Structural Estimation of Callers' Delay Sensitivity in Call Centers

Seyed Emadi
Northwestern University

Baris Ata
Northwestern University

Zeynep Aksini
Koc University

Che-Lin Su
University of Chicago

Abstract

We model callers' decision making process in call centers as an optimal stopping problem. After each period of waiting, a caller decides whether to abandon or to continue to wait. The utility of a caller is modeled as a function of her waiting cost and reward for service. We use a random-coefficients model to capture the heterogeneity of the callers and estimate the cost and reward parameters of the callers using the data of individual calls made to an Israeli call center. We also conduct a series of counterfactual analyses that explore the effects of changes in service discipline on resulting waiting times and abandonment rates. Our analysis reveals that modeling endogenous caller behavior can be important when major changes (such as a change in service discipline) are performed, and that using a model with an exogenously specified abandonment distribution may be misleading.

1 Introduction

Services cannot be stored and frequently cannot be produced without their customers. Thus, waiting is an inevitable part of most service encounters. A growing number of customer contacts take place in call centers, making them a dominant channel for encounters with waiting (Gans et al. (2003), Aksin et al. (2007)). Customers dislike waiting, especially when it is in invisible queues as in call centers. The dislike can be attributed to feelings such as anxiety, ambiguity and a sense of wasted time (Suck and Holling (1997), Leclerc et al. (1995)). A natural consequence of such feelings is that some customers lose their patience and abandon the queue before receiving service. Caller abandonment reflects dissatisfaction and may lead to profit loss of the service provider. It further affects performance metrics such as average waiting times. Understanding caller patience is an essential first step in designing superior service encounters, which motivates the research in this paper. We model wait or quit decisions by callers and estimate their patience by making use of call center data on waiting and abandonment times.

A traditional approach to modeling reneging or abandonment in queues is by considering an exogenous patience time distribution for customers. Thereby, customers abandon the queue when their perceived waiting time exceeds their patience (Gans et al. (2003) and references therein). Frequently, the choice of the patience time distribution is driven by tractability concerns. A distribution is chosen that makes subsequent analysis possible, and its parameters are estimated from historical data. A more recent stream of research focusing on call centers has emphasized the importance of the direct use of data to fit patience time distributions (Brown et al. (2005)).

While the traditional approach lends itself to tractable analysis in many cases, it does not enable explicit modeling of patience. An alternative modeling approach has been to consider wait or quit decisions by callers as the outcome of forward-looking behavior of utility maximizing rational agents. With a utility function that consists of a reward from service and a linear delay cost, forward looking customers either abandon upon arrival (i.e. balk) or not at all. In particular, no caller abandons while waiting. As reviewed in Hassin and Haviv (2003), different assumptions are required to induce rational abandonments while waiting. In Hassin and Haviv (1995), a reward from service that may drop to zero induces rational abandonments. Mandelbaum and Shimkin (2000) incorporate a fault state upon arrival, which means callers arriving in that state will never be served. In this extended model, callers may abandon while waiting because they are worried that they may be trapped in the fault state.

In a similar study, Shimkin and Mandelbaum (2004) consider nonlinear delay costs with no fault state. Under suitable conditions on waiting costs, the authors study the equilibrium in which callers decide upon arrival when to abandon. The abandonment times of the callers are optima of their utility functions. In both Mandelbaum and Shimkin (2000) and Shimkin and Mandelbaum (2004), the authors model the system as a Markovian queue with a general abandonment time distribution (an $M/M/m+G$ queue) and find the waiting time distribution of the callers resulting from the equilibrium between the offered waiting time distribution of the system and the patience time distribution of the callers.

The work of Mandelbaum and Shimkin is an important antecedent of this paper. Indeed, our formulation builds on the following observation made by Shimkin and Mandelbaum (2004): "It is plausible that abandonment decisions are taken online based on the customer's assessment of the current situation and the utility of further wait."

In our model, callers receive a reward from service and incur a delay cost, which is linear in their waiting time along the lines of Naor (1969). Moreover, callers are heterogeneous in their reward and cost parameters, which is captured via a random-coefficients model. No information about the duration of waiting is conveyed to the callers as they wait in the queue, i.e., there is no delay announcement. Callers are forward looking and make wait or abandon decisions dynamically as they wait. To be more specific, we assume that the waiting cost is sunk, i.e., waiting cost incurred in the past are irrelevant to future decisions. Hence, a caller only considers the expected future utility associated with her actions. A caller's utility depends on her reward, delay cost, and idiosyncratic random shocks, representing (external) events that affect the callers' utility. The idiosyncratic shocks correspond to the unobserved variables in the empirical industrial organization literature, see for example Rust (1987), which are observed by the caller but not recorded in the data. Using the terminology that is standard in the operations research literature, each caller solves an optimal stopping problem where stopping corresponds to abandoning. We estimate callers' cost and reward parameters using the maximum likelihood estimation (MLE) approach.

The main contribution of this paper is to develop a simple model which endogenizes and explains the callers' abandonment behavior. Using our model, we estimate the callers' cost and reward parameters and conduct a counterfactual analysis. More specifically, in a series of experiments we change the service discipline of the call center and compare our model (with endogenous abandonments due to forward looking callers) to a model where an exogenously specified abandonment distribution (obtained from the data) is used. For small changes, where we only change the parameters of the existing priority scheme for example, the exogenous modeling appears to be sufficient. However, the comparisons show the importance of endogenizing customer behavior in settings where major policy changes are made.

Our model also makes a methodological contribution to the analysis of queueing systems with abandonments. To the best of our knowledge, this is the first attempt of applying structural estimation approach in the call center operations context. Furthermore, it is the first empirical demonstration of the effect of modeling endogenous customer abandonment behavior in queues. Indeed, our framework can be modified suitably to study various other queueing systems (with abandonments), e.g., those arising in settings such as in the delivery of healthcare services, make-to-order manufacturing, etc.

The rest of the paper is structured as follows. Section 2 reviews the literature. Section 3 characterizes the model of callers' decision making process. Section 4 describes the data. Section 5 explains the estimation method, and provides the estimation results and their interpretation. Section 6 describes the counterfactual analysis. Concluding remarks are offered in Section 7.

2 Literature Review

The behavioral aspects of waiting have been studied extensively. Mostly, waiting is shown to have a negative effect on individuals. Leclerc et al. (1995) study whether people treat waiting as losing monetary utility. In an experiment, they show that individuals' marginal cost of waiting is a concave function when the waiting time is large, e.g. 20 minutes to 5 hours. Suck and Holling (1997) model the effect of waiting time duration and variability on stress caused by waiting. They show that increase in either the duration or variability of the waiting time results in more stressful conditions for the customers. Bitran et al. (2008) study the implications of the psychology of waiting for the design of queuing systems and provide a comprehensive review.

The same waiting experience can have different effects on different people, depending on how it is perceived. Indeed, a stream of research in the behavioral literature analyzes the effect of time perception on callers' behavior. Hornik (1984) studies the difference between the perceived waiting time and the actual waiting time of the customers. The author verifies the existence of this difference empirically. Chebat et al. (1993) state that musical and visual cues, eg. playing music, may decrease customers' perception of the time spent waiting and thus reduce customers' dissatisfaction from waiting. According to experiments in Munichor and Rafaeli (2007), a sense of progressing in the queue enhances customers' mood while waiting. In Zakay (1989), the author suggests that the perceived waiting time is longer when a customer is more conscious about the passage of time. The author also states that conveying the delay information may shorten the perceived waiting time because it decreases the customer's need to pay attention to the passage of time.

A growing literature studies the effect of delay information on callers' behavior and performance of the service center. We refer the reader to Whitt (1999), Guo and Zipkin (2007), Jouini et al. (2011), Armony et al. (2009) and references therein for a detailed account of that literature. In the data, no delay information is provided to callers, consistent with our model.

Apart from the delay information, instruments such as price can be used to control the customers' behavior and decision in waiting situations. Naor (1969) is one of the first papers in the queuing context to model customers as utility maximizing agents whose actions can be modulated via pricing. Naor models a system where imposing tolls affects customers' decision to join the queue or to balk. Mendelson (1985) studies how queueing delay and pricing change the behavior of customers and their arrival rate. The author shows that a manager can maximize the value of the services to the organization or minimize the costs by choosing the proper price and capacity.

A closely related area is the equilibrium analysis of abandonments by rational customers, who maximize their utilities in choosing between waiting and abandoning. Zohar et al. (2002) provide a model of rational abandonments suggesting that customers adapt their patience to their anticipated waiting time. The authors assume that customers' patience follows a parametric distribution, where its parameters are only affected by anticipated waiting time of the customers, and depends on neither customers' utility from receiving service nor their waiting cost. Hassin and Haviv (1995) study the abandonment profile of rational customers in the setting of a single-server Markovian queue with abandonments. The authors assume that customers' waiting cost is linear and customers' utility from service becomes zero if they do not receive service within a fixed time beyond arrival. The authors show that the optimal behavior of the customers is one of the two abandonment profiles: abandoning upon arrival or abandoning when the service utility drops to zero.

As reviewed in the Introduction, Mandelbaum and Shimkin (2000) and Shimkin and Mandelbaum (2004) analyze rational abandonment behavior of impatient customers in a Markovian queue with a general abandonment time distribution (an $M/M/m+G$ queue). In both papers, the authors assume that the waiting cost and the service utility of the callers are given, and customers depending on these parameters, act rationally and decide upon arrival when to abandon if they do not receive service. Our work differs from Mandelbaum and Shimkin (2000, 2004). An important difference is that in our model, callers make their decisions dynamically, not just upon arrival. In essence, each caller solves an optimal stopping problem where "stopping" means abandoning. Another important difference is that we do not undertake a queueing theoretical analysis to derive the equilibrium waiting time. Rather, we deduce the equilibrium distribution of the waiting time from the observed data, and assume that it is common knowledge among the callers and the call center provider;

callers acquire this knowledge through their past experiences of contacting the call center.

We assume callers' utility depends on waiting cost, reward and their idiosyncratic random shocks, which resemble the random utility models one sees in the structural estimation literature, cf. Berry et al. (1995). In that literature, two of the most relevant papers to our work are Rust (1987) and Nair (2007). Rust (1987) studies the estimation of structural parameters of a regenerative optimal stopping model where a maintenance manager in each period of time has to decide between two actions: replacing the engine of a bus and incurring the cost of overhaul, or not replacing the engine and incurring the cost of unexpected failure. In Nair (2007), the author examines the effect of consumers' forward looking behavior on profit of the firms selling video games. The author proposes a dynamic consumer choice model where consumers can buy the product and exit the market, or wait to buy the product at a lower price. The author also models the profit of the firm and suggests that firms may lose profit by not taking the forward looking behavior of the callers into account while setting the prices. Similarly, our counterfactual analysis illustrates, how disregarding endogenous abandonment behavior can lead to erroneous assessment of service levels, while making choices of service discipline in a call center.

3 The Model

In this section, we present a dynamic model of caller's decision process. In each period as callers wait in the queue, they face the decision to either abandon or continue to wait. If a caller chooses to abandon, she will do so immediately at the beginning of the period; if the caller chooses to wait, she will stay in the system for that period. As the caller waits, she may enter service in which case she incurs the waiting cost for that period, receives the reward associated with the service, and exits the queue. Otherwise, the caller incurs waiting cost for that period and then decides again whether to abandon or continue to wait as she enters the next period. We assume callers know the probability of receiving service in period t conditional on not being served yet. Callers also know that they will receive service before period T if they do not abandon. Furthermore, no information about the duration of waiting is conveyed to the callers, i.e. there is no delay announcement.

In our model, callers are forward looking. In each period, they compare the expected utility of waiting, which consists of utilities from the current and future periods and the expected utility of abandonment, and then choose the action that maximizes their expected utility. We assume the waiting cost is sunk, i.e. waiting costs incurred in the past are irrelevant to future decisions. Hence, a caller only considers the expected future utility associated with her actions.

We next describe the model primitives. Let c_i be caller i's cost of waiting for one period and r_i be caller i's reward from receiving service. The callers are heterogeneous in their rewards and waiting costs. More specifically, the reward r_i and the unit waiting cost c_i of caller i are given by

$$\begin{aligned} r_i &= \exp(m_r + \sigma_r y_{1i}), \\ c_i &= \exp(m_c + \sigma_c y_{2i}), \end{aligned} \tag{1}$$

where y_{1i} and y_{2i} are draws from independent and identical standard normal distributions. In other words, callers' reward and cost parameters have log-normal distributions. The parameters m_r and m_c are the means for $\ln(r_i)$ and $\ln(c_i)$, respectively. Similarly, the parameters σ_r and σ_c are the standard deviations for $\ln(r_i)$ and $\ln(c_i)$.[1]

The utility of caller i from choosing action d in period t is given by

$$u(t, r_i, c_i, \varepsilon_{it}(d), d) = v(t, r_i, c_i, d) + \varepsilon_{it}(d), \tag{2}$$

where $\varepsilon_{it}(d)$ denotes the idiosyncratic shock incurred by choosing action d. The term $v(t, r_i, c_i, d)$ is the nominal utility and is independent of the idiosyncratic stochastic shocks. We let $d = 1$ if a caller chooses to abandon in that period and zero otherwise.

Since caller i will exit the queue at the beginning of the period if she chooses to abandon, the nominal utility of caller i abandoning in period t is zero, i.e.

$$v(t, r_i, c_i, 1) = 0. \tag{3}$$

If caller i decides to wait, the nominal utility of waiting is given by

$$v(t, r_i, c_i, 0) = -c_i + \pi(t) r_i + (1 - \pi(t)) \mathbb{E}\left[\max_{d \in \{0,1\}} u(t+1, r_i, c_i, \varepsilon_{i(t+1)}(d), d)\right], \tag{4}$$

where $\pi(t)$ is the probability of receiving service in period t conditional on not being served yet and $\pi(T) = 1$, i.e. all callers receive service within T periods. We assume that $\pi(\cdot)$ is the equilibrium outcome of the system, where callers correctly anticipate the service probabilities based on their past experiences of contacting the call center. Furthermore, the probability of receiving service $\pi(\cdot)$ is common knowledge among the callers. The first term on the right-hand side of (4) is the waiting cost for the current period. The second term is the expected utility from receiving service in period t. Finally, the last term is the future value of waiting. We refer to the expectation in (4) as the integrated value function, denoted by $V(t, r_i, c_i)$. The expectation is taken with respect to the conditional distribution of $\varepsilon_{i(t+1)}$ given ε_{it}, where ε stands for $(\varepsilon(0), \varepsilon(1))$. Assuming $\varepsilon_{it}(d)$ is iid across different callers, periods and actions, we denote caller i's integrated value function as

$$V(t, r_i, c_i) = \int\int \max_{d \in \{0,1\}} u(t+1, r_i, c_i, \varepsilon(d), d) g(\varepsilon(0)) g(\varepsilon(1)) d\varepsilon(0) d\varepsilon(1), \tag{5}$$

where $g(\varepsilon(d))$ is the pdf of the error term $\varepsilon(d)$ for $d = 0, 1$.

Given r_i and c_i, caller i's optimal decision in period t is given by

$$d_{it} = \arg\max_{d \in \{0,1\}} u(t, r_i, c_i, \varepsilon_{it}(d), d). \tag{6}$$

The following proposition (see Appendix A for its proof) characterizes callers' choice probabili-

[1] The mean and standard deviation of callers' rewards are given by $\exp(m_r + \sigma_r^2/2)$ and $\exp(m_r + \sigma_r^2/2)\sqrt{\exp(\sigma_r^2) - 1}$, respectively. Similarly, for callers' costs, these statistics are $\exp(m_c + \sigma_c^2/2)$ and $\exp(m_c + \sigma_c^2/2)\sqrt{\exp(\sigma_c^2) - 1}$.

ties under the assumption that the idiosyncratic shocks have iid type-I extreme value distribution. (see Apendix A for the definition of this distribution). As explained in Rust (1987), this distributional form enables a closed form representation of the choice probabilities.

Proposition 1. *Suppose that the idiosyncratic shocks $\varepsilon_{it}(1)$ and $\varepsilon_{it}(0)$ have iid type-I extreme value distribution. Denoting by $P_{it}(d_{it}; r_i, c_i)$ the probability that caller i chooses action d_{it} in period t, we have*

$$P_{it}(d_{it}; r_i, c_i) = \frac{\exp(v(t, r_i, c_i, d_{it}))}{1 + \exp(v(t, r_i, c_i, 0))}, \tag{7}$$

where

$$v\left(t, r_i, c_i, d_{it}\right) = \begin{cases} 0 & \text{if } \; d_{it} = 1, \\ -c_i + \pi(t) r_i + (1 - \pi(t)) V(t, r_i, c_i) & \text{if } \; d_{it} = 0. \end{cases} \tag{8}$$

Moreover, caller i's integrated value function for $t < T$ is recursively given by

$$V\left(t, r_i, c_i\right) = \log\Big(1 + \exp(-c_i + \pi(t+1) r_i + (1 - \pi(t+1)) V(t+1, r_i, c_i))\Big), \tag{9}$$

and $V(T, r_i, c_i) = 0$.

4 Data

Our data set was generously made available to us by the Service Enterprise Engineering (SEE) lab at the Technion (http://ie.technion.ac.il/Labs/Serveng/). It contains individual call level data as well as agent data from a bank call center for a six month period between April and September 2008. The Call Center operates 24 hours per day, 7 days a week. It processes up to 85,000-90,000 calls a day on weekdays and 15,000-40,000 calls a day on weekends. There are 300-350 agents working in the Call Center on weekdays and 50-175 agents during weekends.

Around 30,000-35,000 calls or 35%-40% of total arrivals are routed according to the agent skills. The rest are IVR/VRU (interactive voice response/voice response unit, representing automated response) calls. The center offers six types of services: private, securities, internet, other languages, loans and solutions. The service type of a call can be observed in the data. Private calls (retail banking) are the largest call type. These are the calls we focus on in our basic analysis. A preliminary look at the data indicates that working days and weekends are significantly different in terms of call traffic, server numbers and wait patterns. In our analysis, we choose to focus on the working day calls.

The data traces each call from its entry to exit. Each call is broken down into subcalls. Entry and exit times from each subcall are available. Calls are distinguished by the route they follow within the call center (directly joining the queue, VRU and then joining the queue, other), and by the outcome of the call (normal termination, transfer, disconnected, on ring, no agent, abandoned short, abandoned, other unhandled). Calls joining the queue directly represent calls transfered from the branches or calls whose customer ID has not been identified. A definition of each outcome is provided in Table 1. Our analysis focuses on the route VRU and joining the queue, and the outcomes normal termination, transfer, abandoned short and abandoned, which consist of more than 80% of the observations. Since our model does not consider multi-stage service and intermediate waits by customers, we restrict our analysis to the first subcall, which consists of waiting in the queue and talking to the first agent. The callers do not receive any delay announcements, but may receive information announcements (working hours, etc.) and marketing announcements, or listen to music.

Code	Outcome	Description
1	Normal terminations	The caller receives service and then terminates the call
2	Transfer	Call was transferred to another agent or unit
3	Disconnected	Customer has terminated the call while on hold or because the agent made log-off
4	On Ring	Agent did not pick up the phone
5	No Agent	Agent has finished his shift without log-off from and phone system continues sending the incoming calls to this agent
11	Abandoned Short	A call placed into queue was abandoned with the wait time less than 5 seconds
12	Abandoned	A call placed into queue was abandoned with wait time larger or equal 5 seconds
13	Other Unhandled	A call placed into queue did not reach the agent by unknown reason (mainly because of hardware malfunctioning)

Table 1: Definitions of the outcomes.

Customers in this call center have different priorities in the queue. There are four levels of priority: high, medium, low, and no priority. The no priority calls are those that cannot be associated with a customer at the point of entry, and are thus treated as having no priority, which corresponds to the lowest priority. We observe the priority group of each caller from the data.

Depending on the caller's priority type, each caller receives a priority point upon arrival. The priority point of a customer is updated dynamically as the customer waits in queue. These priority updates are performed after every 60 seconds of waiting. The updates in priority points occur such that higher priority calls receive higher increases in their priority points relative to lower priority calls. While for the same waiting duration a call with a higher priority type always has higher priority points, a lower priority type caller who has waited a long time may have higher priority points than a newly arriving high type caller due to the dynamic priority point updates. We observe the effect of these dynamic priority increases in waiting time histograms. In particular, we observe peaks at multiples of 60 seconds, corresponding to the dynamic priority point updates. An example for medium priority calls on May 12, 2008, in Figure 1 illustrates the pattern.

Dynamic priority updates are not recorded in the data. Although we know the update mechanism, we do not make use of this in our estimation. Rather, we use the resulting service probabilities directly as estimated from the data. This estimation is described in the following section.

The average arrival pattern for calls on working days are shown in Figure 2. In order to focus on the relatively busy hours of the day, we restrict our analysis to calls between 9 a.m. and 2 p.m. on each working day.

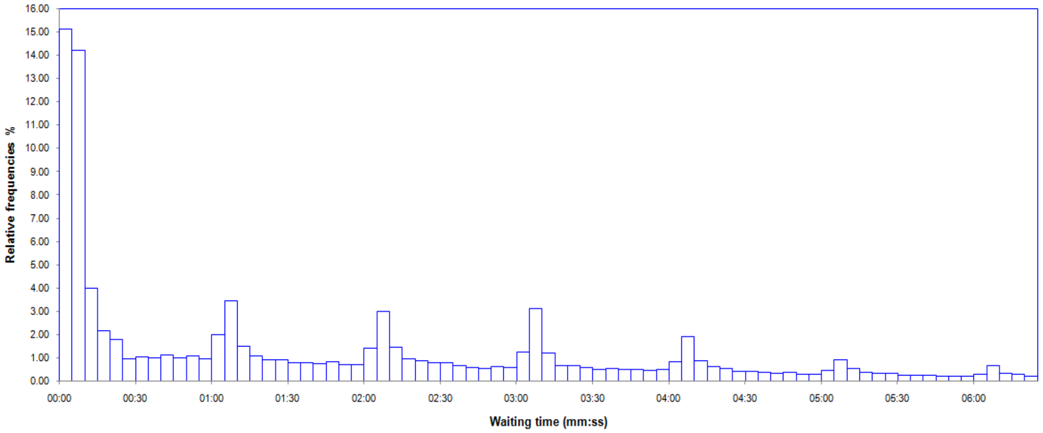

Figure 1: Waiting time histogram for medium priority calls on May 12, 2008.

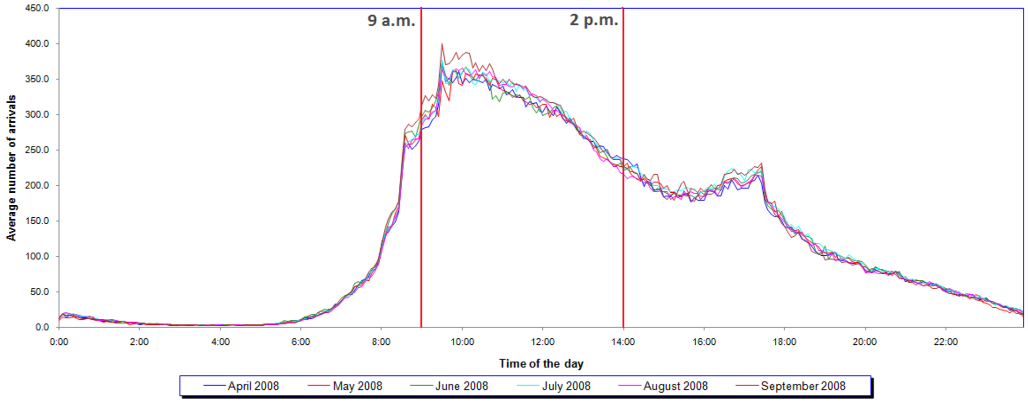

Figure 2: Average arrival pattern for calls in the working days.

The abandonment rate during the day for July 17, 2008, is plotted in Figure 3. This pattern suggests that there may be different staffing patterns during different times of the day. We verify this by making use of available agent data. Figure 4 provides average staff numbers during the day on Mondays (other working days exhibit a similar pattern), showing that the time interval we focus on represents a highly staffed interval, thus ensuring reasonable abandonment rates.

Finally, we focus on calls with a wait duration ranging between zero and 960 seconds. Calls with waiting times longer than 960 seconds constitute fewer than 0.01% of our observations, and have been eliminated to reduce the length of the time horizon in our estimation. Data from weeks with a holiday were excluded from the analysis (these are April 20-26, May 4-10, June 8-14, and September 28-30) as potential outliers.

In summary, our analysis focuses on 1,323,071 calls with the private service type, received on working days during weeks without a holiday in the interval April-September 2008, between 9 a.m. and 2 p.m., having entered the system through the VRU and proceeded to a wait in the queue, and having normal termination, transfer, short abandonment and abandonment as an outcome. We are

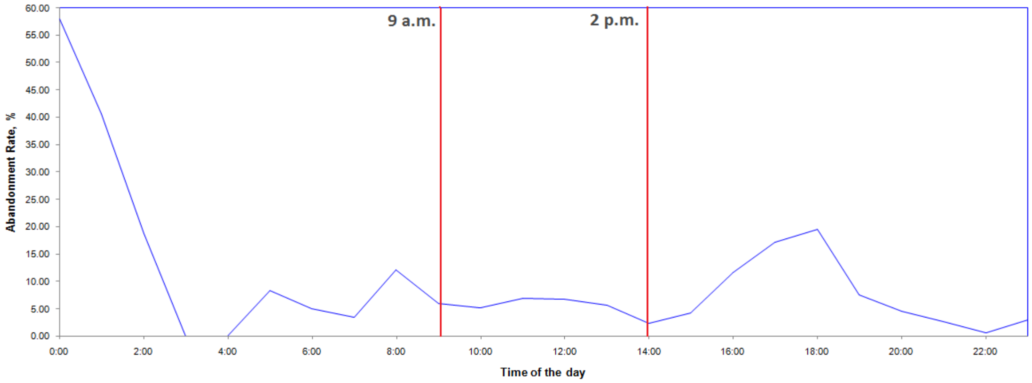

Figure 3: The pattern of average abandonment rate for July 17, 2008.

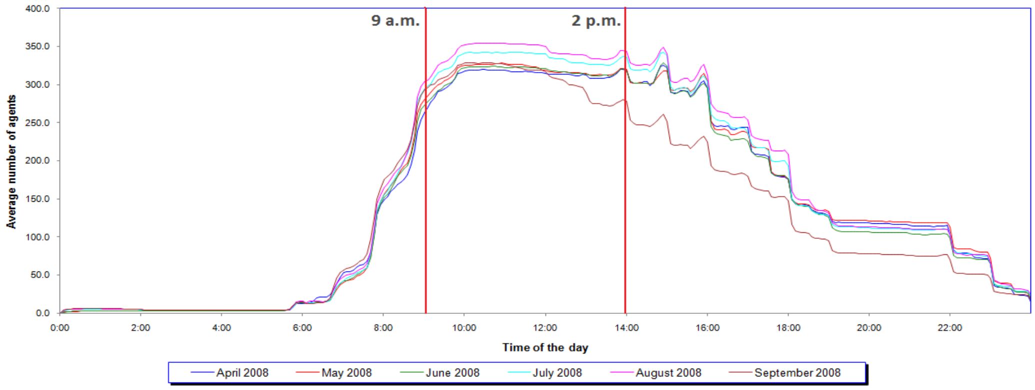

Figure 4: Average agent numbers on Mondays.

focusing on the subcall starting with the wait in the queue and including the encounter with the first agent. The summary statistics for this portion of the data are given in Table 2.

Priority group	Number of observations	Abandonment rate	Average waiting time (sec.)	Average waiting time (abandoned calls) (sec.)	Maximum waiting time (sec.)
High priority	184,722	2.12 %	18.83	71.73	857
Medium priority	516,685	3.68 %	42.19	108.58	958
Low priority	253,963	6.66 %	72.02	123.25	949
No priority	367,701	24.65 %	96.20	100.31	960
Sum	1,323,071				

Table 2: Summary statistics for the portion of the data used in the analysis.

5 Estimation

In this section, we first discuss the identification of callers' parameters from the data. Next, we describe the estimation methodology and results. Finally, we discuss the cross validation and out-of-sample tests to examine the ability of our model to predict callers' abandonment behavior.

Identification. As will be seen below (Figure 5), our data exhibit significant intertemporal variation in the service probabilities $\pi(t)$. This observation along with the fact that waiting times vary across different callers (see Figure 7) allows us to identify the reward and cost parameters, separately. To see the intuition behind this, note from equations (7)-(9) of Proposition 1 that the probability of abandoning in period t depends on the terms $\{\pi(s)r_i - c_i\}_{s=t..T}$, and is given by

$$P_{it}(1; r_i, c_i) = \psi_t(\pi(t)r_i - c_i, \pi(t+1)r_i - c_i, ..., \pi(T)r_i - c_i), \tag{10}$$

where ψ_t is a suitably defined function that does not depend on r_i and c_i. Equation (10) shows that if there were no variation in $\pi(t)$, i.e. $\pi(t) = \pi$ for all t, then we could only identify the difference between πr_i and c_i. This follows because the abandonment probability $P_{it}(1; r_i, c_i)$ would then be solely a function of $\pi r_i - c_i$, which would prevent the identification of the reward and cost parameters separately. However, since callers' waiting time exhibits sufficient variability, we can identify the abandonment probabilities in each period as given in (10) from which we can identify the reward and cost parameters separately given the intertemporal variation in service probabilities $\pi(t)$.

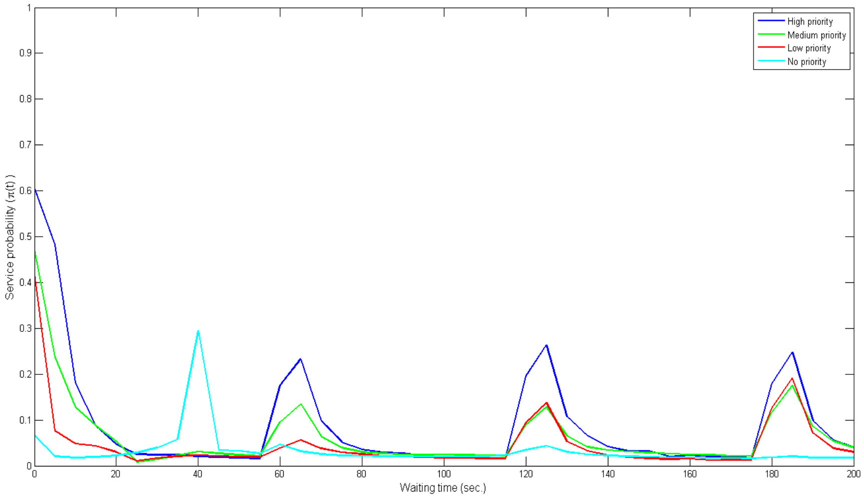

Figure 5: The service probabilities of the priority groups in the data.

Moreover, heterogeneity in callers' cost and reward (i.e. σ_r, σ_c) is identified by the variation in the abandonment behavior of callers in a given period. To see this, consider N callers who have waited for t periods; and recall that the abandonment probability in period t is given by (7). If there is no heterogeneity (i.e. $\sigma_r = \sigma_c = 0$) then each caller has the same abandonment probability. Hence, the total number of abandonments in period t is a Binomial random variable. In contrast, under heterogeneity, callers will have different abandonment probabilities, and total number of abandonments in period t is the sum of N binary random variables where success probabilities are random variables (as determined through r_i, c_i in equations (1) and (7)). Therefore, the total number of abandonments in period t exhibits more variation under heterogeneity. In other words, the degree of variation (or volatility) in callers' abandonment behavior helps us identify the variance parameters σ_r, σ_c. Nonetheless, note that our model is flexible enough to allow $\sigma_r = \sigma_c = 0$. Indeed, we find that this is the case for all but no-priority callers (Table 3).

Estimation Methodology and Results. The estimation of callers' parameters is carried out in two stages. We first estimate the probability of receiving service $\pi(t)$. Next, given the probability of receiving service, we construct the likelihood function of callers' observed decisions in the data and maximize it to estimate the parameters.[2]

We estimate $\pi(t)$, the probability of receiving service in period t conditional on not being served yet, directly from the data. This direct approach allows us to capture all operational aspects of the call center for the interval of analysis. Given the cumulative distribution of a caller's waiting time (time spent in the queue before receiving service), denoted by F, $\pi(t)$ is given by

$$\pi(t) = \frac{F(t+1) - F(t)}{1 - F(t)}. \tag{11}$$

To estimate the waiting time distribution of the callers, we use the Kaplan-Meier estimator (Kaplan and Meier (1958)). This estimator is used to find the survival time distribution when the data is censored, which is the case herein due to the presence of abandonments. See Appendix B for details.

Next, we describe the maximum likelihood estimation problem of callers' parameters given the service probabilities $\pi(t)$ for $t \geq 1$. Recall that callers are indexed by $i = 1, ..., N$, where N is the total number of callers in the data, and that r_i and c_i are given in (1) where y_{1i} and y_{2i} are standard normal random variables. Let τ_i denote the last period in which caller i decides between waiting and abandoning. Also let $\{d_{it} : t = 0, 1, ..., \tau_i\}$ denote the observed actions of caller i where d_{it} is the action of caller i in period t.

Recall that $P_{it}(d_{it}; r_i, c_i)$ denotes the probability of choosing the action d_{it} by caller i in period t. Let $\Theta = (m_r, m_c, \sigma_r, \sigma_c)$ denote the vector of structural parameters to be estimated. With the assumption that r_i and c_i have lognormal distributions, the likelihood of observing the sequence of actions $\{d_{it} : t = 0, 1, ..., \tau_i\}$ by caller i is given by

$$\begin{aligned} \ell_i(\Theta) &= \int\int \prod_{t=0}^{\tau_i} P_{it}(d_{it}; r_i, c_i)\phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i} \\ &= \int\int \prod_{t=0}^{\tau_i} P_{it}(d_{it}; \exp(m_r + \sigma_r y_{1i}), \exp(m_c + \sigma_c y_{2i}))\phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i}, \end{aligned} \tag{12}$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution. The likelihood function of the entire

[2] This is similar to the approach taken in Rust (1987), where the author first estimates the transition probabilities in mileage directly from the data, and then uses those fixed transition probabilities to estimate the structural parameters.

sample is then the product of individual caller's likelihood and is defined as follows:

$$
\begin{aligned}
L(\Theta) &= \prod_{i=1}^{N} \ell_i(\Theta) \\
&= \prod_{i=1}^{N} \int\int \prod_{t=0}^{\tau_i} P_{it}(d_{it}; \exp(m_r + \sigma_r y_{1i}), \exp(m_c + \sigma_c y_{2i}))\ \phi(y_{1i})\ \phi(y_{2i})\ dy_{1i}\ dy_{2i}.
\end{aligned}
\tag{13}
$$

The estimation problem is to choose the structural parameters Θ to maximize the log-likelihood function $\log L(\Theta)$ with the integrated value function (9) as constraints (Su and Judd (2012)). To be more specific, the formulation of the estimation problem is given below.

$$
\begin{aligned}
\underset{\Theta, V(t,\cdot,\cdot)}{\text{maximize}} \quad & \log L(\Theta) = \sum_{i=1}^{N} \log \left(\int\int \prod_{t=0}^{\tau_i} P_{it}(d_{it}; r_i, c_i) \phi(y_{1i}) \phi(y_{2i}) dy_{1i} dy_{2i} \right) \\
\text{subject to} \quad & \text{for all } i = 1, \ldots, N : \\
\forall\, t: \quad & P_{it}(d_{it} = 1; r_i, c_i) = \frac{1}{1 + \exp(-c_i + \pi(t) r_i + (1 - \pi(t)) V(t, r_i, c_i))}, \\
\forall\, t: \quad & P_{it}(d_{it} = 0; r_i, c_i) = \frac{\exp(-c_i + \pi(t) r_i + (1 - \pi(t)) V(t, r_i, c_i))}{1 + \exp(-c_i + \pi(t) r_i + (1 - \pi(t)) V(t, r_i, c_i))}, \\
\forall\, t: \quad & V\left(t, r_i, c_i\right) = \log\Big(1 + \exp(-c_i + \pi\left(t+1\right) r_i + (1 - \pi\left(t+1\right)) V\left(t+1, r_i, c_i\right))\Big), \\
& V(T, r_i, c_i) = 0, \\
& r_i = \exp(m_r + \sigma_r y_{1i}), \\
& c_i = \exp(m_c + \sigma_c y_{2i}), \\
& \sigma_r,\ \sigma_c \geq 0.
\end{aligned}
\tag{14}
$$

In the estimation, we assume that each caller makes the decision every five seconds. Thus, the maximum number of periods in our model is 192 (= 960/5). Since our data is more granular, we truncate the abandonment times downward and service initiation times that happen in a period upward, consistent with our modeling assumptions in Section 3.

We solve the maximum likelihood estimation problem (14) using the nonlinear optimization solver, KNITRO (Byrd et al. (2006)) with AMPL interface. We use 50 randomly generated starting points for finding a better estimate. To approximate the 2-dimensional integration in the likelihood function over y_{1i} and y_{2i}, we use Gauss-Hermite integration (Judd (1998)). We choose 5 points in each dimension and approximate the integral by the weighted sum of the likelihood values at the resulting 25 nodes in the 2-dimensional space associated with the pair (y_{1i}, y_{2i}). For further details, see Appendix C. We also conduct a Monte-Carlo experiment to show that our estimation method can recover the true parameter values; see appendix D for details.

Our empirical analyses focus on four priority groups within the private service group as described in Section 4. For each priority group, the corresponding probability of service $\pi(t)$ is estimated directly from the data. Note that the direct estimation of the service probabilities $\pi(\cdot)$ allows us to capture the interaction between the different priority callers in the queue. We estimate the parameters of each priority group separately. The estimated parameter values and standard errors (shown in parentheses) are reported in Table 3. To compute standard errors, we use the parametric bootstrap method (Horowitz (2001)). We generate 100 simulated data sets with the same size as the real data from the estimates. We then estimate parameters of the simulated data sets and compute the standard errors. We report in Table 4 the mean and standard deviation for callers' rewards and costs for each priority group, which are calculated from the estimates in Table 3 using the formulas in Footnote 1.

Priority group	m_r	m_c	σ_r	σ_c
High Priority	1.842 (0.011)	–2.420 (0.089)	7.16E–06 (0.028)	2.89E–05 (0.156)
Medium Priority	1.820 (0.009)	–3.166 (0.070)	7.39E–06 (0.027)	5.46E–05 (0.140)
Low Priority	1.667 (0.006)	–10.000 (1.517)	5.69E–06 (0.032)	1.09E–03 (0.912)
No Priority	1.426 (0.006)	–7.420 (0.219)	0.152 (0.006)	2.379 (0.079)

Table 3: The estimation results.

Priority group	r-Mean ($)	c-Mean ($/minute)	r-St.Dev.	c-St.Dev.
High Priority	6.309	1.067	4.52E–05	3.09E–05
Medium Priority	6.175	0.506	4.56E–05	2.76E–05
Low Priority	5.299	5.45E–04	3.02E–05	5.91E–07
No Priority	4.211	0.122	0.645	2.057

Table 4: The mean and standard deviation for callers' rewards and costs.

As can be seen in Table 4, mean reward parameters increase with the priority level although they are comparable in magnitude. Similarly, the mean cost parameters are higher for the high and medium priority groups. This suggests the high-priority callers are less patient. The waiting cost is negligible for the low priority group. Recall that the maximum waiting time in the data is about 15 minutes. The negligible cost parameter for the low priority callers suggests that they do abandon not because of high waiting costs but rather because of external events as modeled by the random shocks. Interestingly, the waiting cost for the no-priority callers is nonzero. Recall that the no-priority calls cannot be associated with a customer at the point of entry, and hence, contains a mix of delay-sensitive and non-delay-sensitive callers. The mean waiting cost captures the average of this heterogeneous group and is therefore, higher than that of the low priority group. Note, however, that our random-coefficients model is rich enough to accurately capture this heterogeneity and reflects its implications in the counterfactual analysis.

The negligible variance estimates for the high, medium, and low priority groups suggest that callers in these group are rather homogeneous. In other words, the call center provider was successful in segmenting the callers to these groups. On the other hand, the estimates for the no priority group suggest significant heterogeneity within this group, which is consistent with the fact that callers in this group are not identified by the system and may also be new customers. This observation calls for further efforts to better identify and segment the no-priority group.

Cross Validation and Out-of-sample Tests. We use 10-fold cross validation with stratification to examine the ability of the model to predict the abandonment behavior of the callers (Kohavi (1995)). The validation is done for each priority group in isolation.

Let $P_{aban}(t)$ denote the ex-ante probability of abandoning in period t, which is given by

$$P_{aban}(t) = \begin{cases} (1-F(0)) \int\int P_{i0}(1; r_i, c_i)\phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i} & \text{if } t = 0, \\ (1-F(t))\Big(\int\int \Big(\prod_{s=0}^{t-1} P_{is}(0; r_i, c_i)\Big) P_{it}(1; r_i, c_i)\phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i}\Big) & \text{if } t > 0. \end{cases} \tag{15}$$

The callers' decisions to abandon in each period are independent from each other. Therefore, the predicted number of abandonments in period t has a binomial distribution. Let $m_{aban}(t)$ and $\sigma_{aban}(t)$ denote the mean and standard deviation of this distribution, respectively. Then $m_{aban}(t) = NP_{aban}(t)$ and $\sigma_{aban}(t) = \sqrt{NP_{aban}(t)(1 - P_{aban}(t))}$. Moreover, the predicted number of aggregate abandonments is $\sum_{t=0}^{T-1} m_{aban}(t) = N \sum_{t=0}^{T-1} P_{aban}(t)$. Denote by $a_{aban}(t)$ the actual number of abandonments in period t.

We consider the relative and absolute errors in predicting the aggregate abandonment rates as the performance metrics for the cross validation. Note that

$$\text{Relative Error} = \frac{|\sum_{t=0}^{T-1} m_{aban}(t) - \sum_{t=0}^{T-1} a_{aban}(t)|}{\sum_{t=0}^{T-1} a_{aban}(t)}, \tag{16}$$

$$\text{Absolute Error} = \frac{1}{N}|\sum_{t=0}^{T-1} m_{aban}(t) - \sum_{t=0}^{T-1} a_{aban}(t)|. \tag{17}$$

The averages of the performance metrics across all test sets are shown in Table 5, which show that our estimates are fairly accurate.

Priority group	Relative Error	Absolute Error
High Priority	0.29 %	6.15E–03 %
Medium Priority	0.05 %	1.86E–03 %
Low Priority	0.04 %	2.35E–03 %
No Priority	0.15 %	0.03 %

Table 5: The averages of the performance metrics across all test sets.

A more detailed comparison of the predicted and actual abandonments is provided in Figure 6. In addition to $m_{aban}(t)$ and $\sigma_{aban}(t)$, it also shows $m_{aban}(t) \pm 2\sigma_{aban}(t)$ over time, which helps assess the accuracy of the prediction in relation to the inherent variability of the abandonments, as captured by $\sigma_{aban}(t)$.

In addition to the cross validation study, we also perform several out-of-sample tests to illustrate the accuracy of the estimation. To this end, we first split the data across weeks into two samples. The first half is used to estimate the model, whereas the second half is used for prediction and testing its accuracy. The results for different priority groups are shown in Table 6.

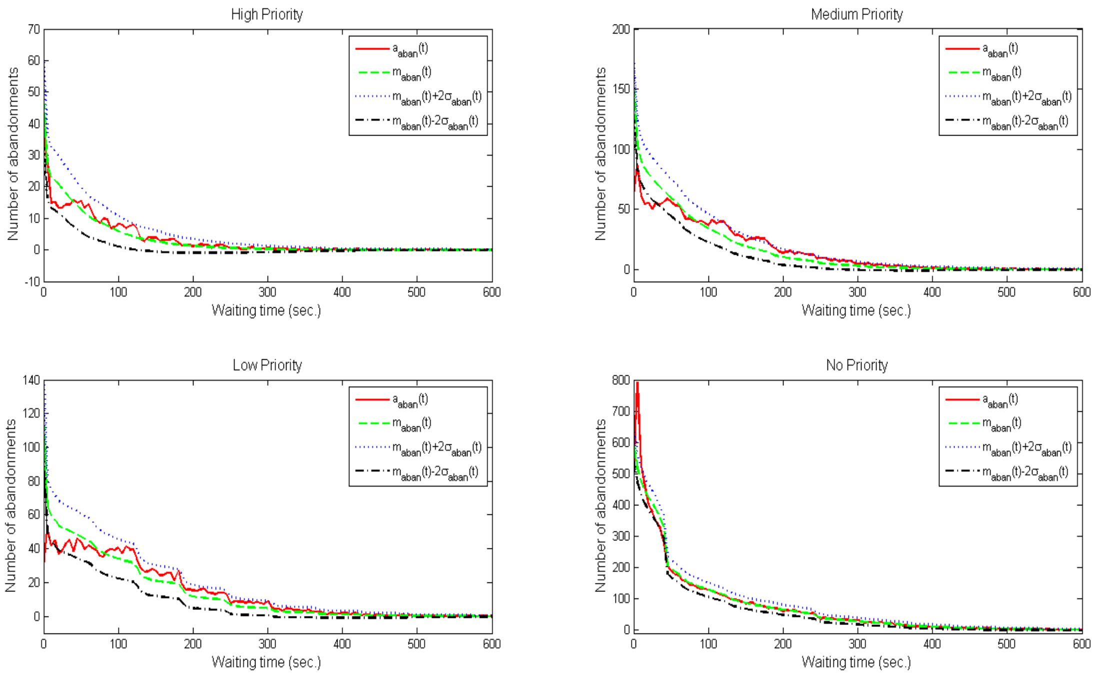

Figure 6: The predicted abandonments $m_{aban}(t)$, actual abandonments $a_{aban}(t)$ and $m_{aban}(t) \pm 2\sigma_{aban}(t)$ over time for the priority groups.

Priority group	Relative Error	Absolute Error
High Priority	20.02 %	0.38 %
Medium Priority	8.91 %	0.26 %
Low Priority	5.47 %	0.32 %
No Priority	3.72 %	1.03 %

Table 6: Out-of-sample testing of the model across different weeks.

As can be seen in Table 6, the estimates from the first half of the data produce fairly accurate predictions for the abandonments observed in the second half of the data. It is interesting to note, however, that because the abandonment rate is small for the high priority group, even a small prediction error is magnified under the relative error metric. Hence, although the relative error may seem high for the high priority group, the corresponding absolute error is small (Table 6).

Next, we repeat the out-of-sample testing for different hours of the day. We use peak-hours data (9 am-2 pm) to estimate the parameters reported in Tables 3 and 4, and use those parameters to predict the abandonments during off-peak hours.[3] More specifically, we consider two off-peak periods: 2-6 pm and 6-10 pm. The prediction results for 2-6 pm and 6-10 pm are shown in Tables 7 and 8, respectively.

Although the predictions of the model for 2-6 pm (based on the peak-hours estimates) are accurate (see Table 7), they are not as accurate for 6-10 pm. However, this discrepancy can be explained by

[3] In the prediction, the service probabilities $\pi(t)$ for the relevant hours are used.

Priority group	Relative Error	Absolute Error
High Priority	14.00 %	0.46 %
Medium Priority	4.04 %	0.18 %
Low Priority	10.90 %	0.76 %
No Priority	9.40 %	2.30 %

Table 7: Out-of-sample testing of the model across peak (9 am-2 pm) versus off-peak (2-6 pm) hours.

Priority group	Relative Error	Absolute Error
High Priority	50.77 %	2.99 %
Medium Priority	15.69 %	1.41 %
Low Priority	16.21 %	2.11 %
No Priority	12.77 %	5.36 %

Table 8: Out-of-sample testing of the model across peak (9 am-2 pm) versus off-peak (6-10 pm) hours.

the differences in caller demographics in different hours. Our hypothesis is that the callers in 2-6 pm are similar to the callers in peak hours, whereas those calling during 6-10 pm are less similar to the peak-hour callers. Therefore, the reward and cost parameters and, consequently, the abandonment behavior of the callers during peak hours are more similar to those of the callers who contacted during 2-6 pm.

To asses the similarity of callers during different hours, we adopt the Bhattacharyya distance (between probability distributions), which is widely used in the information theory literature (Bhattacharyya (1943)). To be specific, for discrete probability distribution p and q over the domain X, the Bhattacharyya distance is given by

$$D_B(p,q) = -\ln(\sum_{x\in X}\sqrt{p(x)q(x)}),$$

see for example, Kailath (1967) and Basseville (1989). In our context, probability distributions p and q correspond to the identity of a random caller during peak hours and off-peak hours, respectively. To be more specific, $p(x)$ denotes the probability that a randomly selected peak-hour caller is caller x. In our data set callers in high, medium and low priority groups are identified. Therefore, we can calculate the distance for those priority groups to assess the similarity of the callers in different hours as shown in Table 9.

Priority group	Distance between peak callers and callers in 2-6 pm	Distance between peak callers and callers in 6-10 pm
High Priority	0.20	0.48
Medium Priority	0.29	0.47
Low Priority	0.35	0.53

Table 9: Bhattacharyya distance for comparing caller similarity in peak versus off-peak hours.

Interestingly, the distances in Table 9 show that callers during peak hours are more similar to those calling during 2-6 pm (in the sense of overlap) than those calling in 6-10 pm. This explains the contrast between the prediction accuracies reported in Tables 7 and 8. In conclusion, the out-of-sample tests provide further support that our model gives good predictions provided that the caller demographics in the two samples are similar.

Building on the estimation results, in the next section, we provide a counterfactual analysis to assess the impact of policy changes.

6 Counterfactual analysis

This section provides a simulation study of the call center using the estimated reward and cost parameters. The ultimate objective is to perform what if analyses to assess the impact of changes in service discipline. The aforementioned assumptions will be maintained throughout this section unless stated otherwise.

The simulation study is constructed along the lines of the usual discrete-event simulations. However, it has a novel feature in that the callers decide dynamically to abandon or to continue to wait by computing their expected utilities under each choice. Consistent with Section 5, in the simulation the callers make their decision to wait or to abandon every 5 seconds. Consequently, the unit of time in the simulation is 5 seconds. The expected utility computation on the callers' part requires the knowledge of the equilibrium service probabilities $(\pi(t),\ t \geq 0)$. Although these can be computed readily from the data for the current service discipline, they need to be recalculated when a new service discipline is considered. Computing these equilibrium service probabilities (for a new policy) seems intractable analytically. Therefore, we use the following iterative procedure which seems to work well. First, given a new policy we simulate the system as if no one abandons to obtain an estimate of service probabilities $(\pi^0(t),\ t \geq 0)$. In the next step, we allow the callers to abandon using $\pi^0(\cdot)$, and simulate the system to get the new estimates of service probabilities $(\pi^1(t),\ t \geq 0)$. We repeat this procedure until the average waiting time and abandonment rate converge for all priority groups.

The first step of the simulation study is to reconstruct the existing *as is* performance of the call center. However, there are challenges to perform this task accurately. The difficulty stems in part from the variation in the data across different days (due to inherent uncertainty). Therefore, we choose to replicate the aggregate performance over all days, which presents challenges too, mainly because it is not immediately clear what number of agents should be used in our simulation. In particular, the number of agents in the data vary across days (and hours within a day). Moreover, the agents handle not just the first subcalls we focus on but also the subsequent subcalls (in addition to other types of calls we do not consider). Consequently, to determine the number of agents we vary the number of agents between 105 and 165. We pick the number of agents to be 133, for which the waiting time and abandonment statistics are closest to those in the data.[4] In what

[4] Under each staffing level being considered, the waiting time and abandonment rate for each priority group is simulated. These values are compared to the values observed in the aggregated data, and a weighted relative error, where weights are taken as the size of the priority groups relative to the size of the entire data, is considered as the comparison metric.

follows, this constitutes the base case for our simulation study.

The simulated system consists of four queues and one pool of agents. Each queue corresponds to one of the priority groups in the data. Recall that the service discipline currently used by the call center is a periodic point-update priority policy. In the simulation, this policy is used to determine priority points of callers as a function of their priority type and waiting time. The periodicity of the priority updates is also reflected in the waiting time histograms of the various groups as shown in Figure 7. The simulation of the current policy yields a similar pattern of periodicity, cf. Figure 8.

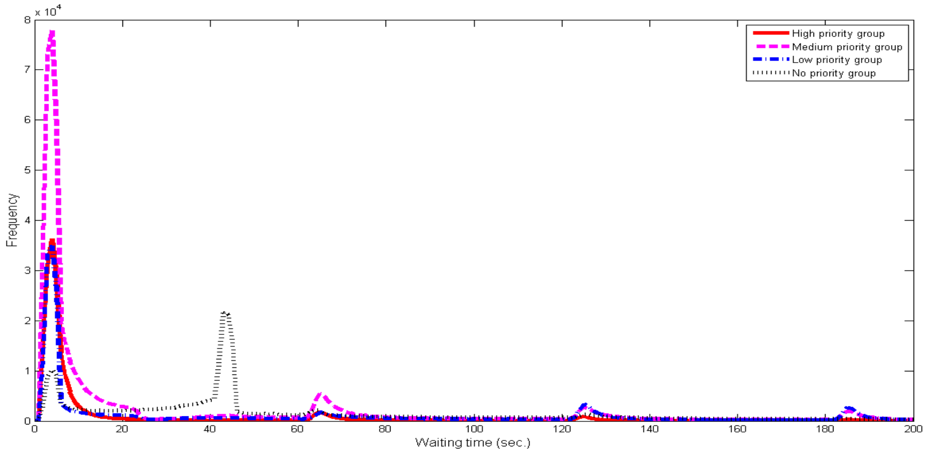

Figure 7: The waiting time histograms of the priority groups in the data.

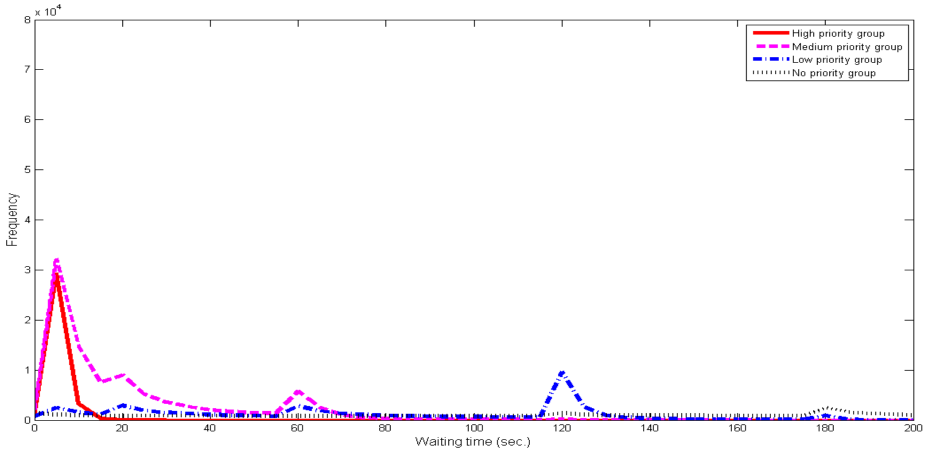

Figure 8: The waiting time histograms of the priority groups in the simulation results of the current policy.

To illustrate the usefulness of our approach, we consider assessing the impact of policy changes to the service discipline. To this end, in addition to the current policy we consider the following policies:[5] First-come-first-served (FCFS) policy, a static (and non-preemptive) priority policy, a threshold policy and the reversed strict priority policy. Under the FCFS policy, calls are served in the order they arrive irrespective of their group. The static priority policy gives the highest priority to the "high-priority" group, next to the "medium-priority" group, then to the "low-priority" group. The lowest priority is given to the "no-priority" group. The threshold policy acts like the static priority policy when the number of "no-priority" calls waiting is less than or equal to the threshold. Otherwise, the "no-priority" calls have the highest priority while the other groups preserve their relative priority levels amongst themselves. Finally, under the reversed strict priority policy, the priority order of the groups in the static priority policy is reversed. The average waiting times and the abandonment rates under these policies are given in the top panel of Table 10. As expected the waiting times under the FCFS policy are similar across different priority groups though the abandonment rates differ.

Endogenous model								
Policy	High priority		Medium priority		Low priority		No priority	
	sec.	% Ab.	sec.	% Ab.	sec.	% Ab.	sec.	% Ab.
Current policy	5.49	0.22	17.13	0.91	64.56	6.25	147.29	37.67
FCFS policy	80.67	7.51	83.23	5.39	83.47	8.18	75.78	23.32
Static priority policy	5.46	0.22	8.47	0.39	24.39	2.46	183.69	41.77
Threshold policy (th = 15)	7.02	0.28	17.82	0.90	265.50	21.55	82.16	24.97
Threshold policy (th = 5)	7.66	0.34	23.36	1.17	517.47	37.78	32.59	11.10
Reversed strict priority policy	89.06	62.98	41.08	2.45	7.68	0.77	5.46	1.94

Exogenous model								
Policy	High priority		Medium priority		Low priority		No priority	
	sec.	% Ab.	sec.	% Ab.	sec.	% Ab.	sec.	% Ab.
Static priority policy	5.37	0.73	8.26	0.86	23.64	2.48	160.61	40.51
Threshold policy (th = 15)	7.07	0.91	17.91	1.89	236.75	24.42	82.49	20.70
Threshold policy (th = 5)	7.54	1.09	22.54	2.31	366.47	38.29	32.61	8.28
Reversed strict priority policy	397.79	57.66	39.55	4.12	8.29	0.83	5.85	1.45

Table 10: Average waiting times and abandonment rates of different caller groups under various service disciplines for the endogenous and the exogenous models. For each group, the first and the second column show the average waiting times and the abandonment rates, respectively.

Recall that the callers are forward looking in our model and their behavior may change as the service discipline changes. To shed light on this, we also consider modeling callers' abandonment behavior using an exogenous time-to-abandon distribution. To this end, we first estimate the exogenous distribution from the data. Since the abandonments are censored (by callers' entering service), we use the Kaplan-Meier estimate. The hazard rates of time-to-abandon are as shown in Figure 9. Treating these as if they were constant, we use a geometric distribution for the time-to-abandon, where the probability of abandonment is estimated using the 25% quartile of the Kaplan-Meier estimate of the cumulative distribution function.[6] The lower panel of Table 10 shows the average

[5] We also considered changing the frequency and the size of priority point updates to the current policy. The frequency changes did not change the average waiting times or abandonment rates. The impact of changes to the size of updates were as one would predict.

[6] Brown et al. (2005) observes that the Kaplan-Meier estimates may be biased under heavy censoring. Therefore, following Brown et al. (2005), we use the first quartile when estimating the probability of abandoning.

waiting times and the abandonment rates resulting from the model with exogenous abandonment distribution under the static priority, the threshold and the reversed strict priority policies.

Comparisons between the endogenous model with strategic customers and the model using an exogenous time-to-abandon lead to the four major observations below.

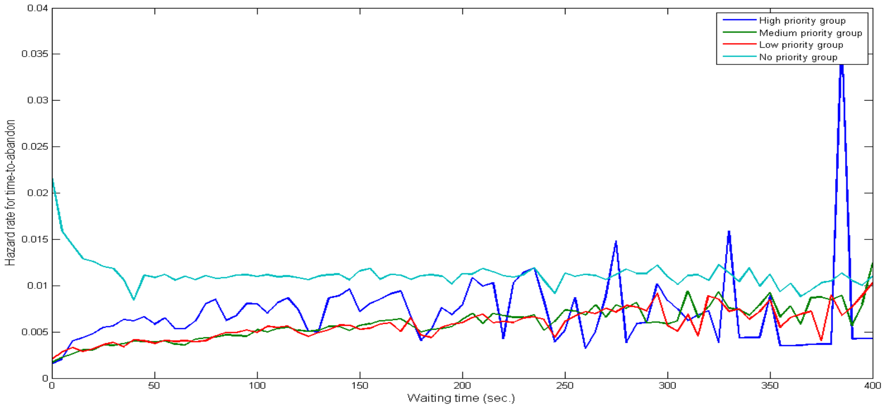

Figure 9: Hazard rates of the priority groups for time-to-abandon in the data.

First, if a caller has a negligible waiting cost, her probability of abandoning decreases as the service probability gets worse. For such callers, the exogenous model will underestimate waiting times relative to the endogenous model, under policies that deteriorate service probability for these callers.

As can be seen from Table 10, the "no-priority" group suffers from long waiting times and high abandonment rates under the current policy or the static priority policy. Recall that callers in this group are unidentified and some are new customers. Therefore, the call center may wish to improve the service quality they receive for retention purposes. Although there is a large number of alternatives for improving the service quality of "no-priority" callers, we focus attention on the threshold policy (described above) for simplicity. Setting the threshold at 15 improves the waiting times and lowers the abandonment rates somewhat for the "no-priority" group. For the "low-priority" group, this leads to significantly higher waiting times and abandonment rates, cf. the top panel of Table 10. (The impact on the other two groups is small.)

Under the threshold policy (with 15 as the threshold) the model with exogenous abandonment distribution underestimates the service degradation to the "low-priority" group (in terms of waiting times, 236.75 sec versus 265.50 sec); see Table 10.[7] Next, we clarify the source of discrepancy for the "low-priority" group (which sheds light onto what happens to other classes as well). Recall that the delay cost c is negligible for the "low-priority" group, cf. Table 4. Substituting $c = 0$ in Equation (9) shows that the integrated value function $V(t) > r$ for all t. Then it is straightforward to conclude from Equations (7)-(8) that as the service quality worsens (i.e. $\pi(t)$ decreases), the probability of abandoning decreases. Intuitively, as the service probability decreases the probabil-

[7] To test the significance of the differences between the results of the exogenous and the endogenous models, we use the two-sample t-test (Snedecor and Cochran (1989)). Under the threshold policy with 15 as the threshold for the average waiting time of the "low-priority" callers, the difference is significant with 90% confidence (t-statistic=1.75).

ity of getting served in later periods increases, and the callers are willing to wait longer to receive service (because their waiting costs are negligible). The decreased abandonment probability and the service degradation lead to higher queue lengths.

Given this observation, comparing the current policy with the threshold policy (with 15 as the threshold) reveals that the service quality gets worse for the "low-priority" calls when switching from the current policy to the threshold policy, cf. the top panel of Table 10. Hence, we expect lower probability of abandoning, i.e. callers abandon more slowly (especially early on). This, in turn, leads to longer queue lengths and waiting times. Figure 10 shows the abandonment probability of the "low-priority" callers under the current policy and the threshold policy with 15 as the threshold. On the other hand, in the model with exogenously given abandonment distributions, callers' abandonment probability is estimated from the current policy, which is higher than that in our model.[8] Thus, in the exogenous model callers abandon sooner which leads to lower waiting times. Therefore, the prediction of the model with exogenous abandonment distribution can be off substantially. This is demonstrated further by setting the threshold to 5 (366.47 sec versus 517.47 sec).[9]

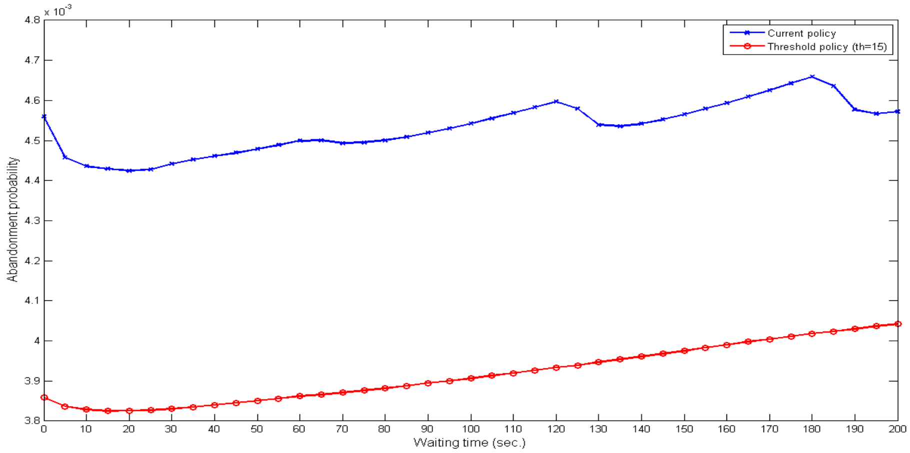

Figure 10: Abandonment probability of the low priority callers in the endogenous model under the current policy and the threshold policy with 15 as the threshold.

Second, if a caller has a significant waiting cost, her probability of abandoning increases as the service quality degrades. For such callers, the exogenous model will overestimate waiting times relative to the endogenous model, under policies that deteriorate service probability for these callers.

The fact that the probability of abandoning goes down as the service quality degrades for the "low-priority" group (in our model) may seem counterintuitive at first. However, what drives this is the fact that the delay cost for the "low-priority" group is negligible. Indeed, if callers have significant delay costs, the implication will be different. More specifically, for the "high-priority" group, the comparisons under the reversed strict priority policy show the model with exogenous

[8] Moreover, the abandonment probability estimated from the data is extrapolated beyond what is observed under the current policy.

[9] Under the threshold policy with 5 as the threshold for the average waiting time of the "low-priority" callers, the difference between the results of the exogenous and the endogenous models is significant with 90% confidence (t-statistic=9.28).

abandonment distribution significantly overestimates the waiting times and somewhat underestimates the abandonment rates (397.79 sec versus 89.06 sec and 57.66% versus 62.98%).[10]

Switching to the reversed strict priority degrades the service quality for the "high-priority" group. The model with exogenous abandonment distributions works precisely as explained above. In contrast, given positive delay costs, callers anticipate the significant future delay costs (embedded in the integrated value function) in our model and they choose to abandon early with high probability. This effect is illustrated in Figure 11, which shows the abandonment probability of the "high-priority" callers under the current policy and the reversed strict priority policy. This effect leads to significantly shorter queue lengths and waiting times than those in the model with exogenous abandonments. Comparing the abandonment rates in the two models requires trading off the counteracting forces: shorter queue lengths but a significantly higher probability of abandoning during each period in our model. The net effect leads to a higher, albeit comparable, abandonment rate in our model. This will be elaborated on further below.

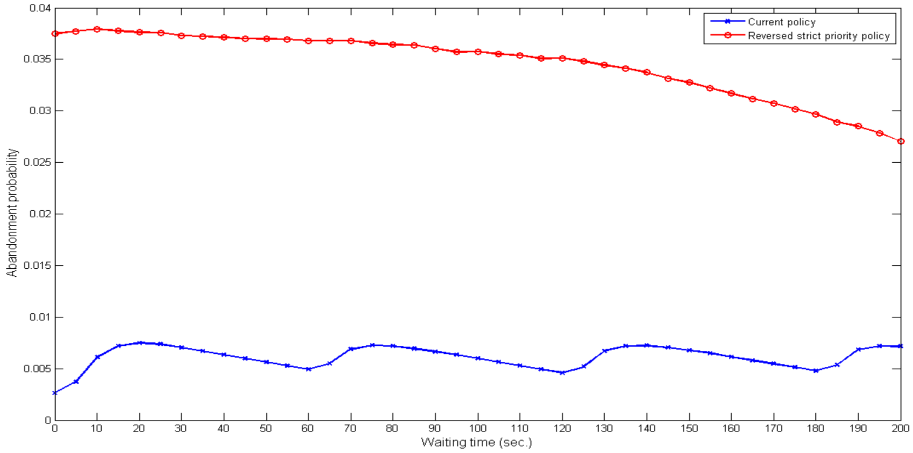

Figure 11: Abandonment probability of the high priority callers in the endogenous model under the current policy and the reversed strict priority policy.

Third, when there is heterogeneity in callers' waiting cost, both the first and the second observations made immediately above are present. In this case, it is the heterogeneity in the cost estimate and its composition that will determine which effect will dominate for such callers.

The comparison of the results for the "no-priority" group under the endogenous versus the exogenous abandonment time distribution reveals a surprising result and exemplifies the usefulness of the random coefficients model. Although the mean waiting cost of the "no-priority" group is positive, the callers in that group do not behave like the callers in the "high-priority" group, who have positive waiting costs too. Note, however, that the waiting cost for the "no-priority" group exhibits significant heterogeneity (whereas that for the "high-priority" group does not). Indeed,

[10] Under the reversed strict priority policy for the average waiting time of the "high-priority" callers, the difference between the results of the exogenous and the endogenous models is significant with 90% confidence (t-statistic=29.87). For the abandonment rate of the "high-priority" callers, the difference is significant with 90% confidence (t-statistic=2.99).

the “no-priority” group can be seen as a mix of callers from “low-priority” and “high-priority” groups qualitatively as far as their delay cost is considered. Hence, we expect to see a decrease in the abandonment probability for those callers who have negligible delay costs under service degradation. On the contrary, we expect to see an increase in the abandonment probability if the caller has a high delay cost.

A simple plot of the probability density function of the waiting cost for the “no-priority” group reveals that the great majority of “no-priority” callers have negligible waiting costs. Hence, we expect their behavior to be similar to those callers in the “low-priority” group (see the first observation made above). Indeed, comparing the average waiting time of the “no-priority” callers for the two models (with the endogenous versus exogenous abandonment distribution) under the static priority rule verifies this intuition (160.61 sec versus 183.69 sec).[11]

In addition, comparing the two models under the threshold policy, we expect the abandonment probability to be higher for the model with the endogenous abandonment distribution because the threshold policy improves the service for the “no-priority” group (relative to the current policy). The threshold policy will ensure that the queue lengths for the “no-priority” group in both models will be close to the threshold, and hence, close to each other. Combining these two suggests the overall abandonment rate will be determined by the (per period) probability of abandoning, which is higher in the model with the endogenous abandonment distribution. Comparing the results for the two models under the threshold policies verifies this intuition; see Table 10 (20.70% versus 24.97% for th=15 and 8.28% versus 11.10% for th=5).[12]

Fourth, the effect of forward looking callers is more prominent in waiting time estimates than abandonment rate estimates. Consider the two forces which contribute to the overall abandonment rate: queue length and the probability of abandoning in a period. For the “low-priority” group, our endogenous model suggests longer queues and lower probability of abandoning whereas the model with exogenous abandonment distribution has shorter queues and higher probability of abandoning. However, the simulation results show that the abandonment rates (which can be approximated by the product of the two) are comparable. This suggests that the waiting time estimates are likely to be off significantly if one ignores the endogenous caller behavior, but the difference in the abandonment rate estimates will be smaller. Nonetheless, when the threshold is 15, the abandonment rate of the “low-priority” group is higher under the exogenous abandonment distribution because the effect of the higher abandonment probability dominates.

In many contexts such as making outsourcing decisions, designing service level agreements and service contracting, the ex-ante performance analysis of the call center by simulation is essential. Our model highlights the importance of modeling callers’ behavior endogenously. Namely, we observe that using a model with exogenously given abandonment distributions may lead to waiting

[11] Under the static priority policy for the average waiting time of the “no-priority” callers, the difference between the results of the exogenous and the endogenous models is significant with 90% confidence (t-statistic=3.99).

[12] Under the threshold policy with 15 and 5 as the threshold for the abandonment rate of the “no-priority” callers, the difference between the results of the exogenous and the endogenous models is significant with 90% confidence (t-statistics=16.86 (th=15) and t-statistics=14.56 (th=5)).

time estimates which can be off significantly. This would be problematic in a setting like call center outsourcing where service level measures on waiting time distributions are used. The estimates of the abandonment rate are less problematic due to the counteracting forces of queue length and the probability of abandoning as explained above. Our modeling approach offers another potential advantage, which is its ability to estimate what happens when extrapolation is needed. Consider a promotional campaign that increases the number of high priority calls significantly. Simulating the system performance in this case may require understanding callers' abandonment patterns in a regime where waiting times are longer than those observed in the data. This is challenging to do non-parametrically whereas our approach can be helpful in studying such situations.

7 Concluding Remarks

This paper studies the patience of callers in call center queues. Callers' valuation for the service obtained and their cost for waiting on hold are empirically estimated from call center data, making use of a structural estimation approach. The valuation for the service, the waiting cost, as well as a random shock that represents external events during the wait in queue, determine a caller's utility over time. Each caller makes wait or abandon decisions based on maximizing this utility. Using actual abandonment decisions in the data, the approach estimates callers' parameters regarding service valuation and waiting costs. The estimation results demonstrate how observed abandonment behavior can be explained with rational agents having linear utility functions, heterogeneous taste parameters, and experiencing idiosyncratic random shocks as they wait.

Understanding customer patience behavior is essential in call center management. The individual level decision modeling approach we take herein allows us to draw a natural bridge between observed behavior (in real data) and subsequent modeling of strategic customers in queues. The estimation can be used within models that explore the management of informational or delay announcements, dynamic routing or priority type choices, and more generally as part of a call center's overall customer relationship efforts.

To illustrate this, the estimation results are used to study the role endogenous abandonment behavior modeling plays in call center performance analysis. A comparison is made between the proposed model with endogenous abandonment behavior and one where the abandonment distribution is exogenously determined from the data, as typically done in the literature. In a series of experiments that contrast the performance under the service discipline in place at the call center, with several different alternatives, it is shown that the two models can lead to significantly different results in terms of waiting time performance. These examples highlight the importance of modeling callers as strategic agents for managerial decisions that are based on caller waiting times (like delay announcements, or service level agreements in outsourcing).

A growing literature in Operations Management deals with models where customers are modeled to be strategic decision makers, cf. Hassin and Haviv (2003). Empirical analyses for such models is mostly lacking in the operations management literature. Our paper illustrates how customer preference parameters can be estimated for such models making use of structural estimation. While we focus on the estimation of a linear utility model in a queuing wait situation, the technique is not restricted to our specific model or setting.

Our analysis points to several future research directions worth exploring. In our estimation, the equilibrium service probability, $\pi(t)$, is estimated directly from the data. While this is a reasonable approach for our estimation, in a call center with delay announcements the equilibrium service probabilities that take into account caller reactions need to be recomputed for counterfactual studies. Also, our model assumes that callers make decisions at discrete time periods. We analyzed the effect of the length of these periods in our estimation, however the question of what decision period length is the most appropriate for a given setting remains to be answered. This is a topic for experimental investigation which is beyond the scope of our analysis. In our model, we assume that callers' waiting cost has a linear form, and that the reward and cost parameters are independent. We also assume that the idiosyncratic shocks have type-I extreme value distribution. It would be worth examining these assumptions in future research.

Acknowledgements

The authors thank the Service Enterprise Engineering (SEE) lab at the Technion (http://ie.technion .ac.il/Labs/Serveng/) for generously providing us with the data and ongoing technical support regarding the data in preparation of this manuscript. We are especially indebted to Prof. Avi Mandelbaum, Dr. Valery Trofimov, Mr. Igor Gavako, and Mrs. Ella Nadjharov who not only provided us with a clean data set, but also helped us understand the intricacies of the data and with further structuring of the data. We have had numerous exchanges with them during the preparation of this manuscript and are indebted to them for their help.

References

Aksin, Z., M. Armony, and V. Mehrotra (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management 16*(6), 665–668.

Armony, M., N. Shimkin, and W. Whitt (2009). The impact of delay announcements in many-server queues with abandonment. *Operations Research 57*(1), 66–81.

Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Process. 18*, 349–369.

Ben-Akiva, M. and S. Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand.* MIT Press, Cambridge, MA.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*, 841–890.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distirbutions. *Bull. Calcutta Math. Soc. 35*, 99–109.

Bitran, G. R., J.-C. Ferrer, and P. Rocha e Oliviera (2008). Managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing & Service Operations Management 10*, 61–83.

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

Byrd, R. H., J. Nocedal, and R. A. Waltz (2006). Knitro: An integrated package for nonlinear optimization. In G. di Pillo and M. Roma (Eds.), *Large-Scale Nonlinear Optimization*, pp. 35–39. Springer-Verlag.

Chebat, J. C., C. Gelinas-Chebat, and P. Filiatrault (1993). Interactive effects of musical and visual cues on time perception: An application to waiting lines in banks. *Perceptual and Motor Skills 77*, 995–1020.

Dahlquist, A. and A. Bjorck (2008). *Numerical Methods in Scientific Computing.* SIAM.

Gans, N., G. Koole, and A. Mandelbaum (2003). Telephone call centers: Tutorial, review and research prospects. *Manufacturing & Service Operations Management 5*, 73–141.

Guo, P. and P. Zipkin (2007). Analysis and comparison of queues with different levels of delay information. *Management Science 53*, 962–970.

Hassin, R. and M. Haviv (1995). Equilibrium strategies for queues with impatient customers. *Operation Research Letters 17(1)*, 41–45.

Hassin, R. and M. Haviv (2003). *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems.* Kluwer Academic Publishers.

Hornik, J. (1984). Subjective vs. objective time measures: A note on the perception of time in consumer behavior. *Journal of Consumer Research 11*, 615–618.

Horowitz, J. L. (2001). The bootstrap. In J. J. Hackman and E. Leamer (Eds.), *Handbook of Econometrics, Vol. 5*, pp. 3159–3228. Amsterdam: Elsevier Science.

Jouini, O., Y. Dallery, and Z. Aksin (2011). Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management 13*, 534–548.

Judd, K. (1998). *Numerical Methods in Economics.* Cambridge, Mass: MIT Press.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Comm. Technology 15*, 52–60.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association 53*, 457–481.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI 1995).

Leclerc, F., B. Schmitt, and L. Dube (1995). Waiting time and decision making: Is time like money? *Journal of Consumer Research 22*, 110–119.

Mandelbaum, A. and N. Shimkin (2000). A model for rational abandonments from invisible queues. *Queueing Systems: Theory and Applications 36*, 141–173.

Mendelson, H. (1985). Pricing computer services: Queueing effects. *Communications of the ACM 28*, 312–321.

Munichor, N. and A. Rafaeli (2007). Numbers or apologies? customer reactions to tele-waiting time llers. *Journal of Applied Psychology 92*(2), 511–518.

Nair, H. (2007). Intertemporal price discrimination with forward-looking consumers: Application to the us market for console video-games. *Quantitative Marketing and Economics 5(3)*, 239–292.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica 37*, 15–24.

Rust, J. (1987). Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica 55(5)*, 999–1033.

Shimkin, N. and A. Mandelbaum (2004). Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems: Theory and Applications 47(1-2)*, 117–146.

Snedecor, G. W. and W. G. Cochran (1989). *Statistical Methods.* Iowa State University Press.

Su, C.-L. and K. Judd (2012). Constrained optimization approaches to estimation of structural models. *Econometrica 80*(5), 2231–2267.

Suck, R. and H. Holling (1997). Stress caused by waiting: A theoretical evaluation of a mathematical model. *Journal of Mathematical Psychology 41*, 280–286.

Whitt, W. (1999). Improving service by informing customers about anticipated delays. *Management Science 45*, 192–207.

Zakay, D. (1989). An integrated model of time estimation. In I. Levin and D. Zakay (Eds.), *Time and human Cognition: A life span perspective.* Amsterdam: North Holland.

Zohar, E., A. Mandelbaum, and N. Shimkin (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science 48*, 566–583.

A Proofs

Proof of Proposition 1. We first derive the formula for choice probabilities $P_{it}(d_{it}; r_i, c_i)$, and then the recursive formula for the integrated value function $V(t, r_i, c_i)$.

Recall that caller i takes action d_{it} if the utility of choosing d_{it} is higher than the utility of taking the reverse action, $1 - d_{it}$, that is

$$\begin{aligned} u(t, r_i, c_i, \varepsilon_{it}(d_{it}), d_{it}) &= v(t, r_i, c_i, d_{it}) + \varepsilon_{it}\left(d_{it}\right) \\ &> v(t, r_i, c_i, 1 - d_{it}) + \varepsilon_{it}\left(1 - d_{it}\right) \\ &= u(t, r_i, c_i, \varepsilon_{it}(1 - d_{it}), 1 - d_{it}). \end{aligned}$$

Therefore, we have

$$\begin{aligned} P_{it}(d_{it}; r_i, c_i) = &\int\int \mathbb{I}_{\{\varepsilon_{it}(d_{it}) - \varepsilon_{it}(1-d_{it}) > v(t, r_i, c_i, 1-d_{it}) - v(t, r_i, c_i, d_{it})\}} \\ &\times g(\varepsilon_{it}(0)) g(\varepsilon_{it}(1)) d\varepsilon_{it}(0) d\varepsilon_{it}(1). \end{aligned} \tag{18}$$

We assume that the idiosyncratic shocks have i.i.d type-I extreme value distribution with scale parameter 1 and location parameter $\beta \in \mathbb{R}$ with the probability density function $\exp(-(\varepsilon(d) - \beta)) \exp(-\exp(-(\varepsilon(d) - \beta))$ for $d = 0, 1$. As will be seen below, for technical convenience we will set $\beta = -\gamma$, where γ is Euler's constant. From (18), by Section 5.2 in Ben-Akiva and Lerman (1985), and the fact that $v(t, r_i, c_i, 1) = 0$, we obtain the formula for the choice probability as follows

$$\begin{aligned} P_{it}(d_{it}; r_i, c_i) &= \frac{\exp\bigl(v(t, r_i, c_i, d_{it})\bigr)}{\exp\bigl(v(t, r_i, c_i, 1)\bigr) \ + \exp\bigl(v(t, r_i, c_i, 0)\bigr)} \\ &= \frac{\exp\bigl(v(t, r_i, c_i, d_{it})\bigr)}{1 \ + \exp\bigl(v(t, r_i, c_i, 0)\bigr)}. \end{aligned} \tag{19}$$

What remains is to derive the recursive formula for the integrated value function. Recall from (5) that the integrated value function is given by

$$V(t, r_i, c_i) = \mathbb{E}\left[\max_{d \in \{0,1\}} u(t+1, r_i, c_i, \varepsilon_{i(t+1)}(d), d)\right],$$

where the expectation is taken over the distribution of $\varepsilon_{i(t+1)}(1)$ and $\varepsilon_{i(t+1)}(0)$. By Section 5.2 in Ben-Akiva and Lerman (1985), $\max_{d \in \{0,1\}} u(t+1, r_i, c_i, \varepsilon_{i(t+1)}(d), d)$ has type-I extreme value distribution with scale parameter 1 and location parameter $\beta + \log(e^{v_1} + e^{v_0})$, where $v_k = v(t + 1, r_i, c_i, k)$, $k = 0,\ 1$. Therefore, we have

$$\begin{aligned} V(t, r_i, c_i) &= \mathbb{E}\left[\max_{d \in \{0,1\}} u(t+1, r_i, c_i, \varepsilon_{i(t+1)}(d), d)\right] \\ &= \beta + \log(e^{v_1} + e^{v_0}) + \gamma. \end{aligned} \tag{20}$$

For technical convenience, we assume that the location parameter for the distribution of the idiosyncratic shocks β is equal to $-\gamma$. Then, by definitions of v_1 and v_0 and (20), it follows that

$$V(t, r_i, c_i) = \log\Big(\exp\big(v(t+1, r_i, c_i, 1)\big) + \exp\big(v(t+1, r_i, c_i, 0)\big)\Big). \tag{21}$$

Substituting the values of the nominal utilities into (21), the integrated value function can be written as follows:

$$V\left(t, r_i, c_i\right) = \log\Big(1 + \exp(-c_i + \pi\left(t+1\right) r_i + (1 - \pi\left(t+1\right))V\left(t+1, r_i, c_i\right))\Big), \tag{22}$$

which provides the recursive formula for the integrated value function. To conclude the proof, note that for period $T-1$ the integrated value function is given by

$$V(T-1, r_i, c_i) = \log\Big(1 + \exp(-c_i + \pi(T-1+1)r_i + (1 - \pi(T-1+1))V(T, r_i, c_i))\Big). \tag{23}$$

Since $\pi(T) = 1$, from (22)-(23), the integrated value functions in period $T-1$ and consequently all earlier periods do not depend on $V(T, r_i, c_i)$, and for convenience we assume that $V(T, r_i, c_i) = 0$ for all i. □

B Kaplan-Meier Estimator

Since some callers abandon the queue, and we can not observe the actual waiting times of all callers, the data is censored. Therefore, we use the Kaplan-Meier estimator to estimate the cumulative distribution of callers' waiting times, which is denoted by $F(t)$.

Recall that N denotes the number of callers in the data. Suppose that $t_1 < t_2 < ... < t_m$ are the ordered waiting times of the callers who receive service, where m is the number of distinct waiting times. Note that $m \leq N$ because some callers may receive service at the same time.

Suppose that n_j callers have not received service or abandoned the queue just prior to $t_j,\ j \in \{1, ..., m\}$. In addition, denote by δ_j the number of callers who receive service at t_j. The conditional probability that a caller receives service after t_j given that the caller has not received service before t_j is given by $q_j = 1 - \delta_j / n_j$. Denote by $S(t)$ the probability that a caller's waiting time exceeds t. The Kaplan-Meier estimation of $S(t)$ for $t \in [t_k, t_{k+1})$ is given by $\hat{S}(t) = \prod_{j=1}^{k} q_j$. Denote by $\hat{F}(t)$ the Kaplan-Meier estimation for $F(t)$. Then, $\hat{F}(t) = 1 - \hat{S}(t)$.

C Gauss-Hermite Integration

To calculate the likelihood function, we need to integrate the products of choice probabilities with respect to y_{1i} and y_{2i}. We use Gauss-Hermite integration to approximate the two dimensional integrations.

Let ω_k and x_k denote the weights and the nodes of the Gauss-Hermite quadrature, respectively. Then, by Equation 7.2.10 in Judd (1998), the expectation of a function $f(y)$ where y is distributed according to $N(\mu, \sigma^2)$ is approximated as follows

$$\begin{aligned}\mathbb{E}(f(y)) &= (2\pi\sigma^2)^{-\frac{1}{2}} \int f(y) \exp(\frac{-(y-\mu)^2}{2\sigma^2}) \\ &= \frac{1}{\sqrt{\pi}} \sum_{k=1}^{l} \omega_k f(\sqrt{2}\sigma x_k + \mu), \end{aligned} \tag{24}$$

where l is the number of nodes used for the approximation. According to Section 7.2 in Judd (1998), the approximation given in (24) is exact for all polynomials with degrees less than or equal to $2l-1$.

Suppose that $f(r_i, c_i)$ is an arbitrary function such that r_i and c_i are given by (1). Then using the Gauss-Hermite integration in (24) and the repeated one dimensional integration given in Equation 7.5.1 in Judd (1998), we have

$$\begin{aligned} & \int\int f(r_i, c_i)\phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i} \\ = & \int\int f(\exp(m_r + \sigma_r y_{1i}), \exp(m_c + \sigma_c y_{2i}))\phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i} \\ = & \frac{1}{\pi} \sum_{k=1}^{l} \sum_{j=1}^{l} \omega_i\omega_j f(\exp(m_r + \sigma_r\sqrt{2}x_k), \exp(m_c + \sigma_c\sqrt{2}x_j)), \end{aligned} \tag{25}$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution. By Theorem 5.4.1 in Dahlquist and Bjorck (2008), the approximation of the integration given in (25) is exact for all $f(r_i, c_i)$ where $f(r_i, c_i) = \sum r_i^m c_i^n,\ m, n \leq 2l-1$ and $m, n \in \mathbb{N}$.

Using the approximation given in (25), the likelihood of the entire sample is given as follows:

$$L = \frac{1}{\pi} \prod_{i=1}^{N} \sum_{k=1}^{l} \sum_{j=1}^{l} \omega_k\omega_j \times \prod_{t=0}^{\tau_i} P_{it}(d_{it}; \exp(m_r + \sigma_r\sqrt{2}x_k), \exp(m_c + \sigma_c\sqrt{2}x_j)).$$

We consider 5 nodes for the approximation, i.e. $l = 5$. The nodes and weights of the Gauss-Hermite quadrature are given in Table 11, cf. Table 7.4 in Judd (1998).

k	x_k	ω_k
1	–2.0202	0.0199
2	–0.9586	0.3936
3	0.0000	0.9453
4	0.9586	0.3936
5	2.0202	0.0199

Table 11: The nodes and weights of the Gauss-Hermite quadrature for $l = 5$.

D Monte-Carlo Experiments

To test the capability of the proposed estimation method to identify the true parameters of the callers, we use Monte-Carlo experiments. To do so, we first generate simulated data sets assuming certain values for the structural parameters. We denote these values by true values. Then, we estimate the parameters of the simulated data sets and construct the 95% confidence intervals, and check if the true values are in the corresponding confidence intervals.

To implement the Monte-Carlo experiment, we consider the following true values for the structural parameters: $m_r = 1.8,\ m_c = -3,\ \sigma_r = 0.2$ and $\sigma_c = 1$. We set the maximum waiting time of the callers T to 120 periods. In addition, for the waiting time distribution and probability of receiving service, $F(t)$ and $\pi(t)$, we use those from the data (suitably truncated), which are estimated using the Kaplan Meier estimator (see Appendix B). We generate 40 simulated data sets such that each data set contains 100,000 callers.

To simulate the abandonment behavior of the callers, for each caller, we draw y_1 and y_2 from the standard normal distribution. Then, we find r and c of the callers making use of the assumed true values of the structural parameters, and consequently can calculate the integrated value function and the nominal utilities of the callers. Next, we add i.i.d. type-I extreme value distributed random shocks to the nominal utilities to find the utilities of waiting and abandoning.

Having the probability of receiving service $\pi(t)$ and the utilities of waiting and abandoning, we can decide if the simulated caller receives service, abandons the queue or continues to wait as follows:

1- Draw a random variable x from the uniform distribution between 0 and 1. If $x \leq \pi(t)$, the caller receives service and we end the procedure.

2- If $x > \pi(t)$, compare the utilities of waiting and abandoning. If utility of abandoning is larger, the caller abandons the queue and we end the procedure, if not, the caller continues to wait and we repeat steps 1 and 2 for the next period.

Table 12 shows the mean, standard deviation, upper and lower bounds of the 95% confidence intervals for the estimated parameters of the simulated data sets. These results as well as a series of extensive Monte-Carlo experiments (available from the authors) show that our estimation method can recover the true parameter values from the data.

Structural parameter	m_r	σ_r	m_c	σ_c
True value	1.80	0.20	–3.00	1.00
Mean (Simulated data)	1.79	0.18	–3.17	1.11
Standard deviation (Simulated data)	0.02	0.04	0.09	0.06
Upper bound of the 95% confidence interval	1.82	0.26	–2.98	1.24
Lower bound of the 95% confidence interval	1.76	0.10	–3.35	0.98

Table 12: Results of the Monte-Carlo experiment.