecast workload accurately. However, most of the existing research has focused on forecasting future call arrival rates, with the exception of Refs 5 and 8.

In this section, we first describe several common characteristics of call arrivals in the section titled “Common Characteristics of Call Arrivals”: time inhomogeneity, arrival rate uncertainty, and interday and intraday dependence. These features are observed empirically and are useful for developing effective forecasting methods. We then review existing call-forecasting methods in the section titled “Call-Forecasting Methods.”

Common Characteristics of Call Arrivals

In most inbound call centers, both the arrival rate of calls and the mix of types of calls entering the system vary over time. Call arrivals have both *predictable* and *unpredictable* components: over very short periods of time, seconds or minutes, the numbers of arriving calls are viewed as stochastic and unpredictable, typically as some form of a Poisson process; however, over longer periods of time–hours, days, weeks, or months–the numbers of arrivals are traditionally viewed as more predictable (see Fig. 5 of Ref. 1 for a hierarchical view of arrival rates).

Call forecasting starts with historical call arrival data, which record the time stamps at which calls arrive to the center and their corresponding service types, in cases where different types of calls are handled by different groups of agents. The common practice in the literature is as follows: first, the center operating period is divided into short time intervals, usually 15 or 30 min, during which the arrival rates are assumed to remain constant; then, the raw data of arrival times are aggregated into numbers of arrivals within the short intervals. The collection of such aggregated call volumes for a given day is referred as the *intraday call volume profile* of that day [9].

Common call-center models and practice then assume that the intraday call volume profile is a realization from a time-inhomogeneous Poisson process, where the underlying arrival rate function varies across different intervals. Brown *et al*. [5] constructed a statistical test, and empirically validated the time-inhomogeneous Poisson assumption using the Israeli call-center data. In addition, they performed a statistical test and claimed that the arrival rate function remains random (or uncertain) given the available covariates of time-of-day, day-of-week, and type of service, and hence is hard to predict. A Poisson process with uncertain arrival rate is referred as a *doubly stochastic Poisson process*, or a *Cox process*.

The phenomenon of arrival rate uncertainty can be manifested empirically through the *overdispersion* of the call volumes during the same time interval across multiple days. Here overdispersion refers to the fact that the variance of the volumes is much larger than the mean. Note that, if the volumes were independent and identically distributed according to a Poisson random variable, then the variance would be equal to the mean. This phenomenon has been recognized before by various authors. Maman [10] recently proposed a theoretical model for overdispersion in call centers and hospital emergency departments, and suggested procedures to empirically estimate the model parameters.

After some initial processing, historical call arrival data, or historical intraday call volume profiles, can be stored as a two-way data matrix, where each row corresponds to the intraday call volume profile for a particular day, and each column records the call volumes during a particular time interval across the days. Alternatively, one could concatenate all the intraday profiles one after another to form a single long profile.

Viewing the historical arrival data as a matrix, there exists a two-way dependence structure that is useful for forecasting future arrival rates. First, there is *interday dependence* among the call volumes across days. For example, Brown *et al*. [5], Shen and Huang [9], and Weinberg *et al*. [11] all reported empirical evidence of dependence within the time series of daily total call volumes. Such dependence can depend on seasonal factors such as day-of-week, month-of-year, and is helpful for forecasting arrival rates several days, weeks, or months ahead.

In addition, there is *intraday dependence* in that call volumes during different time intervals within a given day are positively correlated. For example, Avramides *et al*. [12] developed doubly stochastic Poisson models that reproduced three empirically observed features of call arrival data, *overdispersion*, *time-varying rate*, and *intraday dependence*. Intraday dependence makes it possible to perform within-day dynamic updating, for example, to adjust the forecasts for the afternoon arrival rates on the basis of the call volumes in the morning of the same day. This updating is operationally both feasible and beneficial. As shown in Refs 9, 11, and 13, within-day updating can substantially reduce the forecast error for the rest of the day, and it can be done as early as a couple of hours into the working day to still have a significant effect. In turn, operational benefits would follow from the ability to change agent schedules according to the updated forecast. For example, to adjust for a revised forecast, call-center managers may send people home early or have them take training sessions or work on alternative tasks; or they may arrange agents to work overtime or call in part-time or work-from-home agents. There is some ongoing work that formulates stochastic programming models with recourse actions to actually quantify the practical benefits of within-day updating for scheduling agents, using real call-center data.

Call-Forecasting Methods

Early work on forecasting of call arrivals usually ignores intraday dependence, and focuses on interday dependence among daily total call volumes. Part of the reason is due to lack of relevant data. Forecasting is usually performed through time series forecasting methods such as using autoregressive integrated moving average (ARIMA) models. Transfer functions and exogenous variables are included to incorporate special effects such as advertisement, holidays, and sales promotion. In addition, only point forecasts for future arrival rates are produced. Workforce management in call centers traditionally assumes that these point forecasts are certain. However, the arrival rate uncertainty reported earlier in the section titled “Common Characteristics of Call Arrivals” suggests that the forecasted rates are not realized with probability one.

Recently, with the availability of call-by-call data, several important developments in the call-forecasting literature have been made [5,8,9,11,13–15] that take into account both interday and intraday dependence. The forecasting approaches can be roughly grouped into two categories: model-driven and data-driven. Most of the recent developments consider both point forecasts and distributional forecasts, which capture forecasting uncertainty.

Model-Driven Approaches. The model-driven approaches include Refs 5, 8, 11, 13, and 14. They start with various stochastic models of call arrivals, often as some variations of Poisson processes.

Both Brown *et al*. [5] and Weinberg *et al*. [11] model the call arrivals using a time-inhomogeneous Poisson process with a random rate function. In both papers, a square-root transformation is first applied to the original arrival counts. The motivation is twofold: first, the square-rooted counts roughly have a constant variance; second, they are approximately normally distributed. Brown *et al*. [5] then build a multiplicative two-way mixed-effects model on the transformed counts. The square-root of the arrival rate for each interval is modeled as the product of the daily total of the square-rooted rates and the proportion of arrivals during that time interval. They then assume that the arrival proportion for each interval remains the same for different days of the week, and the daily total square-rooted rate can be forecasted using the total of the square-rooted counts of the previous day.

Weinberg *et al*. [11] extended the work of Brown *et al*. [5] to develop forecasting models for both day-to-day forecasting and within-day updating. As an example of the various forecasting methods, the statistical model proposed by Weinberg *et al*. [11] is presented below. Suppose N_{ij} is the number of arrivals during the jth time interval of the ith day, which is modeled as a Poisson random variable with a random rate λ_{ij} that depends on both day-of-week and time-of-day. A two-way multiplicative Bayesian model is then proposed as follows:

$$\begin{cases} x_{ij} = \sqrt{N_{ij} + 1/4} = \sqrt{\lambda_{ij}} + \epsilon_{ij},\, \epsilon_{ij} \sim N(0, \sigma^2), \\ \sqrt{\lambda_{ij}} = y_i g_{d_i}(t_j), \\ y_i - \alpha_{d_i} = \beta(y_{i-1} - \alpha_{d_{i-1}}) + \eta_i,\, \eta_i \sim N(0, \phi^2), \\ \dfrac{\mathrm{d}^2 g_{d_i}(t_j)}{\mathrm{d}t_j^2} = \tau_{d_i} \dfrac{\mathrm{d}W_{d_i}(t_j)}{\mathrm{d}t_j}, \end{cases}$$

where t_j is the center of the jth time period, $g_{d_i}(\cdot)$ models the *smooth* intraday call arrival pattern that depends on day-of-week $d_i (= 1, \ldots, 5)$, y_i is a random effect with an autoregressive structure adjusting for d_i, and $W_{d_i}(t)$ are independent Wiener processes with $W_{d_i}(0) = 0$, and $\mathrm{var}\{W_{d_i}(t)\} = t$. The model essentially assumes that the time series of daily total square-rooted rates follows a first-order autoregressive model with day-of-week effect, and the arrival proportion changes smoothly across the time intervals and can depend on day-of-week. The authors then choose suitable prior distributions for the model parameters, and develop a Markov chain Monte Carlo (MCMC) algorithm for parameter estimation and forecasting. They also propose a within-day learning algorithm to efficiently perform within-day updating of arrival rates.

Without using the square-root transformation, Shen and Huang [13] directly modeled the historical arrival count matrix as observations from a time series of inhomogeneous Poisson processes, where the time series dependence is across the sequence of days (or rows of the data matrix). The count matrix is then viewed as a multivariate Poisson time series with the dimension being the number of columns in the matrix, or the number of time periods within a day. The dimensionality of the time series is usually high, for example, a 17-h working day consists of 68 time intervals of 15 min each. Hence, the authors start with a Poisson factor analysis on the count matrix to achieve dimension reduction. The estimated factor score series are then modeled using univariate time series models. The forecasts of future factor scores are combined with the extracted factor loadings to yield forecasts of future arrival rate functions. Penalized Poisson regressions on the factor loadings are used to generate dynamic within-day updating of arrival rates.

Unlike in the previous papers, Aldor-Noiman *et al*. [8] considered a univariate response vector by concatenating the square-root-transformed daily volume profiles together. This “long” volume profile (instead of the “fat” count matrix analyzed in the earlier studies) is then modeled using a normal linear mixed-effects model that treats both day and period effects as random effects, both of which are assumed to have a first-order autoregressive correlation structure, and includes exogenous variables such as day-of-week and billing cycle information as fixed effects. A two-stage model estimation algorithm is proposed in order to avoid convergence issues. The authors consider the problem of forecasting both arrival rate as well as workload, under various forecast lead times. In addition, they study the effect of interval length on forecasting accuracy and the effect of the model results on call-center operational performance.

Similar to Weinberg *et al*. [11], Soyer and Tarimcilar [14] also used a Bayesian approach to model call arrivals at an inbound sales call center. For the purpose of developing good marketing strategies, the authors are interested in advertisement-specific analysis of call arrivals to evaluate the efficiency of various advertisements and promotions. Different from the previous studies, the call arrival data are counts of incoming calls during periods of days over the life cycle of an advertisement. The counts are modeled using a modulated time-inhomogeneous Poisson process. A proportional rate model is used to model the underlying arrival rate function as the product of a baseline rate function and an advertisement effect consisting of exogenous variables such as the advertisement characteristics. Similar models have been considered before in the survival analysis literature. Random effects are used to incorporate potential heterogeneity among various advertisements.

Data-Driven Approaches. The data-driven forecasting approaches include Refs 9 and 15. The approaches have advantages in cases where no specific stochastic models for the call arrival processes are made. For example, there exists empirical evidence that calls entering the service queue after a congested IVR did not follow a time-inhomogeneous Poisson process.

Shen and Huang [9] combine the ideas of dimension reduction and time series forecasting; hence the method is similar to the one proposed by them in Ref. 13. Two differences between Refs 9 and 13 are (i) the former does not make any assumption on the arrival process, while the latter assumes that the call arrivals follow a time-inhomogeneous Poisson process; (ii) the former produces forecasts for future call volumes while the latter generates future rate forecasts. The authors [9] treat the square-rooted intraday call volume profiles as a high dimensional vector time series. They first reduce the dimensionality by applying singular value decomposition to the square-rooted count matrix. The dimension reduction allows them to obtain several pairs of "most important" interday and intraday features, which naturally decouple interday and intraday dependence. Once these features are obtained, in a manner similar to Ref. 13, time series models and penalized regression techniques are used to perform interday forecasting and intraday updating. Since no distributional assumptions are made on the arrival counts, nonparametric bootstrapping is proposed to derive distributional forecasts.

Similar to Aldor-Noiman *et al*. [8], Taylor [15] views the historical data as a "long" univariate time series, and compares the empirical performance of several univariate time series methods for forecasting intraday call arrivals. The methods studied include seasonal ARIMA time series models, periodic autoregressive models, exponential smoothing models with two seasonal cycles (for intraday and intraweek), robust exponential smoothing, and dynamic harmonic regression. The author focuses on point forecasting, and considers forecasting lead times ranging from one half-hour ahead to two weeks ahead. No clear winning method is identified, although seasonal ARIMA and exponential smoothing with two seasonal cycles appear to perform better for lead times up to two days ahead.

CUSTOMER PATIENCE AND AGENT SERVICE TIMES

Call-center operations are complicated partly due to the human factors involved in particular, customers and agents. In call centers, it is people, rather than "jobs," that require service and it is people, rather than equipments, that provide the service. Call-by-call operational data also provide opportunities for researchers to empirically study customer and agent behaviors in a call-center environment. Although research in this area is still lacking, interesting and important developments have been made recently. The review below focuses on statistical analysis of customer patience and agent service times.

Customer Abandonment and Patience

As discussed earlier, for customers who want to speak with agents, chances are that they need to wait in the service queue before getting served. Some of them will become impatient, and abandon the system before an agent becomes available. Hence, it is important to understand customer abandonment or patience behavior. In addition, a wrong model of customer patience can lead to wrong staffing [16].

Traditional queueing theory has been naive in its modeling of customer patience, which is often ignored. If at all acknowledged, customers are usually assumed to arrive to the system with independent and identically distributed patience. Furthermore, the patience distribution is most often assumed to be exponential as in the Erlang-A model [6].

Brown *et al*. [5] analyze customer patience empirically at the Israeli call center. They propose to use *the time that a customer is willing to wait before hanging up* as an ancillary measure for the customer's patience. Note that this measure is observed only for a customer who does hang up. For a customer who gets served, we only know that he or she is willing to wait longer than his or her actual waiting time; however, we do not know exactly how much longer, in which case his or her patience is considered to be *right censored*. For statistical analysis of customer patience, the censoring has to be dealt with using survival analysis techniques.

Several interesting empirical observations are revealed by the analysis. First, a clear stochastic ordering emerges among customers seeking different services. For example, stock-trading customers are more patient than customers seeking regular services. Similarly, high priority customers appear to be more patient than regular customers [1]. Second, customers are more likely to abandon the queue after waiting for either only a few seconds or about 60 s. One plausible explanation is that the call-center system plays a "Please wait" message after a customer waits a few seconds and again at 1 min, which apparently prompts the customers to abandon the queue. This also suggests that the patience distribution is not exponential. If it were, the probability for a customer to abandon would remain the same, no matter how long she had waited. Furthermore, the authors demonstrate empirically that the Erlang-A model is very robust against the violation of the exponential assumption of customer patience. The observation of the robustness motivates further research to explore for theoretical justification, which is later provided by Zeltyn and Mandelbaum [16].

Adaptivity of Customer Patience

Traditional queueing theory also assumes that customer patience is independent of system performance, unrelated to past experience or future anticipation of the system. However, Zohar *et al*. [17] provide some empirical evidence to support an inconsistent hypothesis: experienced customers "adapt" their patience on the basis of their expectations regarding system waiting time.

Using the Israeli call-center data, for every 15-min interval during weekdays between 7.00 a.m. and midnight, the *percentage of abandonment* of those customers who wait as well as the *average waiting time* for the same customers are calculated. Surprisingly, for experienced customers, the percentage of abandonment remains at about 38%, regardless of what the average waiting time is, which ranges between 90 and 250 s; on the other hand, for novice customers, the two measures are strongly positive linear dependent. In addition, for experienced customers, the expected patience increases linearly when the system is more congested (and hence needs the customers to wait longer). This also suggests that the experienced customers know what to expect on the basis of their past experience, and when the system needs them to wait longer, they are willing to do so.

Agent Service Times

As providers of service to customers via telephones, call-center agents are another important human component of call-center operations. However, even less empirical research has been devoted to study them.

Traditionally, queueing models assume that service times within a given interval are independent and identically distributed according to an exponential distribution. This assumption is imposed mainly for theoretical simplicity due to the lack-of-memory property of exponential distributions. Although some empirical evidence exists for the exponentiality, service times are not always exponentially distributed. For example, Brown *et al*. [5] analyze the individual call service times in the Israeli call center, and show that they are lognormally distributed, that is, their natural logs are normally distributed. This fact makes the natural logs appealing for statistical analysis. Furthermore, the lognormal distribution provides an excellent fit not only for the overall service time but also when restricted to service types, individual agents, time-of-day, and so on.

Using the call-by-call data, the authors also find out that, in the early portion of the data, a significant amount of calls (7%) are handled in less than 10 s, comparing to 2% of such calls during the latter portion. Service times shorter than 10 s are questionable. As a matter of fact, those short service times are primarily caused by agents who simply hang up on customers to obtain extra rest time. This phenomenon of agents “abandoning” customers is often due to distorted incentive schemes, especially those that overemphasize short average service time or the total number of calls handled by an agent.

FUTURE CHALLENGES

We conclude this article with a brief discussion of several challenging research problems. Empirical analysis of call-by-call operational data should shed light on all of them, influencing how they are modeled by researchers, as well as how they are managed in practice.

The first challenge is about workload forecasting. Most research reviewed in the section titled “Call Forecasting” focuses on forecasting future arrival rates. However, to support agent staffing, a call-center manager requires forecasts of future workload, which combines arrivals with the service times of these arrivals. This is an important but difficult problem for time-varying arrivals, because there exists a time lag between peak arrivals and peak congestion [18]. More work is needed to develop methods that forecast arrival rates and service rates simultaneously. Some interesting contribution has been made recently at the Technion, where a procedure is proposed to estimate the service-time distribution of those abandoning customers.

The second challenge concerns the use of IVRs, which enable customers to use self-service without speaking to an agent. Little is formally known about the statistical or operational properties of IVRs and how they should be managed. At a high level, customer interaction with an IVR can be thought of as a service time, to which queueing-theoretic concepts can be applied. However there has been very limited theoretical work on the use of IVRs. To summarize, the time a customer spends in the IVR, as well as how that time affects subsequent agent service time, are important elements of call-center operations about which little is known.

The third challenge has to do with customer retrial behavior. More specifically, many of the impatient abandoning customers then become so-called retrials, calling again at a later time. Similarly, customers that receive busy signals, as well as those that are served, may become retrials. While there are classical theoretical models of retrial queues, again little is known about the actual retrial behavior of customers. Careful analysis of retrial data should help to shape how retrials are modeled and managed in the future. Some ongoing research at the Technion is analyzing customer retrials using frailty models.

The analyses of IVRs and customer retrials are interesting in and of themselves. They are also important because of their potential impact on the statistical properties of the stream of calls that arrives to agents. Both IVR and retrial behavior have the potential to change important statistical properties of the pattern of arrivals to agents. For example, the calls exiting a congested IVR may not be Poisson.

While waiting, customers usually listen to music, commercials, or updates of their positions in the queue. These messages are referred to as *waiting-time fillers* in consumer psychology. How customers react to such time fillers is another important but underexplored research problem. Brown *et al*. [5] observe empirically that a "Please wait" message actually prompts customers to hang up. Recently, Munichor and Rafaeli [19] used lab experiments to study caller reactions to three types of telephone waiting-time fillers: music, apologies, and information about location in the queue. On the basis of the callers' evaluations after getting their services, the authors conclude that callers are happier when they are given a sense of progress through periodic updates of their position in the queue.

Furthermore, a better empirical understanding of service-time heterogeneity is needed. Empirical research has shown that differences among service times are associated with service types and time-of-day. There is much more work that could and should be done to better understand the drivers of service times. For example, it is interesting to understand how the heterogeneity is caused by agent-specific factors such as learning, forgetting, shift fatigue, and cross-training. Such research findings can complement and benefit recent theoretical work addressing agent heterogeneity in call centers.

Finally, the existence of heterogeneous customer classes and agent types and the employment of skills-based routing introduce additional complications and present many interesting questions. For example, in the context of workload prediction, the prediction depends in some sense on how customers are routed to agents; if one routes more customers to the slow agents, then the workload will be higher. Skills-based routing decisions also affect customers' waiting experience and their patience.

REFERENCES

1. Gans N, Koole G, Mandelbaum A. Telephone call centers: tutorial, review, and research prospects. Manuf Serv Oper Manage 2003;5:79–141.
2. Akşin Z, Armony M, Mahrotra V. The modern call-center: a multi-disciplinary perspective on operations management research. Prod Oper Manage 2007;16:665–688.
3. Mandelbaum A. Call centers: research bibliography with abstracts. Version 7, 2006 May. Technical Report, Technion–Israel Institute of Technology. Available at http://iew3.technion.ac.il/serveng/References/US7_CC_avi.pdf.
4. Batt R, Doellgast V, Kwon H. US call center industry report 2004: national benchmarking report strategy, HR practices & performance. CAHRS Working Paper No. 05–06. Ithaca (NY): Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies; 2005.
5. Brown LD, Gans N, Mandelbaum A, *et al*. Statistical analysis of a telephone call center: a queueing science perspective. J Am Stat Assoc 2005;100:36–50.
6. Garnett O, Mandelbaum A, Reiman M. Designing a call center with impatient customers. Manuf Serv Oper Manage 2002; 4: 208–227.
7. Trofimov V, Feigin P, Mandelbaum A, *et al*. DATA-MOCCA: Data Model for Call Center Analysis. Volume 1: model description and introduction to user interface. Technical Report, Technion–Israel Institute of Technology; 2006. Available at http://iew3.technion.ac.il/serveng/References/DataMOCCA.pdf.
8. Aldor-Noiman S, Feigin PD, Mandelbaum A. Workload forecasting for a call center: methodology and a case study. Ann Appl Stat 2009. In press.
9. Shen H, Huang JZ. Interday forecasting and intraday updating of call center arrivals. Manuf Serv Oper Manage 2008;10:391–410.
10. Maman S. Uncertainty in the demand for service: the case of call centers and emergency departments [Masters Thesis]: Technion–Israel Institute of Technology; 2009. Available at http://iew3.technion.ac.il/serveng/References/Thesis_Shimrit.pdf.
11. Weinberg J, Brown LD, Stroud JR. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. J Am Stat Assoc 2007;102:1185–1199.
12. Avramidis AN, Deslauriers A, L'Ecuyer P. Modeling daily arrivals to a telephone call center. Manage Sci 2004;50:896–908.
13. Shen H, Huang JZ. Forecasting time series of inhomogeneous Poisson processes with applications to call center workforce management. Ann Appl Stat 2008;2:601–623.
14. Soyer R, Tarimcilar MM. Modeling and analysis of call center arrival data: a Bayesian approach. Manage Sci 2008;54:266–278.
15. Taylor J. A comparison of univariate time series methods for forecasting intraday

arrivals at a call center. Manage Sci 2008; 54:253–265.

16. Zeltyn S, Mandelbaum A. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. Queueing Syst 2005;51:361–402.

17. Zohar E, Mandelbaum A, Shimkin N. Adaptive behavior of impatient customers in tele-queues: theory and empirical support. Manage Sci 2002;48:566–583.

18. Feldman Z, Mandelbaum A, Massey WA, *et al*. Staffing of time-varying queues to achieve time-stable performance. Manage Sci 2009;54:324–338.

19. Munichor N, Rafaeli A. Numbers or apologies? Customer reactions to tele-waiting time fillers. J Appl Psychol 2007;92:511–518.