Impact of Delay Announcements in Call Centers: An Empirical Approach

Zeynep Aksin
College of Administrative Sciences and Economics, Koc University, 34450 Sariyer-Istanbul, Turkey,
zaksin@ku.edu.tr
Baris Ata
Booth School of Business, University of Chicago, Chicago, Illinois 60637,
baris.ata@chicagobooth.edu
Seyed Morteza Emadi
Kenan-Flagler Business School, University of North Carolina at Chapel Hill, North Carolina 27599,
seyed_emadi@kenan-flagler.unc.edu
Che-Lin Su[1]
Booth School of Business, University of Chicago, Chicago, Illinois 60637,
che-lin.su@chicagobooth.edu

Abstract

We undertake an empirical study of the impact of delay announcements on callers' abandonment behavior and the performance of a call center with two priority classes. A Cox regression analysis reveals that in this call center, callers' abandonment behavior is affected by the announcement messages heard. To account for this, we formulate a structural estimation model of callers' (endogenous) abandonment decisions. In this model, callers are forward-looking utility maximizers and make their abandonment decisions by solving an optimal stopping problem. Each caller receives a reward from service and incurs a linear cost of waiting. The reward and per-period waiting cost constitute the structural parameters that we estimate from the data of callers' abandonment decisions as well as the announcement messages heard. The call center performance is modeled by a Markovian approximation. The main methodological contribution is the definition of an equilibrium in steady state as one where callers' expectation of their waiting time, which affects their (rational) abandonment behavior, matches their actual waiting time in the call center, and its characterization as the solution of a set of non-linear equations. A counterfactual analysis shows that callers react to longer delay announcements by abandoning earlier, that less patient callers as characterized by their reward and cost parameters react more to delay announcements, and that congestion in the call center at the time of the call affects caller reactions to delay announcements.

Keywords: Delay announcement, Abandonment, Structural estimation, Equilibrium

1 Introduction

Delay announcements provide estimates of the waiting time to prospective customers in invisible queues, inform callers about their chances of receiving service and influence their decision to keep waiting or to abandon. In turn, callers' abandonment behavior affects system performance. Thus,

[1]Che-Lin Su passed away in July 2015. We lost a wonderful friend and the field lost one of its pioneers.

understanding the impact of delay announcements on the system performance is an integral part of designing modern call centers. In this context, we study the following questions: What is the impact of delay announcements on callers' behavior in terms of waiting and abandoning decisions? Do different caller types react differently to delay announcements? Does the same caller react to announcements differently when facing different levels of congestion? How does callers' abandonment behavior affect the system performance if the call center manager announces delays? We undertake an empirical approach to investigate these questions and develop a methodological framework to characterize the equilibrium in steady-state of the system that facilitates our analysis.

We first empirically explore the question of whether delay announcements affect callers' abandonment behavior. A series of Cox regression analyses reveals that both the content and the sequence of announcement messages have a statistically significant impact on callers' abandonment behavior. Furthermore, caller characteristics and congestion levels in the call center matter. This analysis shows the impact of delay announcements observed in the data under the current policy of the call center; however it cannot be used to investigate the impact on the system under different announcement messages or under different operating conditions. The reason for this is that as announcement messages, caller composition or operating characteristics change, the abandonment behavior will change. Empirically demonstrating the role of announcements on caller abandonment behavior in call centers constitutes the first important contribution of this paper.

To investigate the impact of delay announcements on the system (even for announcements different from what we observe in the data), we model callers' abandonment decisions endogenously as in the optimal stopping model introduced in Aksin et al. (2013). The main difference in the estimation, for the current paper's setting with delay announcements, is that callers' hazard rates of entering service, i.e. service probabilities, are a function of delay announcement histories. Since callers' choices when facing delay announcements are explicitly modeled, this allows the possibility to explore abandonment behavior under different announcement schemes.

Much of the extant literature models callers' patience as exogenous to the call center operations and takes its parameters as given. In contrast, we model callers' patience endogenously and estimate its primitives (i.e. structural parameters) using data from a call center with two priority classes: high and low. In our data set, callers receive delay announcements that contain information about their positions in the queue and the waiting time of the longest waiting caller every 60 seconds.

Given the callers' parameters, we then predict the abandonment behavior of the high and low priority callers in settings where the pattern or the information content of the announcements differ from those in the data. We first use the optimal stopping model of callers' abandonment behavior to derive callers' abandonment time distributions from their anticipated waiting time distributions. We develop the optimal stopping model (Section 4) for a case that callers receive a delay announcement message every one minute. For simplicity, in the subsequent analyses (Sections 5 and 6) we assume that callers receive only one announcement message upfront upon arrival. Building on Whitt (2005), we then approximate the system by a Markovian model with state dependent abandonment rates. Using this Markovian approximation, we represent the callers' waiting time

distributions as functions of their abandonment time distributions.

We combine these components to define the equilibrium in steady-state of the system as one in which callers' waiting time distributions used in the derivation of the abandonment time distributions (based on the optimal stopping model) match the waiting time distributions derived from the Markovian approximation. This definition corresponds to the rational expectation equilibrium in the system, where callers' expectation about their waiting time, which affects their (rational) abandonment behavior, matches their actual waiting time in the call center. This is where the main methodological contribution of our paper lies: the definition, characterization and computation of the complex equilibrium in steady-state.

To highlight the contributions of this paper further, it is helpful to compare it with its antecedent Aksin et al. (2013) that introduced the optimal stopping model for studying callers' abandonment behavior under no delay announcements. From an application perspective, the main difference of this paper from Aksin et al. (2013) is the presence of delay announcements and corresponding generalizations of the model and the estimation framework. More importantly, from a methodological perspective, this paper makes a novel contribution by characterizing the system equilibrium in steady-state analytically. This characterization is enabled by the proposed Markovian approximation of the underlying queueing system, and allows us to circumvent the iterative simulation approach taken in Aksin et al. (2013) to find the new equilibrium in steady-state of the system subsequent to a change in the system policy. In the proposed framework, the equilibrium is obtained by solving a set of non-linear equations. An iterative simulation based approach may not be appropriate in a setting with delay announcements due to the additional layer of complexity the announcement messages bring: callers' abandonment behavior will depend on both the announcement message and the system operational state (as shown by the Cox regression results), making the resulting search a multi-dimensional one where the approach of Aksin et al. (2013) may not converge. In what follows, we use "equilibrium" and "equilibrium in steady-state" interchangeably.

Finally, the practical contribution of our paper comes from a series of counterfactual analyses, where we explore the effect of different delay announcements on the system performance. In particular, we focus on the call center manager's choice of the granularity of the information contained in the announcements. For example, the manager can inform the callers about the exact number of callers who are waiting to be served (full information on system occupancy case).[2] Alternatively, the manager may choose to announce a range for the number of callers in the queue (partial information on system occupancy case). In addition, we explore the impact of delay announcements under different priority policies. In our counterfactual study, callers who hear that the queue is long (short) abandon more (less) and leave the system sooner (later) compared to the case when no delay information is provided. Moreover, increasing the granularity of the information contained in the announcements leads to a smoother change in callers' behavior across different system states. The results also show that less patient callers are more sensitive to delay information, and as a result their abandonment behavior changes more significantly. Thus the counterfactual results confirm

[2] Our full information on system occupancy is similar to the partial information case in Guo and Zipkin (2007).

that abandonment behavior changes as a function of customer characteristics, the messages heard, and the operating conditions in the call center, underlining the importance of an approach that can take all of these factors into account when analyzing delay announcements.

The rest of the paper is organized as follows: We provide an overview of the related literature in Section 2. Section 3 describes the data and a Cox regression analysis to illustrate the impact of delay announcements on callers' abandonment behavior, the details of which are given in Appendix A. Section 4 presents a model for callers' abandonment behavior, the estimation framework and results. Section 5 lays out a methodological framework to study the impact of delay announcement in call centers. In Section 6, we discuss the counterfactual analysis. Section 7 concludes the paper. Proofs are provided in Appendix B. A simulation study to illustrate the identification of our model is described in Appendix C. Appendix D demonstrates a method to improve the state space collapse approximation. A comparison of the optimal stopping model for callers' abandonment behavior and a simpler model is given in Appendix E, and finally, Appendix F shows a validation of the equilibrium computation via simulation.

2 Literature Review

A natural consequence of customers' dissatisfaction with waiting is that some customers may lose patience and abandon the queue. The traditional approach to model customers' abandonments considers an exogenous distribution for customers' abandonment time (patience time). If customers' actual waiting time exceeds their abandonment time, they abandon the system; see Gans et al. (2003) and references therein, and also Ward (2012) for an overview of the literature on asymptotic analysis of queueing systems with abandonments.

An alternative approach is to model callers as forward looking decision makers who make wait or quit decisions to maximize their utility (see Hassin and Haviv (2003) and references therein). In Hassin and Haviv (1995), such callers abandon the system if their reward from receiving service drops to zero. Mandelbaum and Shimkin (2000) and Shimkin and Mandelbaum (2004) assume that upon arrival callers choose an optimal abandonment time to maximize their utility. Callers abandon the system if their actual waiting time exceeds their optimal abandonment time. Aksin et al. (2013) and this paper are outgrowths of the Ph.D. dissertation Emadi (2013). Emadi (2013) first formulates the endogenous abandonment model in Aksin et al. (2013) and uses it to study delay announcements, as presented in the current paper. Aksin et al. (2013) model callers' endogenous abandonment behavior as an optimal stopping problem, in which callers make the decision between waiting and abandoning not only upon arrival but also during the waiting encounter. This feature resembles the optimal stopping model for callers' decisions studied herein. However, unlike Aksin et al. (2013), we incorporate the impact of delay announcements on callers' decision. Under a rational expectations framework, we derive a set of nonlinear equations to characterize the equilibrium system performance in the presence of delay announcements. Afèche and Sarhangian (2015) provide an analytical characterization of equilibrium abandonment behavior in an $M/M/1$ observable

queue with two customer classes, in which no delay announcement is needed as customers observe the state of the system.

The current paper is among the first in operations management combining the empirical estimation of choice related parameters of forward looking customers, and the subsequent analysis of an operational problem in the presence of these customers. In our setting the estimation involves estimating the parameters of a caller's utility function, while the operational problem studied is that of predicting the effect of delay announcements in a queue. Yu et al. (2015) also study delay announcements empirically, subsequent to an estimation of caller characteristics. In order to show the effect of announcements on abandonment behavior in the data, Yu et al. (2015) first perform pairwise comparisons of survival curves under different announcement messages. These comparisons do not lead to definitive results regarding the impact of announcements. Our analysis, which makes use of a Cox regression, shows that one cannot look at the effect of an announcement message on abandonment behavior in isolation, since this effect is shown to depend on the content and sequence of messages heard as well as the operating environment at the time of the call. Their subsequent analysis parallels Aksin et al. (2013), making use of an iterative simulation based computational approach. The main difference of their structural estimation analysis is that they consider models that allow for changes in waiting cost parameters as a function of delay announcement. In our approach, the reward from service and the waiting cost are customer primitives, which are not affected by the environment. Lu et al. (2013) analyze how waiting in queue in the context of a retail store affects customers' purchasing behavior. They estimate the impact of customer service levels on purchase incidence and choice decisions from a data set. Vulcano et al. (2010) estimate customers' choice related parameters empirically and then study the subsequent revenue management problem under choice based models. In pricing problems, such an approach of empirical estimation of consumer parameters to predict profit and welfare performance is more common. For some recent examples, see Nair (2007), Hendel and Nevo (2013) and Li et al. (2014).

The study of delay information and its effect in queues has been ongoing for some time. Feigin (2006) provides a descriptive analysis of data from a bank call center, illustrating that aggregate hazard rates change at announcement points. Such changes in behavior affect the system performance, which can be studied through queueing analysis. The role of revealing or suppressing queue length information on customer choices in joining queues is first analyzed by Hassin (1986). One of the earlier models to study the impact of waiting time information or delay announcement in queues is by Whitt (1999a). Whitt analyzes the effect of providing state information on the performance of a single class Markovian model, assuming callers' patience threshold has an exogenously given exponential distribution. The parameter of this distribution is not affected by the announcements. Jouini et al. (2009) provide a multi-class extension of Whitt's model. Guo and Zipkin (2007) show how different levels of delay information, e.g. providing system occupancy versus providing the exact delay, affect the performance of an $M/M/1$ system. Both Whitt (1999a) and Guo and Zipkin (2007) assume that customers either abandon immediately after hearing the delay announcements, or remain in the queue until they enter service. In other words, abandonments while waiting do not

arise in their models. This assumption is relaxed in Armony et al. (2009) and Jouini et al. (2011).

In Armony et al. (2009), callers may abandon after hearing the delay announcements. The authors assume that the probability of balking and the distribution of callers' patience thresholds are exogenously specified and depend on the delay estimates given to the callers. Jouini et al. (2011) also allow abandonments. Instead of specifying exogenous patience thresholds, the authors explicitly model callers' behavior changes after hearing the delay announcements. They also assume an exponential distribution for callers' initial patience thresholds. This distribution remains exponential after callers hear the announcements, but its parameter changes based on an underlying behavioral patience update model. They suggest an iterative fixed point algorithm to find the parameter of the patience threshold distribution in the equilibrium state. Similar to this latter paper, we endogenize callers' reactions to delay announcements; however, we place no restrictions on their form, and derive them from callers' estimated patience primitives within a structural estimation framework.

Since providing delay information affects callers' abandonments which in turn affects system performance, a natural question to investigate is what information to provide in the delay announcements. Whitt (1999a) considers announcing the state of the system and the remaining service time of each customer in the system. Consequently, the callers learn the distribution of their delay time upon arrival. In Jouini et al. (2011), the authors assume that the announcements are made as a chosen percentile of the delay distribution. In Armony et al. (2009), the authors study the effect of announcing one particular real time delay estimate (announcing the last customer's realized delay) on system performance. The performance of real time delay estimators in different settings are analyzed in Ibrahim and Whitt (2009). State dependent delay estimators are discussed in Whitt (1999b). Xu et al. (2007) study ticket queue systems, where each customer is issued a numbered ticket upon arrival, and the number currently being served is displayed. The information provided in ticket queue systems differs from the actual queue length due to customers' abandonments. Guo and Zipkin (2007) consider two types of announcements: partial information (announcing system occupancy) and full information (announcing the exact waiting time).

3 Data

In this section, we first describe the data set. Next, we present a Cox regression analysis to illustrate the impact of delay announcements on callers' abandonment behavior.

3.1 Data Description

Our data set is the operational data of a bank call center spanning all twelve months of 1999.[3] The call center processes up to 100,000-120,000 calls per month. We only focus on callers who seek to talk to an agent (35 percent of all callers) and who contacted the call center on weekdays. Around 70 percent of callers who talked to an agent asked for the retail banking service. We only consider these callers in our analysis. In order to focus on relatively busy hours of the day when the call

[3]The data set was made available to us by the Service Enterprise Engineering (SEE) lab at Technion-Israel Institute of Technology (http://ie.technion.ac.il/Labs/Serveng/).

volume is relatively stable, we focus on calls between 9 a.m. and 4 p.m. Before August, there was only one pool of agents. Starting from August a new pool was added to the call center. We consider observations before August in our analysis.

Customers in this call center have different priorities. There are two priority classes: high and low. The low priority callers join the end of the queue, while the high priority callers are advanced in the queue by 90 seconds. Service is then delivered on a first-come-first-served (FCFS) basis. Consequently, a newly arriving high priority caller will enter service sooner than all the low priority callers who have arrived in the last 90 seconds.

The data traces each call from initiation to termination. The calls can be broken down to three stages: VRU (Voice Response Unit) interaction, waiting in the queue and talking to an agent (service stage). The entry and exit times for each stage is recorded in the data. In addition, the identification number of the callers and the name of the agent who served the caller are observed in the data. The observable entries for each call are listed in Table 1.

Customer ID and ID of the agent who served the call
Priority class and type of service requested
Date and time of entering and exiting the VRU, queue and the service stage
Outcome of the call (entered service or abandoned?)

Table 1: The observable entries in the data.

The callers receive a delay announcement upon arrival and every 60 seconds after arrival. The announcements inform callers about their relative positions in the queue, accompanied by the waiting time of the longest waiting caller. The relative position of the callers is calculated based on the number of available agents. For example, if there are s agents available, the callers in positions 1 to s from the head of the queue are told they are "First" in the queue; callers in positions $s+1$ to $2s$ are told they are "Second" in the queue, etc.

We cannot observe the announcements given to the callers in the data; however, we recover them using the observable entries in Table 1 and the rule for the announcements. In particular, given callers' priority classes and their arrival times to the queue, we find callers' positions in the queue and the amount of time they have been waiting. In addition, to find the number of available agents, we divide the day into 15 minute intervals. An agent is considered available in an interval if she serves at least one caller in that interval. Using the number of available agents, callers' positions and their waiting times at any particular time during the day, we derive the type of announcements that were given to the callers.

After deriving the announcements given to the callers, we observe that the first part of the announcement (the relative position in the queue) is between 1 and 5. We consider callers whose relative position in the queue does not exceed 3 for our analysis. This portion of the data constitutes more than 99.7 percent of the total callers. For this portion of the data, the second part of the announcement (the waiting time of the longest waiting caller) is between 0 and 2700 seconds. In Section 4.2, we observe that the number of announcements contributes directly to the complexity of the maximum likelihood estimation problem. Taking the complexity of the estimation problem into account, we discretize the values for the second part of the announcement to the following six

intervals (in seconds): [0,10], [11,30], [31,90], [91,210], [211,480] and [481,2700]. These intervals are chosen so as to ensure that callers corresponding to each interval have similar waiting times and abandonment patterns.[4] Given these simplifications, callers may hear three different messages in the first part of the announcement (the relative position in the queue) and six different messages in the second part of the announcement (the waiting time of the longest waiting caller). Therefore, the total number of announcement messages J is 18.

Callers who enter service immediately after arrival do not make any abandonment decisions and, consequently, are not considered in our analysis. Moreover, we focus on callers with a wait duration less than 600 seconds, who constitute more than 99.5 percent of callers who had to wait.

In summary, our analysis focuses on 62,718 calls with the following characteristics: the caller asked for the retail banking service, the caller did not enter service immediately after arrival, the call was received on weekdays before August between 9 a.m. and 4 p.m., the caller waited less than 600 seconds and the caller's relative position did not exceed 3. The summary statistics for this portion of the data are given in Table 2. The abandonment rates in this call center are unusually high. This can be attributed to the very small size of this center with around five agents, preventing any statistical economies of scale from taking place.

Priority class	Number of observations	Abandonment rate	Average waiting time (sec.)	Average waiting time of abandoned calls (sec.)
High priority	41,401	17.86 %	92.36	62.07
Low priority	21,317	32.08 %	103.86	63.13

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

As shown in Table 2, the difference between the average waiting times of the high and low priority callers is 11.5 seconds even though the high priority callers are advanced by 90 seconds. In this call center 24% of customers enter service upon arrival, i.e. do not wait at all. Consequently, the queue does not get long and the advancement of the high priority callers does not have a large effect most of the time, which leads to a small difference between the average waiting times of the high and low priority callers.

3.2 Illustrating the Impact of Delay Announcements in the Data

Before proceeding with any further analysis regarding the structural estimation of callers' abandonment behavior under delay announcements, we empirically explore whether delay announcements have an impact on callers' abandonment behavior. This is not straightforward and care needs to be taken in the analysis. First of all, abandonment time data is right censored. The call center being studied makes multiple delay announcements timed at every sixty seconds. Patience is not only affected by the content of these multiple messages, but also by the sequence of messages being heard. In other words, the multiple delay announcements cannot be analyzed independently. These features imply the need for a survival analysis with multiple ordered events, where survival

[4]We partition the values for the second part of the announcement to a fine grid of subintervals, and find the abandonment rates and average waiting times of the callers corresponding to each subinterval. Then we merge the adjacent subintervals for which callers abandonment rates and average waiting times are relatively close. This leads to the six aforementioned intervals.

time represents time to abandonment and the multiple ordered events are the delay announcements being heard every sixty seconds as callers wait. Finally, as described earlier, the call center has two types of priority for callers, and like any call center, will experience different congestion levels as a function of call arrivals and staffing, as well as time of day effects. In order to clearly understand the effect of announcements on patience of callers, one needs to control for other confounding effects such as type of caller or call center congestion. To perform this analysis we resort to a Cox regression analysis where control variables for priority class, arrival rates, and time of day effects are included.

We order the announcement messages lexicographically based on the first part (the relative position in the queue) and the second part (the waiting time of the longest waiting caller) of the message; and label them with indices from 1 to 18. We then define the main independent variables of the analysis as the announcement message heard upon arrival, followed by the incremental value of the announcement message heard at each subsequent announcement event. To capture the incremental effect, all announcement variables following the one upon arrival are defined as the difference between two messages heard in subsequent announcements. Thus, if a caller hears an improving message (progress), the variable will be negative; whereas if the caller is informed of a deteriorating situation, this variable will be positive. The magnitude of the variable will represent the extent of the progress or deterioration.

The results of the analysis show that callers who hear a message with a higher index (representing a worse combination of position in the queue and wait of the longest waiting caller) abandon earlier. In addition, callers who hear announcement messages with increasing index values at multiples of 60 seconds (indicating a worsening condition) abandon earlier. To be more specific, callers who hear an announcement indicating a longer delay upon arrival become less patient and abandon earlier. In addition, callers who see a deteriorating delay condition abandon earlier. These results show that both the content of the announcement messages and their sequence plays a role in shaping callers' abandonment decisions. Furthermore, the significance of the control variables for caller priority and arrival rates (as a proxy for congestion) suggests that callers' patience primitives and the operations of the call center affect their abandonment decisions. The details of this analysis are given in the online Appendix A.

The Cox regression analysis shows that the announcement messages currently implemented in this call center (the independent variables in the regression) indeed impact callers' abandonment behavior (the dependent variable in the regression). However, we cannot use this analysis to find the impact of a different set of independent variables on callers' abandonment behavior. In other words, if we change the content or type of the announcements in the data, we cannot use the results of the Cox regression to find callers' abandonment behavior in the new setting, and consequently, cannot find the impact on call center performance measures. To overcome this issue, we proceed with a structural estimation approach explained in the remainder of the paper.

4 A Model for Callers' Abandonment Behavior Under Delay Announcements

In this section, we first present an optimal stopping model for callers' abandonment behavior under delay announcements. Next, we describe the estimation methodology and results. Our analysis in this section allows each caller to hear multiple delay announcements during her wait, consistent with our data. However, in Sections 5 and 6 we focus attention on the case with a single delay announcement for simplicity, where each caller hears the delay announcement upon arrival.

4.1 A Model of Callers' Abandonment Behavior

We model callers' decision making process as an optimal stopping problem, where abandoning corresponds to "stopping." Callers are forward looking. In each period, callers compare their expected future utilities from waiting and abandoning, and choose the action that maximizes their utility. If a caller decides to abandon, she leaves the system immediately. Otherwise, she remains in the system and faces a similar decision between abandoning and waiting in the next period unless she enters service.

To account for callers' preferences, we assume that callers' utilities depend on two parameters: the reward from receiving service and the delay cost per unit of time, denoted by r_i and c_i, respectively, for caller i. Callers may have different reward and cost parameters. To model caller heterogeneity, we assume that the reward and cost parameters have the following log-normal distributions:

$$r_i = \exp(m_r + \sigma_r y_{1i}) \text{ and } c_i = \exp(m_c + \sigma_c y_{2i}), \tag{1}$$

where y_{1i} and y_{2i} are i.i.d. standard normal random variables. Given (1), the set of structural parameters that characterize the distribution of callers' preferences is $\Theta = (m_r, \sigma_r, m_c, \sigma_c)$.

We assume that callers receive a delay announcement from a set of J possible announcements upon arrival and every L periods after that. In particular, if a caller has not entered service or abandoned until period kL, she receives her $(k+1)^{th}$ announcement at period kL. We denote the $(k+1)^{th}$ announcement, received in period kL, by $j_k \in \{1, ..., J\}$. In the periods with no delay announcement ($t \neq kL, k = 0, 1, 2, ...$) callers make their abandonment decision at the beginning of the period. In the periods with a delay announcement ($t = kL, k = 0, 1, 2, ...$), callers first receive the delay announcement and then make the abandonment decision. For simplicity, we assume that every delay announcement is conveyed instantaneously. Figure 1 shows the order of events.

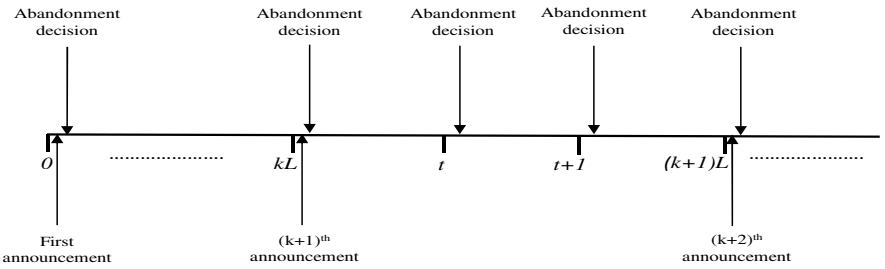

Figure 1: The order of events.

Let $\mathcal{H}_k = (j_0, j_1, ..., j_k)$ denote the sequence of announcements received by a caller who has received $(k+1)$ announcements. Suppose that all callers enter service before time mL.[5] Consequently, the maximum number of announcements a caller may receive is m. Denote by $a_{\mathcal{H}_k}$ the ex-ante probability that a caller receives the announcement history $\mathcal{H}_k$. Moreover, for $t \in \{kL, ..., (k+1)L-1\}$, let $\pi_{\mathcal{H}_k}(t)$ denote the service probability in period t, i.e. the probability of entering service conditional on not having entered service before t, for callers with announcement history $\mathcal{H}_k$. We assume that the probabilities $a_{\mathcal{H}_k}$ and $\pi_{\mathcal{H}_k}(\cdot)$ are common knowledge among callers for all $\mathcal{H}_k \in \{1, ..., J\}^{k+1},\ 0 \leq k < m$. The announcement history $\mathcal{H}_k$ is terminal if no caller with such a delay history hears a further delay announcement. Namely, all such callers either abandon or enter service before the next scheduled announcement. For each such terminal $\mathcal{H}_k$, define $T_{\mathcal{H}_k} \in \{kL, ..., (k+1)L-1\}$ as the maximum waiting time of all callers with that announcement history. Then, $T_{\mathcal{H}_k}$ will help us determine the last relevant period for those callers with announcement history $\mathcal{H}_k$, and facilitate a recursive definition of the callers' valuations of their future decisions, starting from period $T_{\mathcal{H}_k}$.

Next, we define the utility for caller i with announcement history $\mathcal{H}_k$ for $t \in \{kL, ..., (k+1)L-1\},\ 0 \leq k < m$, denoted by $u_{\mathcal{H}_k}$, as a function of the action d chosen by the caller. We let $d = 1$ if the caller abandons and $d = 0$ if she decides to wait. The utility function is given by

$$u_{\mathcal{H}_k}(t, r_i, c_i, \epsilon_{it}(d), d) = v_{\mathcal{H}_k}(t, r_i, c_i, d) + \epsilon_{it}(d), \tag{2}$$

where $\epsilon_{it}(d)$ denotes the random shock incurred by choosing action d. The random shocks can be attributed to external events that may shift a caller's utility and increase her willingness to either abandon or wait. The term $v_{\mathcal{H}_k}(t, r_i, c_i, d)$ in (2) is the nominal utility. If a caller decides to abandon, she will leave the system immediately. Since the waiting cost incurred in the past is sunk, the nominal utility of abandoning is zero, i.e.

$$v_{\mathcal{H}_k}(t, r_i, c_i, 1) = 0, \tag{3}$$

whereas the nominal utility from waiting at $t \in \{kL, ..., (k+1)L-1\}$ has the following form

$$v_{\mathcal{H}_k}(t, r_i, c_i, 0) = -c_i + \pi_{\mathcal{H}_k}(t) r_i + (1 - \pi_{\mathcal{H}_k}(t)) \mathbb{E}\Big[\max_{d \in \{0,1\}} u_{\mathcal{H}_k}(t+1, r_i, c_i, \epsilon_{i(t+1)}(d), d) \Big]. \tag{4}$$

The first term on the right hand side of (4) is the waiting cost, and the second term is the expected

[5] This subsumes the special case of a single announcement when L is sufficiently large so that $m = 1$.

utility of receiving service in the current period. The third term is the expected utility of not receiving service in the current period and making the optimal decision in the next period.

We refer to the expectation in (4) as the integrated value function and denote it by $V_{\mathcal{H}_k}(t, r_i, c_i)$. The expectation is calculated by integrating the maximum utility in the next period over the distribution of the random shocks. We can rewrite (4) as follows

$$v_{\mathcal{H}_k}(t, r_i, c_i, 0) = -c_i + \pi_{\mathcal{H}_k}(t) r_i + (1 - \pi_{\mathcal{H}_k}(t)) V_{\mathcal{H}_k}(t, r_i, c_i). \tag{5}$$

The optimal action of caller i at time t, denoted by d_{it}, maximizes her utility. That is,

$$d_{it} = \arg \max_{d \in \{0,1\}} u_{\mathcal{H}_k}(t, r_i, c_i, \epsilon_{it}(d), d),\ t \in \{kL, ..., (k+1)L - 1\},\ k = 0, 1, ..., m-1. \tag{6}$$

Adopting the type-I extreme value assumption for the distribution of the random shocks, we derive the closed form representations of the integrated value functions and choice probabilities in Proposition 1; see the online Appendix B for its proof.

Proposition 1. *Suppose that the idiosyncratic shocks $\epsilon_{it}(1)$ and $\epsilon_{it}(0)$ have iid type-I extreme value distribution. Denoting by $P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i)$ the probability that caller i with delay announcement history $\mathcal{H}_k$ takes action $d_{it} \in \{0, 1\}$ in period t, for each $k = 0, 1, ..., m-1$, we have*

$$P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i) = \frac{\exp(v_{\mathcal{H}_k}(t, r_i, c_i, d_{it}))}{1 + \exp(v_{\mathcal{H}_k}(t, r_i, c_i, 0))},\ t \in \{kL, ..., (k+1)L - 1\}, \tag{7}$$

where $v_{\mathcal{H}_k}(t, r_i, c_i, 1)$ and $v_{\mathcal{H}_k}(t, r_i, c_i, 0)$ are given by (3) and (5), respectively. Moreover, the integrated value function $V_{\mathcal{H}_k}(t, r_i, c_i)$ is given by

$$V_{\mathcal{H}_k}(t, r_i, c_i) = \begin{cases} \log\left(1 + \exp(v_{\mathcal{H}_k}(t+1, r_i, c_i, 0))\right) & \text{if } t \neq (k+1)L - 1, \\ \sum\limits_{\mathcal{H}_{k+1} = (\mathcal{H}_k, j),\, j \in \{1, ..., J\}} \frac{a_{\mathcal{H}_{k+1}}}{a_{\mathcal{H}_k}} \log\left(1 + \exp(v_{\mathcal{H}_{k+1}}(t+1, r_i, c_i, 0))\right) & \text{if } t = (k+1)L - 1. \end{cases} \tag{8}$$

For every terminal announcement history $\mathcal{H}_k$, we have $V_{\mathcal{H}_k}(T_{\mathcal{H}_k}, r_i, c_i) = 0$ for all r_i and c_i.

4.2 Estimation methodology and results

The cost and reward parameters of each priority class are estimated separately using a two-stage approach. First, we estimate service probabilities $\pi_{\mathcal{H}_k}(\cdot)$ and the ex-ante probabilities of receiving different announcement histories $a_{\mathcal{H}_k}$ for all announcement histories directly from the data. Then, given the service probabilities and the probabilities of receiving announcements, we construct the likelihood of observed callers' actions in the data and maximize it to find the callers' parameters.

To estimate the service probabilities, we first use the Kaplan-Meier estimator (Kaplan and Meier (1958)) to estimate the waiting time distribution $F_{\mathcal{H}_k}(t)$, due to censoring because of abandonments. We assume that the waiting time distribution in the data is the equilibrium outcome. Given the waiting time distribution $F_{\mathcal{H}_k}(t)$ corresponding to delay announcement history $\mathcal{H}_k$ (obtained from our data under the current call center operations), the service probability $\pi_{\mathcal{H}_k}(t)$ is calculated as

$$\pi_{\mathcal{H}_k}(t) = \frac{F_{\mathcal{H}_k}(t+1) - F_{\mathcal{H}_k}(t)}{1 - F_{\mathcal{H}_k}(t)},\ t \in \{kL, ..., (k+1)L-1\},\ k = 0, 1, ..., m. \tag{9}$$

To find ex-ante probabilities of receiving different announcement histories $a_{\mathcal{H}_k}$, we count the number of callers that receive the announcement message sequence in $\mathcal{H}_k$ and divide it by the total number of callers in the data. We next discuss the identification of our model.

Identification. Our data exhibits significant variation in the service probabilities across different periods and in waiting times across different callers. This variation is illustrated in Figure 2, which shows the waiting time histogram and service probabilities $\pi_{\mathcal{H}_k}(t)$ for callers who receive type 1 announcement upon arrival, i.e. callers who hear that their relative position is one and the waiting time of the longest waiting caller is in [0,10]. The variation in the service probabilities and waiting times allows us to identify the reward and cost parameters separately.

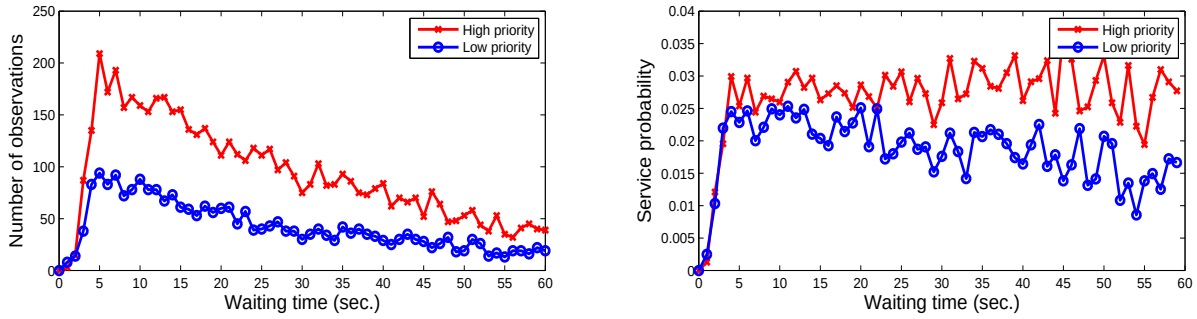

Figure 2: The waiting time histogram and service probabilities for callers who hear upon arrival that their relative position is one and the waiting time of the longest waiting caller is in [0,10].

To be specific, as shown in Figure 2, callers' waiting patterns exhibit significant variation which enables us to identify the abandonment probabilities in different periods for different delay announcement histories. In addition, using equations (3)-(5) and (7)-(8), we can write the abandonment probability of the callers with announcement history $\mathcal{H}_k$ as follows

$$P_{it}^{\mathcal{H}_k}(1; r_i, c_i) = \psi_{it}^{\mathcal{H}_k}\left(\left\{\pi_{\mathcal{H}_q}(t')r_i - c_i : k \le q \le m-1, t' \ge t, t' \in \{qL, ..., (q+1)L-1\}\right\}\right), \tag{10}$$

where $\psi_{it}^{\mathcal{H}_k}$ is a suitably defined function. Equation (10) shows that if there were no variation in the service probabilities, i.e. $\pi_{\mathcal{H}_q}(t) = \pi$ for all t and $\mathcal{H}_q$, then the abandonment probability is solely a function of $\pi r_i - c_i$. In this case, we cannot identify the reward and cost parameters separately, i.e. only $\pi r_i - c_i$ is identified. However, the inter-temporal variation of the service probabilities in the data allows us to identify the reward and cost parameters separately. Moreover, callers receive different delay announcements in our data set. Consequently, callers who have the same belief about the waiting time distribution ex-ante may receive different announcement messages at the next announcement epoch, which according to our analysis in Section 3.2 would lead to different beliefs about the waiting time distribution, and consequently, different service probabilities. As a result, delay announcements further increase the variation in the service probabilities (beyond what is shown in Figure 2). Thus, presence of delay announcements in our data set helps further with the identification of the reward and cost parameters separately.

Heterogeneity in callers' reward and cost parameters (σ_r and σ_c) is identified by the variation in callers' abandonment behavior in a given period. To see the intuition behind this, suppose that there are N callers who have waited for t periods. If there is no heterogeneity ($\sigma_r = \sigma_c = 0$), then all callers have the same reward and cost parameters, and hence, the same abandonment probability. Consequently, the total number of abandonments in period t is a Binomial random variable. On the other hand, when callers are heterogeneous ($\sigma_r \neq 0$ or $\sigma_c \neq 0$) the total number of abandonments in period t is the sum of N binary random variables with different success probabilities, which are themselves random. Therefore, under heterogeneity, the total number of abandonments exhibits more variation. Consequently, we can identify σ_r and σ_c from the variation in abandonment times across different callers.

In addition, the abandonment probabilities are more sensitive to variation in the cost parameter than variation in the reward parameter. To see this, note that by (10), increasing r_i by ε will impact the arguments of $\psi_{it}^{\mathcal{H}_k}$ (i.e. $\pi_{\mathcal{H}_q}(\cdot)r_i - c_i$) by $\pi_{\mathcal{H}_q}\varepsilon$. However, increasing c_i by ε will impact the arguments of $\psi_{it}^{\mathcal{H}_k}$ by $-\varepsilon$. Therefore, the variation in the number of callers' abandonments and its sensitivity to the reward and cost parameters across different periods helps us identify σ_r and σ_c separately. Also, note that our model is flexible enough to allow for cases where the heterogeneity across the callers is negligible; i.e. $\sigma_r = \sigma_c = 0$. In the estimation results (Table 3), we will see that this is the case for the low priority callers and for the cost parameter of the high priority callers.

Maximum likelihood estimation problem. Suppose that callers are indexed by $i = 1, ..., N$, where N is the total number of callers in the data. Recall that the reward and cost parameters of caller i are r_i and c_i, respectively, given by (1). Let W_i denote the last period in which caller i decides between waiting and abandoning. Moreover, let $\{d_{it} : t = 0, 1, ..., W_i\}$ denote the sequence of actions of callers i, where d_{it} is the action chosen in period t.

Given that the announcements are made every L periods, the number of announcements heard by caller i is $\lfloor W_i/L \rfloor + 1$. Recall that $\mathcal{H}_k$ and $P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i)$ denote the announcement history and the choice probability for $t \in \{kL, ..., (k+1)L-1\}$ and $k = 0, ..., m-1$. The likelihood of observing the sequence of actions $\{d_{it} : t = 0, 1, ..., W_i\}$ by caller i is given by

$$\ell_i(\Theta) = \int\int \prod_{k=0}^{\lfloor \frac{W_i}{L} \rfloor} \prod_{t=kL}^{\min(W_i,(k+1)L-1)} P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i)\ \phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i}. \tag{11}$$

The likelihood of the entire sample is the product of each individual caller's likelihood and has the following form:

$$L(\Theta) = \prod_{i=1}^{N} \int\int \prod_{k=0}^{\lfloor \frac{W_i}{L} \rfloor} \prod_{t=kL}^{\min(W_i,(k+1)L-1)} P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i)\ \phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i}. \tag{12}$$

To estimate the structural parameters Θ, we maximize the log-likelihood function $\log L(\Theta)$ subject to the following constraints: the integrated value functions given by (8), the abandonment probabilities given by (7) and the nominal utilities given by (3)-(5) for all $i = 1, ..., N$, $k =$

$0, ..., m-1,\ t \in \{kL, ..., (k+1)L-1\},\ \mathcal{H}_k \in \{1, ..., J\}^{k+1}$, and (1).

To evaluate the likelihood function, we need to compute the integrated value functions corresponding to all $\mathcal{H}_k \in \{1, ..., J\}^{k+1}$, $k = 0, ..., m-1$, at all $t \in \{kL, ..., (k+1)L-1\}$. Therefore, the number of equations for the integrated value function, which contributes directly to the complexity of the maximization problem, is $L\sum_{k=0}^{m-1} J^{k+1} = LJ(J^m - 1)/(J-1)$. Consequently, the complexity of the estimation problem increases by the number of announcement messages J, number of scheduled announcements m and number of periods between two consecutive announcements L.

We use the non-linear optimization solver KNITRO (Byrd et al. (2006)) with AMPL interface to solve the maximum likelihood estimation problem. We solve the problem for 100 randomly generated starting points and find the estimates corresponding to the highest likelihood value. To approximate the two dimensional integrations in the likelihood function, we use the Gauss-Hermite integration method (Judd (1998), Chapter 7.2) considering five nodes in each dimension. We also conduct a Monte-Carlo experiment to illustrate the capability of our estimation method to recover true parameter values; see the online Appendix C for the details.

Estimation results. We estimate the parameters of the high and low priority groups separately. In other words, we solve the MLE problem for each group in isolation. We assume that callers make their decisions between waiting and abandoning every 5 seconds. Since our data is more granular, consistent with our modeling assumptions in Section 4.1, we truncate the abandonment times downward and the service initiation times upward. The estimated parameters of the callers and their standard errors (shown in parenthesis) are reported in Table 3. To compute the standard errors, we use the parametric bootstrap method (Horowitz (2001)). Table 4 shows the mean and standard deviation for callers' rewards and cost parameters, which are calculated from the estimates in Table 3.[6]

Priority group	m_r	m_c	σ_r	σ_c
High priority	1.856	–2.336	0.202	3.44E–05
	(0.012)	(0.043)	(0.015)	(0.101)
Low priority	1.734	–2.326	3.37E–05	9.91E–06
	(0.009)	(0.030)	(0.035)	(0.081)

Table 3: The estimation results.

Priority group	r-Mean (\$)	c-Mean (\$/minute)	r-St.Dev.	c-St.Dev.
High priority	6.527	1.161	1.333	4.79E–04
Low priority	5.661	1.173	1.91E–04	1.39E–04

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

As can be seen in Table 4, the reward parameter of the high priority callers is larger than that of the low priority callers. Therefore, the high priority callers value service more than the low priority callers. This observation along with the fact that the high and low priority callers have approximately the same cost parameters shows that the low priority callers are less patient than the high priority callers. This can also be confirmed by the summary statistics of the observations in the data in Table 2. As can be seen in Table 2, the average waiting time of the low priority callers

[6]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 $\exp(m_c + \sigma_c^2/2)\sqrt{\exp(\sigma_c^2) - 1}$.

is only 11.5 seconds longer than the average waiting time of the high priority callers; however, the low priority callers abandon twice as much as the high priority callers. The standard deviation of the reward parameter for the high priority callers is positive. This shows that the high priority callers are heterogeneous in their rewards from receiving service.

Given the estimation results in Table 4, the ratio of the reward and cost parameters (r/c) for the high and low priority callers are 5.62 and 4.83, respectively, even though as shown in Table 2 the average of callers' observed abandonment times in the data is around 1 minute for both the high and low priority callers. At first, this may seem problematic. However, given that the abandonment times of the callers are censored (heavily by the service process), the observed abandonment times in the data are not an indicator of callers' mean patience threshold, and cannot be compared with r/c directly. An analysis (available from the authors) that estimates the mean of callers' abandonment time using the Kaplan-Meier estimator (Kaplan and Meier (1958)) and corrects its bias using the Jack-Knife method in Stute and Wang (1994) shows that the bias corrected mean of the abandonment time are 5.55 and 5.16 for the high and low priority callers, respectively, which are fairly close to r/c.

Out of Sample Testing. To examine the ability of our model to predict callers' abandonment behavior, we perform out-of-sample tests. We split the data for each priority group to two sets: calls between 9 am and 12 pm, and calls between 12 pm and 4 pm. We consider one of these sets as the training data, and the other one as the test data. We first estimate the parameters of the callers in the training data, which will be used to predict the number of abandonments in the test data. To do so, we estimate the service probabilities and the probabilities of receiving different announcement histories in the test data. Then, using those with the parameter estimates from the training data, we predict the number of abandonments in the test data.

Let $P^{\mathcal{H}_k}_{aban}(t)$ denote the probability that a caller in the test data receives announcement history $\mathcal{H}_k$ and abandons in period $t \in \{kL, ..., (k+1)L-1\}$.[7] The predicted number of calls that the corresponding callers receive announcement history $\mathcal{H}_k$ and abandon in period t is $NP^{\mathcal{H}_k}_{aban}(t)$, where N is the total number of calls in the test data. Also, let $Q^{\mathcal{H}_k}_{aban}(t)$ denote the actual number of calls that the corresponding callers with announcement history $\mathcal{H}_k$ abandon in period t. Then the total number of predicted and actual abandonments are $\sum_{\mathcal{H}_k}\sum_{t=kL}^{(k+1)L-1} NP^{\mathcal{H}_k}_{aban}(t)$ and $\sum_{\mathcal{H}_k}\sum_{t=kL}^{(k+1)L-1} Q^{\mathcal{H}_k}_{aban}(t)$, respectively. To examine the accuracy of the prediction, we consider the relative and absolute errors given by

$$\text{Relative Error} = \frac{|\sum_{\mathcal{H}_k}\sum_{t=kL}^{(k+1)L-1} Q^{\mathcal{H}_k}_{aban}(t) - \sum_{\mathcal{H}_k}\sum_{t=kL}^{(k+1)L-1} NP^{\mathcal{H}_k}_{aban}(t)|}{\sum_{\mathcal{H}_k}\sum_{t=kL}^{(k+1)L-1} Q^{\mathcal{H}_k}_{aban}(t)}, \tag{13}$$

[7] The probability $P^{\mathcal{H}_k}_{aban}(t)$ has the following form:

$$P^{\mathcal{H}_k}_{aban}(t) = a_{\mathcal{H}_k} \int\int \Big((1-\pi_{\mathcal{H}_k}(t))P^{\mathcal{H}_k}_{it}(1;r_i,c_i)\Big)\Big(\prod_{q=0}^{k}\prod_{s=qL}^{\min(t-1,(q+1)L-1)}(1-\pi_{\mathcal{H}_q}(s))P^{\mathcal{H}_q}_{is}(0;r_i,c_i)\Big)\,\phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i}.$$

The term $a_{\mathcal{H}_k}$ on the right hand side is the probability of receiving announcement history $\mathcal{H}_k$. The first term inside the integral is the probability of not receiving service and abandoning in period t. The second term inside the integral is the probability of not receiving service and not abandoning in periods before t.

$$\text{Absolute Error} = \frac{1}{N}|\sum_{\mathcal{H}_k}\sum_{t=kL}^{(k+1)L-1} Q^{\mathcal{H}_k}_{aban}(t) - \sum_{\mathcal{H}_k}\sum_{t=kL}^{(k+1)L-1} NP^{\mathcal{H}_k}_{aban}(t)|. \tag{14}$$

Table 5 shows the relative and absolute errors of prediction for different combinations of the training and test data for the high and low priority classes. The actual abandonment rates for the high priority callers in (9 am-12 pm) and (12 pm-4 pm) are 18.42% and 17.44%, respectively. For the low priority callers in (9 am-12 pm) and (12 pm-4 pm), the actual abandonment rates are 33.18% and 33.16%, respectively.

Training set	Test set	Relative Error	Absolute Error
High priority (9 am-12 pm)	High priority (12 pm-4 pm)	13.94 %	2.43 %
High priority (12 pm-4 pm)	High priority (9 am-12 pm)	3.68 %	0.68 %
Low priority (9 am-12 pm)	Low priority (12 pm-4 pm)	7.88 %	2.61 %
Low priority (12 pm-4 pm)	Low priority (9 am-12 pm)	11.42 %	3.79 %
Average across all tests		9.23 %	2.38 %

Table 5: The relative and absolute errors in predicting the abandonment rates.

As can be seen in Table 5, our model is capable of predicting the abandonment rates with an average absolute error less than 2.5% and an average relative error less than 10%.

To test the performance of our model in predicting callers' abandonment behavior for a case where the training set and the test set do not exhibit time of the day effect and are similar in terms of callers' abandonment behavior, we focus on observations between 2pm and 4pm. We split the data to observations in January-April, and observations in May-July. Table 6 shows the absolute and relative errors in predicting the abandonment rates for different combinations of the training and test sets.

Training set (2pm-4pm)	Test set (2pm-4pm)	Relative Error	Absolute Error
High priority (Jan.-April)	High priority (May-July)	5.57 %	1.04 %
High priority (May-July)	High priority (Jan.-April)	1.90 %	0.36 %
Low priority (Jan.-April)	Low priority (May-July)	1.63 %	0.60 %
Low priority (May-July)	Low priority (Jan.-April)	1.59 %	0.57 %
Average across all tests		2.67 %	0.64 %

Table 6: The relative and absolute errors in predicting the abandonment rates for cases with a lower level of time of the day effect.

Comparison of the results in Tables 5 and 6 indicates that the average of the errors decreases by more than 50% for the case without a time of the day effect. In other words, the out-of-sample predictions are more accurate if the training and test sets are similar in terms of callers' abandonment behavior.

We also compared the prediction power of our stopping time model with a simpler (myopic) decision making model as an alternative model for callers' abandonment behavior. We demonstrate that our model has a better performance. See the online Appendix E for more details.

5 A Framework to Study the Impact of Delay Announcements

In this section, we develop a framework to study the impact of delay announcements on the performance of a system with two priority classes. For simplicity, we assume callers receive a delay

announcement only once upon arrival. The announcement provided by the call center manager affects callers' anticipation of their waiting time distributions and their chances of receiving service. This in turn influences callers' abandonment behavior, in particular the probability that callers abandon at a given time while waiting. Moreover, a change in callers' abandonment behavior affects the system performance and, consequently, callers' waiting time distributions. We are particularly interested in finding the equilibrium of the system, where callers' anticipation of their waiting time distributions, which are affected by the announcements, match their actual experience in the system. The analysis of this system is complex and requires several substantial steps.

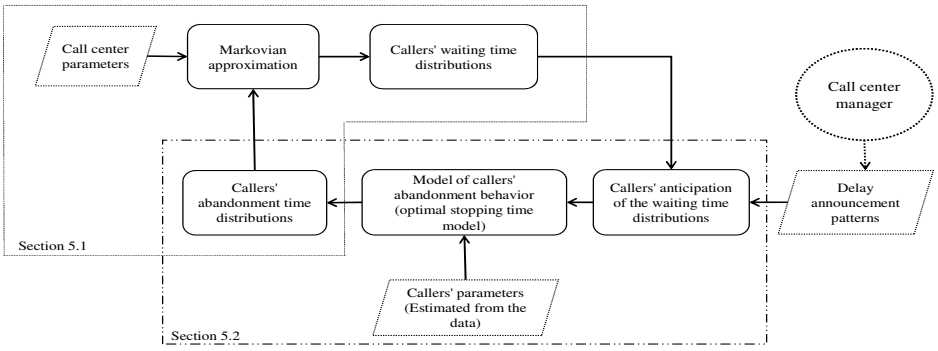

Figure 3: The building blocks for the equilibrium characterization.

Figure 3 shows different blocks for the analysis of the delay announcements' impact on the system and the sections of the paper that describe the analyses involving the corresponding blocks. We use a Markovian approximation of a two-class queueing system to derive callers' waiting time distributions from callers' abandonment time distributions and call center parameters such as the arrival and service rates. This Markovian approximation is described in Section 5.1.

Moreover, we represent the abandonment time distributions as endogenous functions of callers' reward and cost parameters and their anticipated waiting time distributions (using the optimal stopping model of Section 4.1). This representation is laid out in Section 5.2. In addition, in Section 5.2, we derive a set of equations that characterizes the equilibrium of the system for different choices of delay announcement messages.

5.1 A Markovian approximation of the call center operations

In this section, we introduce a Markovian approximation to simplify the underlying queuing analysis. Using this approximation, we derive the steady-state probabilities of the system and the waiting time distributions of the callers given their abandonment time distributions.

Preliminaries. Consider a call center with two priority classes: high and low. Subscripts "h" and "l" will be used to represent the high and low priority callers. Suppose that both the high and low priority callers are served by a single pool of agents that consists of s agents. The interarrival times for the high and low priority callers have i.i.d. exponential distributions with rates λ_h and λ_l, respectively. Denote by $\lambda = \lambda_h + \lambda_l$ the total arrival rate to the system. The service times for the high and low priority callers have i.i.d. exponential distributions with rates μ_h and μ_l, respectively.

We initially assume that $\mu_h = \mu_l = \mu$. Later in this section, we relax this assumption.

The sequencing policy in the call center is a point updating policy. There are two separate queues: high priority queue and low priority queue. The high (low) priority callers join the end of the high (low) priority queue. The priority point of the low priority callers upon arrival is zero, however, the high priority callers obtain τ points upon arrival. Callers' points increase by their waiting times. In particular, a low priority caller who has been waiting for t time units has t points; however, a high priority caller who has been waiting for t time units has $t + \tau$ points. If an agent becomes available, the agent compares the points of the callers at the head of the high and low priority queues and selects the caller with the higher points to render service. Note that a newly arriving high priority caller has higher points than any low priority caller who has arrived in the past τ time units. The average number of the low priority callers who arrive in τ time units is $\lambda_l \tau$. Consequently, a newly arriving high priority caller may pass up to approximately $[\lambda_l \tau]$ low priority callers. We let $[x]$ denote the nearest integer to x.

Analyzing the evolution of this two class system requires tracking the number of callers in both the high and low priority queues. To decrease the dimension of the tracking process, we use the state-space collapse approximation discussed next.

State space collapse approximation. Let n_h and n_l denote the number of the high and low priority customers in the queue, respectively. The total number of callers in the queues is then $n = n_h + n_l$. Inspired by the heavy traffic literature (Harrison (1988), Bramson (1998)), we use the state space collapse approximation to represent n_h and n_l as functions of the total queue length n. To this end, we approximate the expected time between successive arrivals of the high and low priority callers by $1/\lambda_h$ and $1/\lambda_l$, respectively. Assuming that abandonments are rare compared to service completions, the delay experienced by the caller at the head of the high priority queue is approximated by n_h/λ_h; and she has $n_h/\lambda_h + \tau$ priority points. Similarly, the delay of the caller at the head of the low priority queue is approximated by n_l/λ_l; and she has n_l/λ_l priority points.

Using the state space collapse approximation, we assume that the priority points of the two callers at the head of the two priority queues are approximately equal.[8] Therefore, given n, we find the values of n_h and n_l that minimize the difference between the points of the callers at the head of the high and low priority queues. In other words, we solve the following minimization problem:

$$\min_{n_h, n_l} \left| \frac{n_h}{\lambda_h} + \tau - \frac{n_l}{\lambda_l} \right| \text{ subject to } n = n_l + n_h,\ 0 \le n_l, n_h \le n. \tag{15}$$

The solution of (15) is given by

$$n_h = \max\left(0, \left[\frac{\lambda_h}{\lambda_h + \lambda_l}(n - \lambda_l \tau)\right]\right) \text{ and } n_l = n - n_h. \tag{16}$$

Under large abandonment rates, the state-space collapse approximation given in (16) may not be as accurate as desired. For such cases, we propose a heuristic refinement of the approximation

[8] If this were not the case, then one of the classes would have strict priority over the other. Then, in the heavy traffic limit, that class becomes empty instantaneously. Therefore, whenever both queues are non-empty, they should have the same priority in the heavy traffic limit. Motivated by this, we look for solutions in which the difference in the priorities of the first customer in each queue is minimized.

in the online Appendix D that adjusts the solution of (15) and improves the accuracy of the approximation.

State dependent waiting time and abandonment time distributions. The call center manager announces the total number n of callers waiting in the two queues upon arrival. Therefore, we let $F_n^h(t)$ and $F_n^l(t)$ denote the virtual waiting time distributions for the high and low priority callers, respectively, who arrive when there are n callers in total in the high and low priority queues and all agents are busy.[9] Similarly, we let $Z_n^h(t;r,c)$ and $Z_n^l(t;r,c)$ denote the abandonment time distributions for the high and low priority callers, respectively, who arrive when there are n callers in the two queues, and have reward and cost parameters r and c. Also let $\Gamma_n^h(t;r,c)$ and $\Gamma_n^l(t;r,c)$ denote the hazard rates of $Z_n^h(t;r,c)$ and $Z_n^l(t;r,c)$, respectively. Since the callers' reward and cost parameters are not observable, we use the expected abandonment time distributions and the expected hazard rates in the Markovian approximation. To this end, let $G_n^h(t)$ and $G_n^l(t)$ denote the expected abandonment time distributions for the high and low priority callers, respectively, who arrive when there are n callers in total in the two queues. Also let $H_n^h(t)$ and $H_n^l(t)$ denote the expected hazard rates for the high and low priority callers, respectively, who arrive when there are n callers in total in the two queues. That is, for $\eta \in \{h,l\}$

$$G_n^\eta(t) = \int\int Z_n^\eta(t;r,c)\phi(y_1)\phi(y_2)dy_1dy_2 \quad \text{and} \quad H_n^\eta(t) \quad = \int\int \Gamma_n^\eta(t;r,c)\phi(y_1)\phi(y_2)dy_1dy_2, \quad (17)$$

where $r = \exp(m_r^\eta + \sigma_r^\eta y_1)$, $c = \exp(m_c^\eta + \sigma_c^\eta y_2)$ and y_1, y_2 are i.i.d. standard normal random variables as in (1). In this section, we assume that $G_n^\eta(t)$ and $H_n^\eta(t)$ are known for all n and $\eta \in \{h,l\}$. Then, we use the Markovian approximation to find the steady-state probabilities of the system and the virtual waiting time distributions $F_n^\eta(t),\ \eta \in \{h,l\}$.

Steady-state probabilities. To derive the steady-state probabilities, we build on Whitt (2005) and approximate the system by a birth-and-death process with state-dependent abandonment rates. The abandonment rate from the system is the sum of the abandonment rates of all callers in the two queues. Recall that a caller's abandonment time distribution (and the corresponding hazard rate) depend on the total number of callers in the two queues upon her arrival.

Under the assumption that abandonments are rare compared to service completions, the caller who is the i_h^{th} caller from the end of the high priority queue (caller i_h) has been waiting for approximately i_h/λ_h time units. Let $m_h(i_h,n)$ denote the total number of callers in the two queues upon caller i_h's arrival given that there are n callers in the two queues currently. We approximate $m_h(i_h,n)$ as follows:

$$m_h(i_h,n) = [s\mu\frac{i_h}{\lambda_h}] + (n_h - i_h) + (n_l - [\lambda_l\frac{i_h}{\lambda_h}]). \quad (18)$$

Note that given the total number of callers n in the two queues, we solve (15) to derive the corresponding n_h and n_l. The first term on the right hand side of (18) is the number of callers who have entered service since caller i_h's arrival. The second and the third terms are the numbers

[9] Note that $F_{\mathcal{H}_k}(t)$ in Section 4.2 denotes the waiting time distribution corresponding to a caller with announcement history $\mathcal{H}_k$, while $F_n(t)$ in this section denotes the virtual waiting time distribution of callers depending on the state of the system upon arrival.

of the high and low priority callers, respectively, who are currently in the queue and were in the system upon caller i_h's arrival. The term $[\lambda_l i_h/\lambda_h]$ approximates the number of low priority callers who joined the system after caller i_h's arrival. As a result, the abandonment time distribution of caller i_h is $G^h_{m_h(i_h,n)}(\cdot)$. Similarly, given that caller i_h has been waiting for approximately i_h/λ_h time units, her (expected) abandonment rate is $H^h_{m_h(i_h,n)}(i_h/\lambda_h)$.

Analogously, the total number of callers in the two queues upon arrival of the ${i_l}^{th}$ caller from the end of the low priority queue (caller i_l), denoted by $m_l(i_l, n)$, is approximated by

$$m_l(i_l, n) = [s\mu\frac{i_l}{\lambda_l}] + (n_l - i_l) + (n_h - [\lambda_h\frac{i_l}{\lambda_l}]). \tag{19}$$

Consequently, the abandonment time distribution of caller i_l is $G^l_{m_l(i_l,n)}(\cdot)$, and her (expected) abandonment rate is $H^l_{m_l(i_l,n)}(i_l/\lambda_l)$.

Let δ_n denote the total abandonment rate from the system when the total number of callers in the high and low priority queues is n. The abandonment rate is zero when the queues are empty, i.e. $\delta_0 = 0$. For $n \geq 1$, the total abandonment rate is the sum of abandonment rates of all callers in the high and low priority queues, given by

$$\delta_n = \sum_{i_h=1}^{n_h} H^h_{m_h(i_h,n)}(\frac{i_h}{\lambda_h}) + \sum_{i_l=1}^{n_l} H^l_{m_l(i_l,n)}(\frac{i_l}{\lambda_l}). \tag{20}$$

We approximate the stochastic process governing the number of callers in the system by a birth-and-death process. The birth rate is the total arrival rate to the system denoted by $\lambda = \lambda_h + \lambda_l$. When there are k callers in the system, the death rate μ_k is given by

$$\mu_k = \begin{cases} k\mu & \text{if } \ 1 \leq k \leq s, \\ \\ s\mu + \delta_{k-s} & \text{if } \ k \geq s+1. \end{cases} \tag{21}$$

Let γ_k denote the steady-state probability of having k callers in the system. The steady-state probabilities are given by the balance equations:

$$\lambda\gamma_k = \mu_{k+1}\gamma_{k+1}, k = 0, 1, 2, ..., \quad \text{and} \quad \sum_{k=0}^{\infty}\gamma_k = 1.$$

Approximating the virtual waiting time distribution. Following Whitt (2005), we approximate the waiting time of the served calls by a sum of exponential random variables, whose rates depend on the service rate of the system and the abandonment rates of the callers present in the high and low priority queues.

To derive the virtual waiting time distributions $F^h_n(t)$ and $F^l_n(t)$, consider a caller who joins the system when the total number of callers in the high and low priority queues is n. We track the evolution of the system until all callers in the high and low priority queues with priority points higher than the caller of interest enter service or abandon. Then, the caller of interest will be the next to enter service.

To this end, it is essential to establish the order in which the existing high and low priority callers enter service. Although it is virtually impossible to figure out the exact order of callers, we

approximate it by sorting the callers based on their (approximate) priority points. Recall that the (approximate) priority point of the high priority caller i_h (from the end of the high priority queue) is $i_h/\lambda_h + \tau$. Similarly, the priority point of the low priority caller i_l is i_l/λ_l. Figure 4 shows the priority points of the callers in the high and low priority queues.

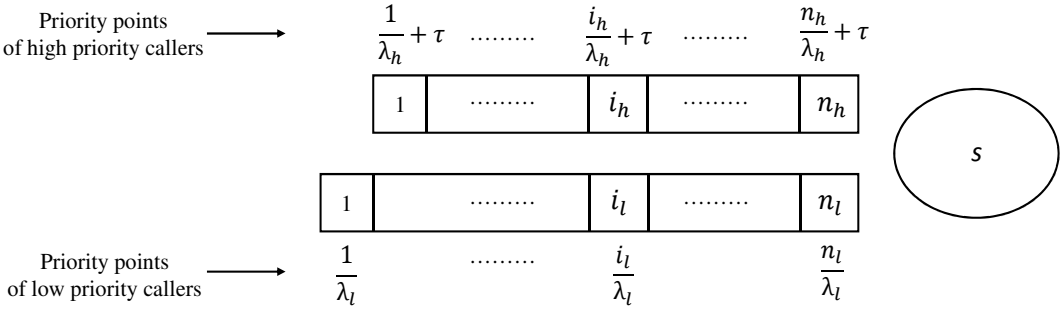

Figure 4: The priority points of the callers in the high and low priority queues.

For the high priority callers, let $O_h(i_h, n)$ denote the order in which caller i_h enters service. We have

$$O_h(i_h, n) = n - (i_h - 1) - \min(\lfloor \lambda_l(\frac{i_h}{\lambda_h} + \tau)\rfloor, n_l), \tag{22}$$

where $\lfloor \lambda_l(i_h/\lambda_h + \tau)\rfloor$ is the largest integer less than or equal to $\lambda_l(i_h/\lambda_h + \tau)$. The second and the third terms on the right hand side of (22) are the number of the high and low priority callers with lower priority points than caller i_h, respectively. Similarly, for the low priority callers, the order in which caller i_l enters service, denoted by $O_l(i_l, n)$, is given by

$$O_l(i_l, n) = n - (i_l - 1) - \min(\lfloor \lambda_h(\frac{i_l}{\lambda_l} - \tau)\rfloor^+, n_h), \tag{23}$$

where $\lfloor \lambda_h(i_l/\lambda_l - \tau)\rfloor^+ = \max(0, \lfloor \lambda_h(i_l/\lambda_l - \tau)\rfloor)$. The second and the third terms on the right hand side of (23) are the number of the low and high priority callers with lower priority points than caller i_l, respectively. The function O_l (O_h) gives the order of entering service for the low (high) priority callers across all callers in the two queues. For example, a low priority caller that has a lower order number than a high priority caller enters service sooner. Also, if a caller has the order number 1, then she is the first caller to enter service irrespective of her priority class.

We next derive the virtual waiting time distribution for the high priority callers. We approximate each inter-departure time (including abandonments) by exponential random variables. To be more specific, each inter-departure time is modeled as a minimum of several exponential random variables, and hence, its rate is the sum of the rates of those exponential random variables.

Suppose that a new high priority caller arrives when there are n ($n > [\lambda_l\tau]$) callers in total in the queues.[10] Recall that the high priority callers are advanced in the system. In particular, the new high priority caller can pass approximately $[\lambda_l\tau]$ low priority callers. Therefore, the new high priority caller will enter service after $n - [\lambda_l\tau] + 1$ departures. The time between the new high priority caller's arrival and the first departure is modeled as an exponential random variable with

[10]When $n \le [\lambda_l\tau]$, the solution of (15) is $n_h = 0$ and $n_l = n \le [\lambda_l\tau]$. In this case, the new high priority caller passes all callers in the low priority queue and will be the next customer who enters service. Therefore, her waiting time distribution is an exponential random variable with rate $s\mu$.

rate $x_1^h(n)$ given by

$$x_1^h(n) = s\mu + \sum_{i_h=1}^{n_h} H^h_{m_h(i_h,n)}(\frac{i_h}{\lambda_h}) + \sum_{i_l=[\lambda_l\tau]+1}^{n_l} H^l_{m_l(i_l,n)}(\frac{i_l}{\lambda_l}). \tag{24}$$

The first term on the right hand side of (24) is the service rate. The second term is the abandonment rate of the high priority callers who are ahead of the new high priority caller. The last term is the abandonment rate of the low priority callers who are not passed by the new high priority caller.

To consider subsequent departures, we make two assumptions about the system dynamics as in Whitt (2005). First, given an ordering of callers, we assume that they depart the queue in that order irrespective of whether they enter service or abandon. Second, we assume that each inter-departure time is $1/\lambda$, i.e the total departure rate from the queues is λ. Consequently, after each departure we add $1/\lambda$ to the time the callers have been waiting to calculate their abandonment rates, i.e. the hazard rates of the abandonment time distributions. Under these assumptions, the time between the first and the second departures is approximated by an exponential random variable with rate $x_2^h(n)$ given by

$$x_2^h(n) = s\mu + \sum_{i_h=1}^{n_h} H^h_{m_h(i_h,n)}(\frac{i_h}{\lambda_h} + \frac{1}{\lambda})\ \mathbb{I}_{\{O_h(i_h,n)\geq 2\}} + \sum_{i_l=[\lambda_l\tau]+1}^{n_l} H^l_{m_l(i_l,n)}(\frac{i_l}{\lambda_l} + \frac{1}{\lambda})\ \mathbb{I}_{\{O_l(i_l,n)\geq 2\}}. \tag{25}$$

Note that the indicator function in (25), $\mathbb{I}_{\{.\}}$, is one if the order of receiving service for the corresponding caller is greater than or equal to two, i.e. the corresponding caller is not the first one to depart. Repeating the above procedure, we can show that the time between the $(q-1)^{th}$ and q^{th} departure is modeled as an exponential random variable with rate $x_q^h(n)$ given as follows:

$$\begin{aligned} x_q^h(n) = s\mu &+ \sum_{i_h=1}^{n_h} H^h_{m_h(i_h,n)}(\frac{i_h}{\lambda_h} + (q-1)\frac{1}{\lambda})\ \mathbb{I}_{\{O_h(i_h,n)\geq q\}} \\ &+ \sum_{i_l=[\lambda_l\tau]+1}^{n_l} H^l_{m_l(i_l,n)}(\frac{i_l}{\lambda_l} + (q-1)\frac{1}{\lambda})\ \mathbb{I}_{\{O_l(i_l,n)\geq q\}}. \end{aligned} \tag{26}$$

The waiting time of the high priority caller of interest is approximated by the sum of exponential random variables with rates $x_q^h(n)$, $1 \leq q \leq n - [\lambda_l\tau] + 1$. We characterize the waiting time distributions via their Laplace transforms. Then, we invert the Laplace transforms using a numerical transform inversion to derive the waiting time distributions. The Laplace transform of $F_n^h(t)$, denoted by $\mathcal{L}_{F_n^h(t)}(z)$, is given by

$$\mathcal{L}_{F_n^h(t)}(z) = \frac{1}{z} \prod_{q=1}^{n-[\lambda_l\tau]+1} \frac{x_q^h(n)}{z + x_q^h(n)}. \tag{27}$$

To invert the Laplace transforms, we use the Euler method introduced in Whitt and Abate (1995).

We next derive the virtual waiting time distribution for a low priority caller who arrives when there are n callers in total in the two queues. Because this low priority caller of interest may be passed by some high priority callers, we cannot ignore the future arrivals as we did in the derivation of the waiting time distribution for the high priority callers. Let $b_h(n)$ denote the number of the

high priority callers who will pass the low priority caller of interest during her stay in the system. Under the assumption that each departure takes approximately $1/\lambda$ time units, the average waiting time of the low priority caller of interest is $(n+b_h(n)+1)/\lambda$. The low priority caller of interest will be passed by high priority callers who arrive in the next τ time units. Therefore, the low priority caller of interest will be passed by at most $\lambda_h\tau$ high priority callers, and if her average waiting time $(n+b_h(n)+1)/\lambda$ is less than τ by only $\lambda_h(n+b_h(n)+1)/\lambda$ callers. Hence, the number of the high priority callers who pass the low priority caller is $\lambda_h \min((n+b_h(n)+1)/\lambda, \tau)$ which should be equal to $b_h(n)$ by definition. Therefore, we have

$$b_h(n) = \lambda_h \min(\frac{n+b_h(n)+1}{\lambda}, \tau). \tag{28}$$

The solution of (28) has the following form (see Lemma 1 in the online Appendix B):

$$b_h(n) = \begin{cases} \lambda_h\tau & \text{if } \ n \geq \lambda_l\tau - 1, \\ \\ \frac{\lambda_h}{\lambda_l}(n+1) & \text{if } n < \lambda_l\tau - 1. \end{cases} \tag{29}$$

We divide callers who enter service sooner than the low priority caller of interest into four groups, which are illustrated in Figure 5. The first group consists of the high priority callers in the queue upon arrival of the low priority caller of interest. The order of receiving service for these callers is $O_h(i_h, n),\ i_h \in \{1, ..., n_h\}$. The second group includes the low priority callers who will not be passed by any high priority callers. The order of receiving service for these callers is $O_l(i_l, n),\ i_l \in \{[\lambda_l\tau], ..., n_l\}$. The third group consists of the low priority callers who will be passed by some high priority callers. If no high priority caller passes these group of callers, their order of receiving service would be $O_l(i_l, n),\ i_l \in \{1, .., [\lambda_l\tau]\}$. However, because these callers are passed by some high priority callers, their order of receiving service is $\widetilde{O}_l(i_l, n)$ given by[11]

$$\widetilde{O}_l(i_l, n) = O_l(i_l, n) + ([\lambda_l\tau] - i_l)[b_h(n)]/[\lambda_l\tau],\ i_l \in \{1, .., [\lambda_l\tau]\}.$$

The fourth group consists of the high priority callers who pass the low priority caller of interest. The order of receiving service for these callers is $\widetilde{O}_h(i'_h, n)$ given by[12]

$$\widetilde{O}_h(i'_h, n) = n - [\lambda_l\tau] + (\frac{\lambda_l}{\lambda_h} + 1)i'_h + 1,\ i'_h \in \{1, ..., [b_h(n)]\}.$$

[11] Recall that the low priority caller of interest will be passed by $[b_h(n)]$ high priority callers. Moreover, the $[\lambda_l\tau]^{th}$ caller from the end of the low priority queue will not be passed by any high priority caller. Consequently, caller i_l ($i_l \leq [\lambda_l\tau] - 1$) will be passed by approximately $([\lambda_l\tau] - i_l)[b_h(n)]/[\lambda_l\tau]$ high priority callers, which is a linear interpolation between 0 and $[b_h(n)]$.

[12] Caller i'_h joins the system approximately i'_h/λ_h time units after the low priority caller of interest arrives. Therefore, $\lambda_l i'_h/\lambda_h$ new low priority callers have arrived since the arrival of the low priority caller of interest. Given that each high priority caller passes approximately $[\lambda_l\tau]$ low priority callers, caller i'_h passes $[\lambda_l\tau] - \lambda_l i'_h/\lambda_h - 1$ low priority callers who were in the system before the low priority caller of interest arrives (out of the total n callers in the queues). In addition, $i'_h - 1$ high priority callers arrived before caller i'_h and will enter service sooner than caller i'_h. Therefore, the total number of callers who enter service sooner than caller i'_h is $n + i'_h - 1 - ([\lambda_l\tau] - \lambda_l i'_h/\lambda_h - 1) = n - [\lambda_l\tau] + (\lambda_l/\lambda_h + 1)i'_h$, i.e. her order of receiving service is $n - [\lambda_l\tau] + (\lambda_l/\lambda_h + 1)i'_h + 1$.

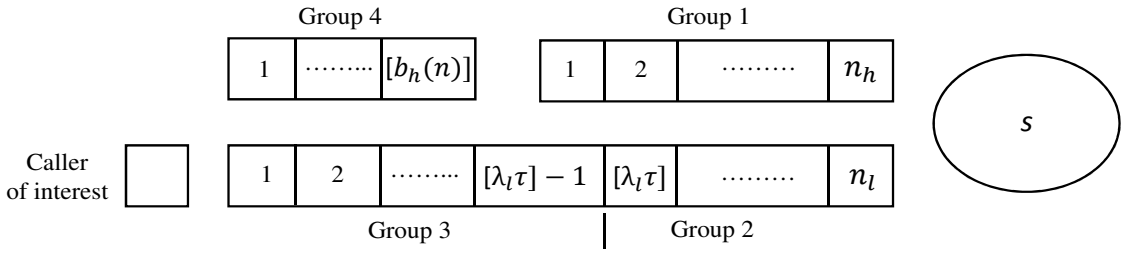

Figure 5: Four groups of callers that enter service sooner than the low priority caller of interest.

Given the four groups defined above, we approximate the time between the $(q-1)^{th}$ and q^{th} departures after arrival of the low priority caller of interest by an exponential random variable with rate $x_q^l(n)$ given by

$$\begin{aligned} x_q^l(n) = s\mu &+ \sum_{i_h=1}^{n_h} H^h_{m_h(i_h,n)}\left(\frac{i_h}{\lambda_h} + (q-1)\frac{1}{\lambda}\right) \mathbb{I}_{\{O_h(i_h,n)\geq q\}} \\ &+ \sum_{i_l=[\lambda_l\tau]}^{n_l} H^l_{m_l(i_l,n)}\left(\frac{i_l}{\lambda_l} + (q-1)\frac{1}{\lambda}\right) \mathbb{I}_{\{O_l(i_l,n)\geq q\}} \\ &+ \sum_{i_l=1}^{[\lambda_l\tau]-1} H^l_{m_l(i_l,n)}\left(\frac{i_l}{\lambda_l} + (q-1)\frac{1}{\lambda}\right) \mathbb{I}_{\{\widetilde{O}_l(i_l,n)\geq q\}}, \\ &+ \sum_{i'_h=1}^{[b_h(n)]} H^h_n\left(-\frac{i'_h}{\lambda_h} + (q-1)\frac{1}{\lambda}\right) \mathbb{I}_{\{\widetilde{O}_h(i'_h,n)\geq q\}}. \end{aligned} \tag{30}$$

The first term on the right hand side of (30) is the service rate. The second to fifth terms correspond to the abandonment rates of the callers in the first to fourth groups explained above, respectively. Every high priority caller in the fourth group will find n callers in the queues upon arrival, and consequently, her hazard rate will be $H_n^h(\cdot)$. The reason is that by our assumption the total departure rate from the system is approximately λ, which is also the total arrival rate. Consequently, the total number of callers in the queues remains approximately fixed at n. In addition, we allow that the argument of H_n^h, i.e. $(-i'_h/\lambda_h + (q-1)/\lambda)$ which is the amount of time caller i'_h has been waiting, takes negative values.[13] This happens if caller i'_h has not arrived yet. Therefore, we assume that the hazard rate functions are zero for negative arguments, i.e. $H_n^h(-t) = 0$ for $t > 0$.

The waiting time for the low priority caller of interest is approximated by the sum of exponential random variables with rates $x_q^l(n)$, $1 \leq q \leq n + b_h(n) + 1$. The Laplace transform of $F_n^l(t)$ denoted by $\mathcal{L}_{F_n^l(t)}(z)$ is given by

$$\mathcal{L}_{F_n^l(t)}(z) = \frac{1}{z} \prod_{q=1}^{n+[b_h(n)]+1} \frac{x_q^l(n)}{z + x_q^l(n)}. \tag{31}$$

Relaxing the assumption that $\mu_h = \mu_l$. Recall that we assumed that abandonments are

[13] Caller i'_h in the fourth group will arrive i'_h/λ_h time units after the low priority caller of interest arrives, and $(q-1)^{th}$ departure occurs approximately $(q-1)/\lambda$ time units after the low priority caller of interest arrives. Therefore, caller i'_h has been waiting for $(-i'_h/\lambda_h + (q-1)/\lambda)$ time units.

rare compared to service completions. An implication of this assumption is that a caller who has entered service is a high priority caller with probability λ_h/λ and is a low priority caller with probability λ_l/λ. Consequently under the assumption that abandonments are rare compared to service completions, an arbitrary busy agent serves the high priority callers with probability λ_h/λ and serves the low priority callers with probability λ_l/λ. Hence, the long run average service rate of an arbitrary busy agent denoted by μ is given by $\mu = (\lambda_h\mu_h + \lambda_l\mu_l)/\lambda$. Substituting $\mu = (\lambda_h\mu_h + \lambda_l\mu_l)/\lambda$ in the preceding derivations gives us the desired approximation for the case of unequal service rates. That is, we replace μ in the derivation of the steady-state probabilities, equation (21), and the derivation of the waiting time distributions, equations (24)-(26) and (30), by $(\lambda_h\mu_h + \lambda_l\mu_l)/\lambda$.

5.2 The equilibrium in steady-state of the system

This section lays out a framework for computing the steady-state of the system in equilibrium for different decisions of the call center manager regarding the delay announcements.

A caller who joins the system when there is an idle agent directly enters service and does not receive any delay announcements. However, when all agents are busy, the caller hears a delay announcement upon arrival that contains information about the number of callers who are waiting to be served, i.e. the total number of callers in the high and low priority queues. Callers trust the information provided by the call center manager. Recall that in our framework for computing the new equilibrium, we assume for simplicity that each caller receives a delay announcement upon arrival; and no further announcements are made to her.

The call center manager can choose three levels of granularity for information contained in the announcements: no information on system occupancy, full information on system occupancy and partial information on system occupancy.[14] In the no information case, the manager does not provide any information. In the full information case, the manager informs the callers of the exact number of callers waiting to be served. In the partial information case, the manager does not provide the exact number of callers waiting to be served but provides a range. For example, the manager may announce: "There are fewer than three callers waiting to be served" or "There are more than three callers waiting to be served."

Different announcements for the high and low priority groups may affect their behavior differently. Therefore, the call center manager may choose the announcements for the two priority groups differently. Letting $\mathcal{S} = \{0, 1, 2, ...\}$ denote the state space for the number of callers waiting to be served, for each priority group $\eta = l, h$, the manager chooses the number of announcement messages J_η and a partition A_1^η, A_2^η,.., $A_{J_\eta}^\eta$ of $\mathcal{S}$, where $A_j^\eta = \{l_{j-1}^\eta + 1, l_{j-1}^\eta + 2, ..., l_j^\eta\}$ with $l_0^\eta = -1 < l_1^\eta < l_2^\eta < ... < l_{J_\eta-1}^\eta < l_{J_\eta}^\eta = \infty$. We refer to each such partition $(A_1^\eta,..., A_{J_\eta}^\eta)$ as an announcement message partition. For example, $J_\eta = 1$ and $A_1^\eta = \mathcal{S}$ corresponds to the no information case, whereas $J_\eta = \infty$ and $A_j^\eta = \{j\}$ corresponds to the full information case. To

[14] For brevity, we omit the "on system occupancy" term and use "no information", "partial information" and "full information" terms instead.

facilitate the analysis to follow, given an announcement message partition $(A_1^\eta,...,A_{J_\eta}^\eta)$, we define the corresponding announcement type function $\Delta_\eta(\cdot)$ that maps $\mathcal{S}$ to $\{A_1^\eta,...,A_{J_\eta}^\eta\}$: $\Delta_\eta(n) = j$ if $n \in A_j^\eta$ for $n \geq 0$.

In what follows, we first explain the derivation of callers' abandonment time distributions given their parameters and virtual waiting time distributions. Also note that the preceding analysis (in Section 5.1) derives the (virtual) waiting time distributions from the given abandonment time distributions. Combining these we next derive a set of equations that characterizes the steady-state of the system in equilibrium for different announcement message partitions chosen by the call center manager for the high and low priority callers. This is done by ensuring that the two aforementioned distributions, i.e. the derivation of waiting time distributions from abandonment time distributions and vice versa, yield distributions which are consistent with each other.

Derivation of callers' abandonment time distributions. We derive the (expected) abandonment time distribution $G_n^h(t)$ for a high priority caller who arrives when there are n callers waiting in the two queues. The derivation for the low priority callers is similar and omitted for brevity. We first calculate the high priority caller's anticipation of her waiting time distribution, which depends on the delay announcement message. Then, we use the optimal stopping model to derive her abandonment probabilities and (expected) abandonment time distribution.

This high priority caller receives a type $\Delta_h(n)$ announcement message. That is, she is told that the number of callers who are waiting to be served is an element of $A^h_{\Delta_h(n)}$. Therefore, to calculate her anticipated waiting time distribution given the message $\Delta_h(n)$, denoted by $F^h(t; A^h_{\Delta_h(n)})$, we take the expectation over the set $A^h_{\Delta_h(n)}$:

$$F^h(t; A^h_{\Delta_h(n)}) = \frac{\sum_{k \in A^h_{\Delta_h(n)}} \gamma_{s+k} F_k^h(t)}{\sum_{k \in A^h_{\Delta_h(n)}} \gamma_{s+k}}, \tag{32}$$

where γ_{s+k} is the steady-state probability of having $s+k$ callers in the system (s callers in service and k callers in the queues) and $F_k^h(t)$ is the virtual waiting time distribution for callers who find k callers in the queue and all agents busy upon arrival. Note that $F^h(t; A^h_{\Delta_h(n)})$ is different from the high priority caller's actual waiting time distribution $F_n^h(t)$ except in the full information case.

The high priority caller makes her abandonment decisions based on the optimal stopping model outlined in Section 4. Suppose that her reward and cost parameters are r and c. Moreover, let $\Theta^h = (m_r^h, \sigma_r^h, m_c^h, \sigma_c^h)$ denote the structural parameters of the high priority callers. The high priority caller receives an announcement only once at $t = 0$. Consequently, her announcement history is $\mathcal{H}_0 = (\Delta_h(n))$. The caller's abandonment behavior is affected by her anticipated probability of receiving service. Since the caller believes that her waiting time distribution is $F^h(t; A^h_{\Delta_h(n)})$, her anticipated probability of receiving service corresponds to the hazard rate of $F^h(t; A^h_{\Delta_h(n)})$. We denote the hazard rate of $F^h(t; A^h_{\Delta_h(n)})$ by $\pi^h(t; A^h_{\Delta_h(n)})$. Note that this is different from the service probability $\pi_n^h(t)$ which is the actual probability of receiving service (as opposed to the caller's anticipation of it). We use (7)-(8) to calculate callers' abandonment probabilities based on their anticipated probability of receiving service $\pi^h(t; A^h_{\Delta_h(n)})$ and their reward and cost parameters.

We next derive the caller's abandonment time distribution denoted by $Z_n^h(t;r,c)$ and its expectation denoted by $G_n^h(t)$ using the abandonment probabilities. For $t=0$, we have

$$Z_n^h(0;r,c) = (1-F_n^h(0))P_{i0}^{\Delta_h(n)}(1;r,c). \tag{33}$$

The right hand side of (33) is the product of two terms: the probability of not receiving service, and the abandonment probability at $t=0$. For $t \geq 1$, the abandonment time distribution of caller i is given by

$$Z_n^h(t;r,c) = Z_n^h(t-1;r,c) + (1-Z_n^h(t-1;r,c))(1-\pi_n^h(t))P_{it}^{\Delta_h(n)}(1;r,c), \tag{34}$$

where $\pi_n^h(t)$ is the hazard rate of $F_n^h(t)$. The first term on the right hand side of (34) is the probability of abandoning at or before period $t-1$. The second term is the probability of not abandoning before period t, not receiving service at time t, but deciding to abandon in period t. Using (17), (33) and (34), we have

$$\begin{aligned} G_n^h(0) &= (1-F_n^h(0)) \int\int P_{i0}^{\Delta_h(n)}(1;r,c)\phi(y_1)\phi(y_2)dy_1dy_2, \\ G_n^h(t) &= G_n^h(t-1) + (1-\pi_n^h(t)) \int\int (1-Z_n^h(t-1;r,c))P_{it}^{\Delta_h(n)}(1;r,c)\phi(y_1)\phi(y_2)dy_1dy_2, \end{aligned} \tag{35}$$

where $r = \exp(m_r^\eta + \sigma_r^\eta y_1)$, $c = \exp(m_c^\eta + \sigma_c^\eta y_2)$ and y_1, y_2 are i.i.d. standard normal random variables as in (1).

Finding the equilibrium in steady-state of the system. We wish to determine F_n^η, G_n^η and Z_n^η ($\eta \in \{h,l\}$) in equilibrium. Suppose that T is the maximum waiting time of the callers. Also, let T_j^η be the maximum waiting time of the callers of class η who heard message j. Note that $T = \max_{\eta,j} T_j^\eta$.[15] For simplicity, we express the waiting and abandonment time distributions as $(T+1)$-vectors: for $\eta = h,l$, $F_n^\eta = [F_{n,t}^\eta]_{t=0}^T$, $G_n^\eta = [G_{n,t}^\eta]_{t=0}^T$ where $F_{n,t}^\eta = F_n^\eta(t)$ and $G_{n,t}^\eta = G_n^\eta(t)$.

Then the following four sets of equations characterize the steady-state of the system in equilibrium:

- Callers' anticipated waiting time distributions:

$$j \in \{1,...,J_\eta\}, 0 \leq t \leq T : F^\eta(t; A_{\Delta_\eta(n)}^\eta) = \frac{\sum_{k \in A_{\Delta_\eta(n)}^\eta} \gamma_{s+k} F_{k,t}^\eta}{\sum_{k \in A_{\Delta_\eta(n)}^\eta} \gamma_{s+k}}, \tag{36}$$

- Derivation of callers' abandonment time distributions using the optimal stopping model: we first derive the abandonment time distribution for a caller, say caller i, with parameters r and c using equations (33) and (34):

$$\begin{aligned} Z_n^\eta(0;r,c) &= (1-F_{n,0}^\eta)P_{i0}^{\Delta_\eta(n)}(1;r,c), \\ t \geq 1 : Z_n^\eta(t;r,c) &= Z_n^\eta(t-1;r,c) + (1-Z_n^\eta(t-1;r,c))(1-\pi_n^\eta(t))P_{it}^{\Delta_\eta(n)}(1;r,c), \end{aligned} \tag{37}$$

Then, we derive the expected abandonment time distributions using equation (35), where the distribution of callers' parameters for $\eta = h,l$ are given by (1):

[15] To compute T_j^η, we first set $T_j^\eta = T$ for some large T and find the system equilibrium. Then we update T_j^η for all j, η based on the resulting waiting time distributions. In our numerical experiments, the values T_j^η converge rapidly.

$$
\begin{aligned}
G^{\eta}_{n,0} &= (1-F^{\eta}_{n,0})\int\int P^{\Delta_\eta(n)}_{i0}(1;r^{\eta}_i,c^{\eta}_i)\ \phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i},\\
T_{\Delta_\eta(n)} \geq t \geq 1 :G^{\eta}_{n,t} &= G^{\eta}_{n,t-1} + (1-\pi^{\eta}_n(t))\int\int (1-Z^{\eta}_n(t-1;r^{\eta}_i,c^{\eta}_i))\\
&\qquad \times P^{\Delta_\eta(n)}_{it}(1;r^{\eta}_i,c^{\eta}_i)\ \phi(y_{1i})\phi(y_{2i})dy_{1i}dy_{2i},
\end{aligned}
\tag{38}
$$

To derive the abandonment probabilities in equations (37) and (38) for type j announcement ($\Delta_\eta(n)=j$ for $n\in A^{\eta}_j$), we use equation (7), where the integrated value functions are given by (8):

$$
t\geq 0 :P^{j}_{it}(1;r^{\eta}_i,c^{\eta}_i) = \frac{1}{1+\exp\Big(-c^{\eta}_i+\pi^{\eta}(t;A^{\eta}_j)r^{\eta}_i+(1-\pi^{\eta}(t;A^{\eta}_j))V_j(t,r^{\eta}_i,c^{\eta}_i)\Big)},
\tag{39}
$$

$$
\begin{aligned}
\pi^{\eta}(t;A^{\eta}_j) &= \frac{F^{\eta}(t+1;A^{\eta}_j)-F^{\eta}(t;A^{\eta}_j)}{1-F^{\eta}(t;A^{\eta}_j)},\\
t\leq T_j-1 :V_j(t,r^{\eta}_i,c^{\eta}_i) &= \log\Big(1+\exp(-c^{\eta}_i+\pi^{\eta}(t+1;A^{\eta}_j)r^{\eta}_i\\
&\qquad +(1-\pi^{\eta}(t+1;A^{\eta}_j))V_j(t+1,r^{\eta}_i,c^{\eta}_i))\Big),\\
V_j(T_j,r^{\eta}_i,c^{\eta}_i) &= 0,
\end{aligned}
\tag{40}
$$

where $V_j(t,r^{\eta}_i,c^{\eta}_i)$ is the value function for a caller who heard message j.

- Derivation of the steady-state probabilities using the Markovian approximation:

$$
\begin{aligned}
n\geq 1 :\delta_n &= \sum_{i_h=1}^{n_h} H^{h}_{m_h(i_h,n)}(\frac{i_h}{\lambda_h}) + \sum_{i_l=1}^{n_l} H^{l}_{m_l(i_l,n)}(\frac{i_l}{\lambda_l}),\\
k\leq s :\mu_k &= k\mu,\ k\geq s+1 :\mu_k = s\mu+\delta_{k-s},\ k\geq 1 :\gamma_k = \frac{\lambda^k}{\prod_{q=1}^{k}\mu_l}\gamma_0,\ \sum_{k=0}^{K}\gamma_k = 1.
\end{aligned}
\tag{41}
$$

- Derivation of the waiting time distributions using the Markovian approximation:

$$
\begin{aligned}
n\geq 0, 0\leq t\leq T :F^{h}_{n,t} &= \mathcal{L}^{-1}\left\{\frac{1}{z}\prod_{q=1}^{n-[\lambda_l\tau]+1}\frac{x^{h}_q(n)}{z+x^{h}_q(n)}\right\}(t),\\
n\geq 0, 0\leq t\leq T :F^{l}_{n,t} &= \mathcal{L}^{-1}\left\{\frac{1}{z}\prod_{q=1}^{n+[b_h(n)]+1}\frac{x^{l}_q(n)}{z+x^{l}_q(n)}\right\}(t),
\end{aligned}
\tag{42}
$$

where $m_h(i_h,n)$, $m_l(i_l,n)$, $x^{h}_q(n)$ and $x^{l}_q(n)$ are given by equations (18), (19), (26) and (30), respectively. Note that H^{η}_n is computed by (17) from Γ^{η}_n, which is the hazard rate of, and hence, computed from Z^{η}_n. In addition, we have $\mu=(\lambda_h\mu_h+\lambda_l\mu_l)/\lambda$. In (41), K denotes the maximum number of callers allowed to enter the system. We choose K such that the blocking probability is negligible.

Given the announcement partitions, the call center parameters, and the set of structural parameters for the high and low priority callers, we solve the above sets of equations simultaneously to find the system equilibrium.[16] This ensures that the waiting time distributions ($F^{\eta}_{n,t},\ \eta\in\{h,l\}$)

[16]To solve this system of equations, we use the KNITRO solver (Byrd et al. (2006)) with AMPL interface. We solve an optimization problem in which the objective function is 0 (or, a constant that does not depend on the variables) and the

that are used to calculate the anticipated waiting time distributions by the first set of equations (Equation (36)), match the distributions that are derived using the Markovian approximation by the fourth set of equations (Equation (42)).

6 Counterfactual analysis

In this section, we conduct simulation experiments to assess the impact of changing the announcement partition (i.e. the granularity of the information) and the relative priority of the two classes on the performance of the call center described in Section 3. To compute the new equilibrium of the system under these changes (using the structural parameters estimated in Section 4.2), we rely on the framework presented in Section 5.2. The results presented in this section may not apply to call centers with different settings/operations.

To study a system that represents the call center in the data, we estimate the arrival and service rates from the data.[17] Given that the number of agents in the data varies across days and hours within a day, it is not obvious a priori what number of agents should be used in the analysis. To determine that, we consider the call center performance under different staffing levels and choose the number of agents that results in average waiting times and abandonment rates that are closest to those observed in the data. This leads to a choice of 5 agents. We set the maximum number of callers allowed in the system including the new arrival (K in Equation (41)) to 19, which corresponds to at most 13 callers in the queue at the time of a new arrival. This ensures a negligible blocking probability (in the order of 10^{-6}).

We conduct two sets of experiments: In the first set of experiments, we compute the equilibrium of the system under the current priority policy for four announcement partitions given below:

- Announcement partition with one subset: No information is provided to the high and low priority callers. This case can be thought of as announcing the message that the queue length is less than or equal to 13, i.e. $A_1^\eta = \{0, ..., 13\}$ for $\eta \in \{h, l\}$, which, of course, has no valuable informational content for the callers.
- Announcement partition with two subsets: The announcement for the high and low priority callers is one of two possible messages. Specifically, we set $A_1^\eta = \{0, ..., 9\}$ and $A_2^\eta = \{10, ..., 13\}$ for $\eta \in \{h, l\}$. This particular choice is made because the aggregate performance metrics reveal that callers arriving when the queue length is 10 or larger suffer from long waits. In contrast, when the queue length is 9 or shorter, the corresponding waiting times are moderate to low.
- Announcement partition with three subsets: The announcement for the high and low priority callers is one of three possible messages. We set $A_1^\eta = \{0, ..., 4\}$, $A_2^\eta = \{5, ..., 9\}$ and $A_2^\eta = \{10, ..., 13\}$ for $\eta \in \{h, l\}$, corresponding to low, moderate and high congestion, respectively.
- Announcement partition with fourteen subsets: Full information is provided to the high and low priority callers. That is, we have $A_1^\eta = \{0\}$, $A_2^\eta = \{1\}$,..., $A_{14}^\eta = \{13\}$ for $\eta \in \{h, l\}$.

constraints are the set of equations that defines the equilibrium.

[17] We find the inter-arrival and service times during busy hours (9 a.m. to 4 p.m.) and use their average to find the arrival rate and service rate in our counterfactual analysis.

Table 7 shows the summary statistics of the high and low priority callers for different announcement partitions.[18] In Table 7, P(A), E(W), E(W|A) and E(W|S) denote the probability of abandonment, average waiting time, average waiting time for the abandoned calls and average waiting time for the served calls, respectively. As can be seen in Table 7, the impact of providing more granular information on most of the aggregate performance metrics is not significant. The only metric that is significantly impacted is E(W|A), which decreases as the information provided becomes more granular.

Announcement	High priority				Low priority			
partition with	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
One subset (No inf.)	10.20%	49.99	48.85	60.06	27.47%	57.40	55.69	61.93
Two subsets	10.18%	49.93	48.91	58.99	27.45%	57.48	55.97	61.47
Three subsets	10.07%	48.48	48.03	52.49	26.66%	56.91	56.77	57.31
Fourteen subsets (Full inf.)	9.98%	47.37	47.09	49.84	26.10%	54.46	54.83	53.42

Table 7: Summary statistics for different announcement partitions under the current policy of the call center.

To understand the impact of delay announcements further, consider Figures 6 and 7, which show the probability of abandonment and the average waiting time of the abandoned calls for different announcement partitions, respectively. As can be seen in Figures 6 and 7, providing delay information helps callers make better decisions in the sense that callers who receive information that the queue length is long abandon more and leave the system sooner compared to the case with no information. Moreover, callers who receive information that the queue length is short abandon less and stay longer in the system. The net effect on the probability of abandonment P(A) is negligible; however, it is significant on the average waiting time of the abandoned calls E(W|A).

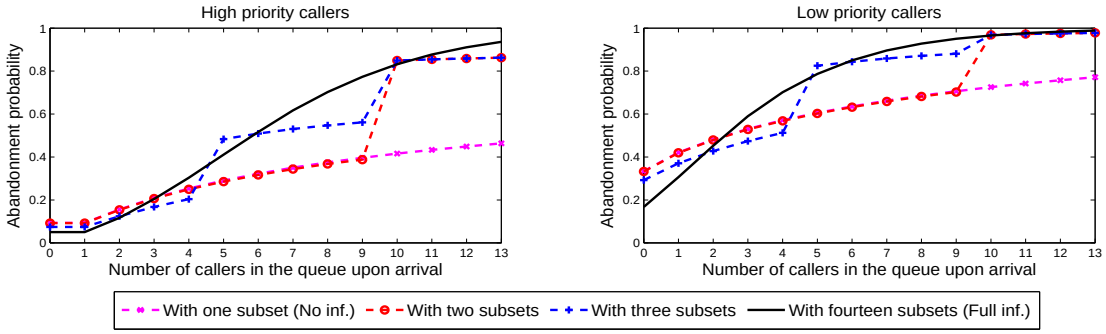

Figure 6: The abandonment probability of the callers depending on the number of callers in the queue upon arrival for different announcement partitions under the current policy of the call center.

Figures 6 and 7 also show that for a given announcement partition heard by a caller, the change in the callers' behavior is "continuous" in the system state, i.e. queue length, as that varies within the particular announcement partition. In contrast, if callers arrive in states which belong to different sets of the announcement partition, they receive different information and the change in their behavior is not continuous in the queue length. This is manifested by a "jump" in the graphs for

[18] Because the abandonment rates of the high and low priority queues are high, when finding the equilibrium in steady-state of the system, we adjust the solution of the state space collapse approximation appropriately using the procedure explained in the online Appendix D. We also validate the accuracy of our equilibrium computation method. See the online Appendix F for the details.

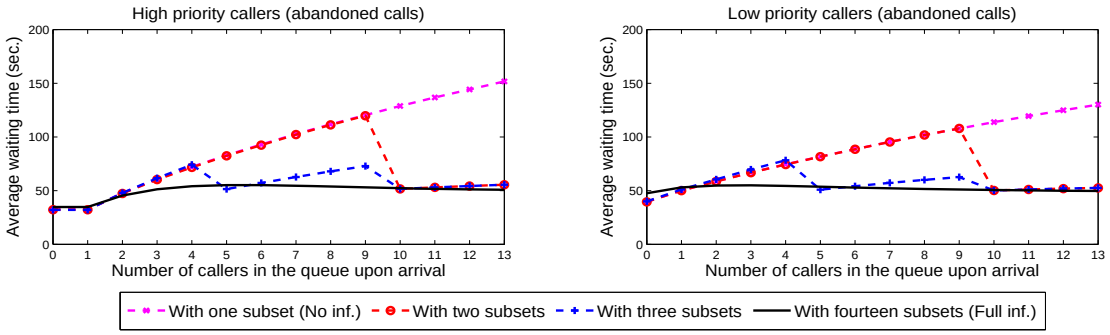

Figure 7: The average waiting time of the abandoned calls depending on the number of callers in the queue upon arrival for different announcement partitions under the current policy of the call center.

the abandonment probability and average waiting time.

In the second set of experiments, we investigate the impact of providing delay information on the system under priority policies different from that found in the data. We consider the following priority policies: First-come-first-served policy (FCFS), a strict (and non-preemptive) priority policy, which gives strict priority to the "high priority" customers, and the reversed strict priority policy, which gives strict priority to the "low priority" customers. Our framework is flexible enough to incorporate these policies by setting τ, i.e. the time parameter by which callers in a particular class are advanced in a queue, sufficiently large. (Setting $\tau = 1800$ in the current policy closely approximates the strict priority rules.)

Table 8 shows the summary statistics of the system under different priority policies for the no information and full information cases. Comparing the results for the no information and full information cases leads to three insights. First, providing delay information does not impact the callers who receive very good service quality. As can be seen in Table 8, the summary statistics of the high priority callers under the strict priority policy and the summary statistics of the low priority callers under the reversed strict priority policy are not affected significantly by the delay information. In these cases, irrespective of the system state, callers receive very good service quality. Therefore, providing delay information does not affect callers' anticipation of the service quality much and, consequently, does not affect their behavior significantly.

Priority policy	High priority				Low priority			
(Announcement partition)	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Reversed priority (No info.)	21.83%	67.08	66.84	67.96	3.80%	24.68	24.23	36.18
Reversed priority (Full info.)	19.38%	59.85	61.50	53.00	3.76%	24.41	23.95	36.18
FCFS (No info.)	12.92%	55.85	54.92	62.09	18.64%	50.04	47.02	63.25
FCFS (Full info.)	12.89%	52.91	53.12	51.54	18.03%	46.74	45.59	51.93
Strict priority (No info.)	3.73%	29.03	28.61	39.75	46.11%	47.08	34.10	62.24
Strict priority (Full info.)	3.70%	28.34	27.91	39.41	39.30%	45.36	40.32	53.16

Table 8: Summary statistics for the no information and full information cases under different priority policies.

Second, providing delay information decreases the average waiting time of the abandoned calls when the service quality is either mediocre or poor. This is the case for the high priority callers under the reversed strict priority policy, the low priority callers under the strict priority policy and

both of them under the FCFS policy. For these cases, the callers who hear that the queue length is long, and are not patient enough to wait for entering service, leave the system sooner, which results in a lower average waiting time for the abandoned calls.

Finally, providing delay information affects the impatient callers more than the patient ones. Under the reversed strict priority policy, the abandonment rate of the high priority callers is not affected significantly from the delay information. In contrast, the change in the abandonment rate of the low priority callers under the strict priority policy is more significant. Recall from the estimation results in Section 4.2 that the low priority callers are less patient than the high priority callers. Therefore, they are more sensitive to the delay information and, consequently, their behavior changes more significantly. To clarify this point, we illustrate in Figure 8 the abandonment probability of the high and low priority callers under the reversed strict priority policy and the strict priority policy, respectively. Figure 8 shows that under delay information the abandonment probability of the low priority callers increases faster as the system gets congested. As can be seen in Figure 8, for the no information case the graphs for the high and low priority callers have roughly the same slope. However, for the full information case, the slope of the graph of the low priority callers is steeper initially than that of the high priority callers. This confirms the fact that that the low priority callers are more sensitive (and more responsive) to delay information as the congestion in the system builds.

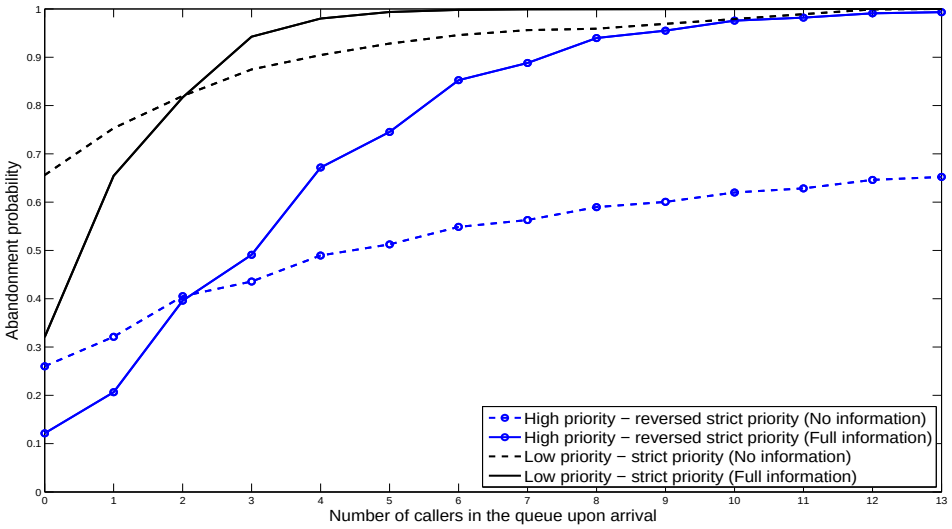

Figure 8: The abandonment probability for the no information and full information cases for the high priority callers under the reversed strict priority policy and the low priority callers under the strict priority policy.

7 Concluding Remarks

This paper introduces a modeling framework to study the impact of delay announcements in a multi-class call center. A model for callers' abandonment behavior under delay announcement is developed where callers' decisions are endogenous to their anticipation of the chances of receiving service and their parameters such as the reward from receiving service and the waiting cost. Callers' reward and cost parameters are estimated from the call center data using maximum likelihood es-

timation. Callers anticipate the service probabilities based on the information contained in the announcements. This gives the call center manager an instrument to affect callers' anticipation and as a result their abandonment behavior by changing the information in the announcements.

A change in callers' behavior affects the evolution of the system. Analyzing this effect requires an extensive queueing analysis which we perform in our setting using a Markovian approximation. In particular, using callers' positions in the queue and the type of announcement they received, we compute their probability of abandonment. Then, we model the system as a birth-and-death process with state dependent abandonment rates, which facilitates the derivation of steady-state probabilities and waiting time distributions.

To illustrate the impact of delay information, a comparison is made between different levels of information granularity under different priority policies. For each case, we solve a set of non-linear equations to find the equilibrium in steady-state of the system. We show that callers indeed react to delay announcements and that providing delay information helps them make better decisions. We find that providing delay information does not significantly affect the aggregate statistics such as the average waiting times and abandonment rates. However, when callers suffer from poor service, providing information induces the callers who are impatient and would eventually abandon, to leave the system sooner. This results in a lower average waiting time for the abandoned calls, and presumably, a higher customer satisfaction. We also find that impatient callers are more sensitive to delay information in the sense that their behavior may change more prominently than that of the patient callers.

In future work, it would be interesting to explore the application of the proposed framework to other service systems where delay information is shared with customers, e.g. health care services. Investigating the robustness of the results to different types of delay announcements, and in different call center operating environments is another direction to pursue. For example call centers with more than two classes and with dynamic priority policies can be analyzed.

Lastly, this paper does not prove the existence and uniqueness of the equilibrium. These questions are tackled rigorously in Ata et al. (2015) and Ata and Peng (2015) for the case with no delay announcement. It would be interesting to investigate these questions under delay announcements too.

Acknowledgements

We are grateful to the Service Enterprise Engineering (SEE) lab at the Technion (http://ie.technion.ac.il/Labs/Serveng/) for providing us with the data. We are especially thankful to Prof. Avi Mandelbaum and Dr. Valery Trofimov.

References

Afèche, P. and V. Sarhangian (2015). Rational abandonment from priority queues: Equilibrium strategy and pricing implications. *Available at SSRN 2679328*.

Aksin, Z., B. Ata, S. Emadi, and C. Su (2013). Structural estimation of callers delay sensitivity in call centers. *Management Science 59*(12), 2727–2746.

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.

Ata, B., P. Glynn, and X. Peng (2015). An equilibrium analysis of a discrete–time M/M/s queue with endogenous abandonments. *Working paper, Booth School of Business, University of Chicago*.

Ata, B. and X. Peng (2015). An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Working paper, Booth School of Business, University of Chicago*.

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

Bramson, M. (1998). State space collapse with applications to heavy traffic limits for multiclass queueing networks. *Queueing Systems 30*, 89–148.

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.

Emadi, S. (2013). *Estimation and Analysis of Callers' Behavior in Call Centers*. Ph. D. thesis, Northwestern University.

Feigin, P. (2006). Analysis of customer patience in a bank call center. *Working paper, Technion Israel Institute of Technology*.

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.

Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous customer populations. In e. W. Fleming, P. L. Lions (Ed.), *Stochastic Differential Systems, Stochastic Control Theory and Their Applications*, pp. 147–186. New York: Vol. 10. The IMA Volumes in Mathematics and Its Applications, Springer-Verlag.

Hassin, R. (1986). Consumer information in markets with random product quality: the case of queues and balking. *Econometrica 54*, 1185–1195.

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.

Hendel, I. and A. Nevo (2013). Intertemporal price discrimination in storable goods markets. *American Economic Review 103*(7), 2722–2751.

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.

Hosmer, D. W., S. May, and S. Lemeshow (2008). *Applied survival analysis*. Wiley-Interscience.

Ibrahim, R. and W. Whitt (2009). Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management 11*(2), 397–415.

Jouini, O., Y. Dallery, and Z. Aksin (2009). Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics 120*, 389–399.

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.

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

Li, J., N. Granados, and S. Netessine (2014). Are consumers strategic? structural estimation from the air-travel industry. *Management Science 60*, 2114– 2137.

Lu, Y., A. Musalem, M. Olivares, and A. Schilkrut (2013). Measuring the effect of queues on customer purchases. *Management Science 59*(8), 1743–1763.

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

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.

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.

Stute, W. and J.-L. Wang (1994). The jackknife estimate of a Kaplan–Meier integral. *Biometrika 81*(3), 602–606.

Vulcano, G., G. van Ryzin, and W. Chaar (2010). Choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management 12*(3), 371–392.

Ward, A. R. (2012). Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science 16*(1), 1–14.

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

Whitt, W. (1999b). Predicting queueing delays. *Management Science 45*, 870–888.

Whitt, W. (2005). Engineering solution of basic call-center model. *Management Science 51*, 221–235.

Whitt, W. and J. Abate (1995). Numerical inversion of laplace transforms of probability distributions. *ORSA Journal on Computing 7*, 36–43.

Xu, S., L. Gao, and J. Ou (2007). Service performance analysis and improvement for a ticket queue with balking customers. *Management Science 53*(6), 971–990.

Yu, Q., G. Allon, and A. Bassamboo (2015). How do delay announcements shape customer behavior? an empirical study. *Management Science, Forthcoming*.

Online Appendix for "Impact of Delay Announcements in Call Centers: An Empirical Approach"

A Illustrating the Impact of Delay Announcements in the Data

In this section, we illustrate the impact of delay announcements on callers' abandonment behavior observed in the data using the Cox regression analysis (Hosmer et al. (2008)). The Cox regression analysis can be used to find the impact of a set of covariates (independent variables) on the hazard rate of customers' survival time distribution. In this paper, we are interested in customers' abandonment behavior. Consequently, we consider customers' abandonment times as their survival times.

Suppose that $h_i(t)$ is the hazard rate of the abandonment time distribution for caller i with the set of independent variables X_i. The vector X_i includes variables that may impact the caller's abandonment behavior, e.g. the announcement messages. In the Cox regression analysis we assume that

$$h_i(t) = h_0(t) \ \exp(X_i \beta^T), \tag{43}$$

where $h_0(t)$ is the baseline hazard function that can be any function of t as long as $h_0(t) > 0$. The function $h_0(t)$ does not have to be specified. The vector β is the vector of coefficients of the independent variables, which captures the impact of the independent variables on callers' abandonment hazard rates. Note that the exponent term in (43) does not involve a time variable. Consequently, the ratio of hazard functions of two callers does not depend on time and $h_0(t)$. To be more specific, the ratio of hazard function of callers i and j is given by

$$\frac{h_i(t)}{h_j(t)} = \exp((X_i - X_j)\beta^T). \tag{44}$$

As can be seen in (44) in the Cox regression analysis it is assumed that the ratio of hazard functions of two callers only depend on the difference between their independent variables. This assumption is called the proportional hazard rate assumption. Next, we explain the set of independent variables and present the regression results.

Recall that each announcement message has two parts. The first part is the relative position in the queue, which is an integer value between 1 and 3. The second part is the waiting time of the longest waiting caller. As explained in Section 3, we discretize the values for the second part to six intervals: [0,10], [11,30], [31,90], [91,210], [211,480] and [481,2700]. Consequently, we have the total of 18 announcement messages. We order the announcement messages lexicographically based on the first part and the second part of the message; and label them with indices from 1 to 18. For example, the index for the announcement message with first part=3 and second part=[0,10] is 13. Note that an announcement message with a higher index indicates a longer delay.

Denote by $a_1 = \left[a_1^i\right]$ the vector of the indices of the announcement messages callers heard upon

arrival, where i indexes different callers. Moreover, for $k \geq 2$ denote by $a_k = \left[a_k^i\right]$ the vector of the differences between the indices of the announcement messages callers receive at $60k$ seconds and $60(k-1)$ seconds. As explained in Section 3 we focus on calls with waiting times less than 600 seconds, consequently, the maximum number of announcement messages a caller may receive is ten and $k \leq 10$. Note that if a_k^i is positive then caller i's expected delay announced at $60k$ seconds is higher than the delay announced at $60(k-1)$ seconds, which indicates a deteriorating condition. This situation can occur for the low priority callers who are passed by some high priority callers because of 90 seconds advancement in the priority policy. Let L denote the dummy variable for the low priority group, A_{rate} denote the arrival rate at the contact time, and $t_h;\ 9 \leq h \leq 14$ denote the dummy variable for time of the day. The variable t_h is equal to 1 if the caller contacts between $h:00$ and $h+1:00$. The set of independent variables in the Cox regression is: $X = (a_l;\ 1 \leq l \leq k$, L, A_{rate} and $t_h;\ 9 \leq h \leq 14)$.

We first perform a standard Cox regression and observe the main independent variables $(a_l;\ 1 \leq l \leq k)$ do not satisfy the proportional hazard rate assumption that is made in the Cox regression analysis, given the time dependent nature of these covariates; see Hosmer et al. (2008), Chapter 6.3, page 205 for details of the proportional hazard test. We verify that these variables are indeed time varying by including interaction terms with time for each one of these variables that fail the proportional hazard test while being significant, and performing the Cox analysis with these interaction terms. For the main independent variables that are significant, their interactions with time are also significant confirming that the hazard rates are not proportional. We do not report these results for brevity here, however note that the results in this regression qualitatively parallel our final results reported below.

Based on these observations we consider an analysis where we split the data and perform a separate Cox regression on each split. More specifically, to find the impact of the announcement messages on callers' abandonment behavior between $60(k-1)$ seconds and $60k$ seconds from arrival, we focus on callers that hear the k^{th} announcement message at $60(k-1)$ seconds but either abandon or enter service before hearing the next announcement message at $60k$ seconds. For this group of callers, we run a Cox regression on $a_l;\ 1 \leq l \leq k$, L, A_{rate} and $t_h;\ 9 \leq h \leq 14$. All of these regressions satisfy the proportional hazards assumption and show significant impact of announcements (again not reported for brevity), paralleling our main results below. These results suggest that doing a stratified Cox regression (see Hosmer et al. (2008) Chapter 7) where we consider each split described above as a stratum in our analysis is appropriate and furthermore this will enable us to use all of the data in one regression.

To do the stratified Cox regression we define G_{strata} as our strata variable, which is equal to k for callers who hear the k^{th} announcement message at $60(k-1)$ seconds but either abandon or enter service before hearing the next announcement message at $60k$ seconds. In the stratified Cox regression, we allow different strata to have different baseline hazard rates. In our context, this corresponds to estimating a different baseline hazard for callers who hear one additional announcement as we move from $G_{strata} = 1$ to $G_{strata} = 10$ (recall that the maximum number of announcement

messages a caller may receive is ten). Let $h^k(t)$ denote the hazard rate of the abandonment time distribution for callers whose G_{strata} is equal to k (i.e. are in the k^{th} stratum). The function $h^k(t)$ has the following form:

$$h^k(t) = h_{0k}(t)\ \exp(X\beta^T), \tag{45}$$

where $h_{0k}(t)$ is the baseline hazard function for callers in the k^{th} stratum. As can be seen in (45), we assume that callers in different strata have different baseline hazard function, but the same coefficients for the independent variables. Table 9 reports the results of this analysis.

Table 9: The results of the stratified Cox regression.

Variable	**Coefficient**	**(Std. Err.)**
a_1	0.1758 ***	(0.0023)
a_2	0.0526 ***	(0.0062)
a_3	0.0324 ***	(0.0118)
a_4	0.0371 **	(0.0198)
a_5	0.0458	(0.0316)
a_6	0.0491	(0.0620)
a_7	-0.0741	(0.0745)
a_8	0.1703 **	(0.0978)
a_9	0.0901	(0.3227)
a_{10}	0.7483 ***	(0.1917)
L	0.4889 ***	(0.0172)
A_{rate}	0.0137 ***	(0.0005)
t_9	0.0622 ***	(0.0309)
t_{10}	0.0327	(0.0280)
t_{11}	0.0925 ***	(0.0304)
t_{12}	0.0698 ***	(0.0331)
t_{13}	-0.0455	(0.0321)
t_{14}	-0.0083	(0.0311)

[a] *** and ** denote statistically significant at 0.05 and 0.10, respectively.
[b] Stratified by G_{strata}

As can be seen in Table 9, the coefficient for a_1 (the index of the first announcement) is significant and positive. This shows that callers who hear a longer delay announced upon arrival abandon earlier. Similarly, the coefficients for a_2, a_3, a_4, a_8 and a_{10} are positive and significant. Note that these variables capture the incremental change in the announcement index, which are positive if the caller is informed of a deteriorating situation. Consequently, the regression results show that callers who see a deteriorating delay condition abandon earlier. Moreover, the coefficient for L and A_{rate} are positive and significant, which shows that the low priority callers, and callers who arrive when the system is more congested abandon earlier. We qualitatively obtain the same significance and directional results in all models considered (Cox regression with time interaction terms and Cox regression separately run on each split (stratum)).

B Proofs

Proof of Proposition 1. Suppose that caller i's announcement history is $\mathcal{H}_k$. We first derive the formula for choice probabilities $P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i)$, and then the recursive formula for the integrated value function $V_{\mathcal{H}_k}(t, r_i, c_i)$ for $t \in \{kL, ..., (k+1)L-1\}$.

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

$$\begin{aligned} u_{\mathcal{H}_k}(t, r_i, c_i, \varepsilon_{it}(d_{it}), d_{it}) &= v_{\mathcal{H}_k}(t, r_i, c_i, d_{it}) + \varepsilon_{it}(d_{it}) \\ &> v_{\mathcal{H}_k}(t, r_i, c_i, 1-d_{it}) + \varepsilon_{it}(1-d_{it}) \\ &= u_{\mathcal{H}_k}(t, r_i, c_i, \varepsilon_{it}(1-d_{it}), 1-d_{it}). \end{aligned}$$

Therefore, we have

$$\begin{aligned} P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i) = \int\int \mathbb{I}_{\{\varepsilon_{it}(d_{it}) - \varepsilon_{it}(1-d_{it}) > v_{\mathcal{H}_k}(t, r_i, c_i, 1-d_{it}) - v_{\mathcal{H}_k}(t, r_i, c_i, d_{it})\}} \\ \times g(\varepsilon_{it}(0)) g(\varepsilon_{it}(1)) d\varepsilon_{it}(0) d\varepsilon_{it}(1). \end{aligned} \tag{46}$$

We assume that the idiosyncratic shocks have i.i.d type-I extreme value distribution with scale parameter 1 and location parameter $\alpha \in \mathbb{R}$ with the probability density function $\exp(-(\varepsilon(d) - \alpha)) \exp(-\exp(-(\varepsilon(d) - \alpha))$ for $d = 0, 1$. As will be seen below, for technical convenience we will set $\alpha = -\gamma$, where γ is Euler's constant. From (46), by Section 5.2 in Ben-Akiva and Lerman (1985), and the fact that $v_{\mathcal{H}_k}(t, r_i, c_i, 1) = 0$, we obtain the formula for the choice probability as follows

$$\begin{aligned} P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i) &= \frac{\exp\big(v_{\mathcal{H}_k}(t, r_i, c_i, d_{it})\big)}{\exp\big(v_{\mathcal{H}_k}(t, r_i, c_i, 1)\big) + \exp\big(v_{\mathcal{H}_k}(t, r_i, c_i, 0)\big)} \\ &= \frac{\exp\big(v_{\mathcal{H}_k}(t, r_i, c_i, d_{it})\big)}{1 + \exp\big(v_{\mathcal{H}_k}(t, r_i, c_i, 0)\big)}. \end{aligned} \tag{47}$$

We next derive the integrated value function for two cases: $t \neq (k+1)L-1$ and $t = (k+1)L-1$. As the first case, suppose that $t \neq (k+1)L - 1$. Recall from (4) that the integrated value function is given by

$$V_{\mathcal{H}_k}(t, r_i, c_i) = \mathbb{E}\left[\max_{d \in \{0,1\}} u_{\mathcal{H}_k}(t+1, r_i, c_i, \varepsilon_{i(t+1)}(d), d)\right],$$

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 $\alpha + \log(e^{v_1} + e^{v_0})$, where $v_q = v_{\mathcal{H}_k}(t +$

$1, r_i, c_i, q)$, $q = 0,\ 1$. Therefore, we have

$$V_{\mathcal{H}_k}(t, r_i, c_i) = \mathbb{E}\left[\max_{d\in\{0,1\}} u_{\mathcal{H}_k}(t+1, r_i, c_i, \varepsilon_{i(t+1)}(d), d)\right]$$
$$= \alpha + \log(e^{v_1} + e^{v_0}) + \gamma. \tag{48}$$

For technical convenience, we assume that the location parameter for the distribution of the idiosyncratic shocks α is equal to $-\gamma$. Then, by definitions of v_1 and v_0 and (48), it follows that

$$V_{\mathcal{H}_k}(t, r_i, c_i) = \log\Big(1 + \exp\big(v_{\mathcal{H}_k}(t+1, r_i, c_i, 0)\big)\Big),\ t \neq (k+1)L - 1. \tag{49}$$

As the second case, suppose that $t = (k+1)L - 1$. Note that the caller receives a new announcement in the next period. Caller's announcement history in the next period will be $\mathcal{H}_{k+1}$ with probability $a_{\mathcal{H}_{k+1}}/a_{\mathcal{H}_k}$, where $\mathcal{H}_{k+1}$ is concatenation of $\mathcal{H}_k$ with $j \in \{1, ..., J\}$. Consequently, the integrated value function, which is the expected maximum utility in the next period, is given by

$$V_{\mathcal{H}_k}(t, r_i, c_i) = \sum_{\mathcal{H}_{k+1}|\mathcal{H}_k} \frac{a_{\mathcal{H}_{k+1}}}{a_{\mathcal{H}_k}} \mathbb{E}\left[\max_{d\in\{0,1\}} u_{\mathcal{H}_{k+1}}(t+1, r_i, c_i, \varepsilon_{i(t+1)}(d), d)\right]. \tag{50}$$

From (48) and (49), for $t + 1 = (k+1)L$, we have

$$\mathbb{E}\left[\max_{d\in\{0,1\}} u_{\mathcal{H}_{k+1}}(t+1, r_i, c_i, \varepsilon_{i(t+1)}(d), d)\right] = \log\Big(1 + \exp\big(v_{\mathcal{H}_{k+1}}(t+1, r_i, c_i, 0)\big)\Big). \tag{51}$$

By substituting (51) in (50), we obtain

$$V_{\mathcal{H}_k}(t, r_i, c_i) = \sum_{\mathcal{H}_{k+1}|\mathcal{H}_k} \frac{a_{\mathcal{H}_{k+1}}}{a_{\mathcal{H}_k}} \log\Big(1 + \exp\big(v_{\mathcal{H}_{k+1}}(t+1, r_i, c_i, 0)\big)\Big),\ t = (k+1)L - 1. \tag{52}$$

If $\mathcal{H}_k$ is a terminal history, then by definition all callers enter service or abandon by $T_{\mathcal{H}_k}$. Therefore, no caller enters service after $T_{\mathcal{H}_k}$ which means that the expected future value of waiting is zero at $T_{\mathcal{H}_k}$. Consequently, we assume that $V_{\mathcal{H}_k}(T_{\mathcal{H}_k}, r_i, c_i) = 0$. □

Lemma 1. *The solution for* $b_h(n) = \lambda_h \min((n + b_h(n) + 1)/\lambda, \tau)$ *is given by*

$$b_h(n) = \begin{cases} \lambda_h \tau & \text{if } \ n \geq \lambda_l \tau - 1, \\ \\ \frac{\lambda_h}{\lambda_l}(n+1) & \text{if } n < \lambda_l \tau - 1. \end{cases}$$

Proof. We consider two cases: $1)(n + b_h(n) + 1)/\lambda \geq \tau$ and $2)(n + b_h(n) + 1)/\lambda < \tau$.

Case 1. Suppose that $(n+b_h(n)+1)/\lambda \geq \tau$. Then, $b_h(n) = \lambda_h \min((n+b_h(n)+1)/\lambda, \tau) = \lambda_h \tau$. Consequently, by the assumption for case 1, we should have $(n + \lambda_h \tau + 1)/\lambda \geq \tau$. Thus, we should have $n + 1 \geq \lambda\tau - \lambda_h\tau = \lambda_l\tau$. Hence, if $n \geq \lambda_l\tau - 1$, then $b_h(n) = \lambda_h\tau$.

Case 2. Suppose that $(n + b_h(n) + 1)/\lambda < \tau$. Then, $b_h(n) = \lambda_h \min((n + b_h(n) + 1)/\lambda, \tau) = \lambda_h(n + b_h(n) + 1)/\lambda$. Solving for $b_h(n)$, we obtain $b_h(n) = (n+1)\lambda_h/\lambda_l$. Consequently, by the

assumption for case 2, we should have $(n + (n+1)\lambda_h/\lambda_l + 1)/\lambda < \tau$. It is easy to check that this inequality is equivalent to $n < \lambda_l\tau - 1$. Hence, if $n < \lambda_l\tau - 1$, then $b_h(n) = (n+1)\lambda_h/\lambda_l$. This completes the proof. □

C A Monte-Carlo Experiment

We use the Monte-Carlo experiments to test the capability of our estimation method to identify the true parameters of the callers. To conduct these experiments, we first generate simulated data sets considering certain values for the structural parameters. We refer to these values by true values. Next, we estimate the parameters of the simulated data sets and construct the 95% confidence intervals.

We consider the following true values for the structural parameters: $m_r = 1.85$, $m_c = -2.20$, $\sigma_r = 0.35$ and $\sigma_c = 0.20$, which correspond to the mean of the reward parameter, the standard deviation of the reward parameter, the mean of the cost parameter and the standard deviation of the cost parameter equal to 6.49, 1.33, 1.36 and 3.29, respectively. We use the service probabilities $\pi_{\mathcal{H}_k}(\cdot)$ and the probability of announcement histories $a_{\mathcal{H}_k}$ corresponding to the high priority callers in the data for our simulated data generation procedure. We generate 50 simulated data sets which have the same number of observations as the number of the high priority callers in our data set.

To simulate the abandonment behavior of simulated caller i, we draw y_{1i} and y_{2i} from the standard normal distribution. We find r_i and c_i using the assumed true values for the structural parameters. Given r_i and c_i, we calculate the integrated value functions and abandonment probabilities for all announcement histories and at all periods.

Having the service probabilities $\pi_{\mathcal{H}_k}(t)$, the abandonment probabilities $P_{it}^{\mathcal{H}_k}(d_{it}; r_i, c_i)$ and the probabilities of the announcement histories $a_{\mathcal{H}_k}$ allows us to simulate what announcement caller i receives and whether she enters service or abandons the queue.

Table 10 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	1.850	0.350	–2.200	0.200
Mean (Simulated data)	1.845	0.348	–2.218	0.218
Standard deviation (Simulated data)	0.010	0.015	0.033	0.080
Upper bound of the 95% confidence interval	1.865	0.376	–2.153	0.376
Lower bound of the 95% confidence interval	1.824	0.319	–2.283	0.061

Table 10: Results of the Monte-Carlo experiment.

D Improving the State Space Collapse Approximation for Cases with Large Abandonment Rates

In this appendix, using a series of simulation analyses we show that the state space collapse approximation may have a poor performance for cases with large abandonments. To be more specific, for cases with large abandonments, the individual number of the high and low priority callers (n_h and n_l) given the total number of callers in the queues ($n = n_h + n_l$) that are calculated using (16) may not be accurate. To address the inaccuracy of the state space collapse approximation, we propose an approach to change the solution in (16) by introducing a fudge factor for the amount of advancement of the high priority callers τ. This effectively changes the split of n between n_h and n_l. Although there are many ways to do this, adjusting τ in (16) is perhaps the simplest such approach as it lends itself to a one-dimensional search to fine tune our Markovian approximation. This approach is outlined not only for cases with exogenous abandonment time distributions, but also for cases in our counterfactual analysis where callers' abandonment behavior is endogenous and follows an optimal stopping model.

We performed an extensive simulation study to examine the accuracy of the Markovian approximation outlined in Section 5.1. We simulated the system for different specifications that include different arrival rates, service rates, priority policies and abandonment time distributions. For each specification used in the simulation, we calculated callers' average waiting times and abandonment rates from the simulation and compared them to those obtained from using the Markovian approximation. Due to space limitation, we only report results from two cases described below; results for the other cases are qualitatively similar.

Case 1. A system with two priority classes (high and low), where the high priority callers are advanced by $\tau = 90$ seconds. There are 5 agents. The arrival process is Poisson with rates 55.80 (/hr) and 28.80 (/hr) for the high and low priority groups, respectively. The service times are exponentially distributed with rates 19.44 (/hr) and 16.56 (/hr) for the high and low priority groups, respectively. The abandonment times for the high and low priority callers have exponential distributions with rates 6.84 (/hr) and 15.48 (/hr), respectively.

Case 2. A system with the same setting as Case 1 except that the abandonment times for the high and low priority callers have exponential distributions with rates 0.14 (/hr) and 0.31 (/hr), respectively.

Table 11 shows the comparison of the approximation and simulation results for Cases 1 and 2. Note that because the rates of the abandonment time distributions in Case 2 are smaller than those in Case 1, the abandonment rates (P(A)) of the callers in Case 2 are smaller than those in Case 1. The averages of the relative errors across all statistics in Table 11 for Case 1 and Case 2 are 8.33% and 4.33%, respectively. This shows that the accuracy of our Markovian approximation decreases if callers' abandonment rate increases. Next, we explain the reason for this.

Table 12 shows the comparison of the approximation and simulation results for the expected

Case 1	High priority				Low priority			
	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Approximation results	7.73%	40.68	38.34	68.65	23.20%	53.96	47.84	74.23
Simulation results	9.10%	47.13	44.49	73.54	23.84%	55.27	46.74	82.53
Relative error	15.02%	13.69%	13.83%	6.65%	2.66%	2.36%	2.35%	10.05%
Case 2	High priority				Low priority			
	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Approximation results	1.50%	395.75	395.39	419.21	3.88%	450.88	451.83	427.49
Simulation results	1.60%	406.85	406.18	449.06	4.08%	465.05	465.64	451.00
Relative error	6.06%	2.73%	2.66%	6.65%	4.87%	3.05%	2.97%	5.21%

Table 11: Comparison of the average waiting times and abandonment rates for the approximation and the simulation for Case 1 and Case 2.

number of callers in the system E(N), the expected number of callers in the queues $\mathrm{E}(n_h + n_l)$, and the expected ratio of the high priority queue length and the total queue length given that the queue is not empty $\mathrm{E}(n_h/(n_h+n_l)|n_h+n_l>0)$. As can be seen in Table 12, the approximation errors for

Case 1	E(N)	$\mathrm{E}(n_h+n_l)$	$\mathrm{E}(n_h/(n_h+n_l)\|n_h+n_l>0)$
Approximation results	5.21	1.09	0.50
Simulation results	5.33	1.17	0.62
Relative error	2.16%	6.68%	19.35%
Case 2	E(N)	$\mathrm{E}(n_h+n_l)$	$\mathrm{E}(n_h/(n_h+n_l)\|n_h+n_l>0)$
Approximation results	14.66	9.95	0.62
Simulation results	14.72	10.01	0.63
Relative error	0.43%	0.58%	1.59%

Table 12: Comparison of E(N), $\mathrm{E}(n_h+n_l)$ and $\mathrm{E}(n_h/(n_h+n_l)|n_h+n_l>0)$ resulted from the approximation and the simulation for Case 1 and Case 2.

E(N), $\mathrm{E}(n_h+n_l)$ and $\mathrm{E}(n_h/(n_h+n_l)|n_h+n_l>0)$ are higher for the case with larger abandonment rates (Case 1). In addition, for Case 1 the errors in approximating E(N) and $\mathrm{E}(n_h+n_l)$ are less than 7%, but the error in approximating $\mathrm{E}(n_h/(n_h+n_l)|n_h+n_l>0)$ is around 20%. This suggests that for cases with large abandonments, even though the performance of the approximation in capturing the total number of callers in the queues is acceptable, its performance in capturing the individual number of the high and low priority callers n_h and n_l in the queues can be poor.

Recall that in Section 5.1, we use the state space collapse approximation to find the number of the high and low priority callers (n_h, n_l) given the total number of callers in the queues $(n = n_h+n_l)$ by solving equation (15). However, this approximation may not be suitable for cases with large abandonments. Note that the number of the high and low priority queues has a significant impact on derivation of the waiting time distributions. Consequently, an inaccurate state space collapse approximation may lead to a poor approximation of the waiting times and abandonment rates.

We adjust the calculation in the state space collapse approximation such that $\mathrm{E}(n_h/(n_h+n_l)|n_h+n_l>0)$ in the approximation matches that of the simulation. This ensures that on average the (n_h, n_l) pair resulted from the approximation matches the pair in the simulation. Recall that the solution of the state space collapse approximation is given by (16). We can adjust the solution formula by changing either the arrival rates to the system (λ_h and λ_l), or the amount of the high priority callers' advancement τ. Finding the proper amount of adjustment requires a search

procedure which is easier to conduct on one dimension. Therefore, we propose to adjust the solution formula by substituting the amount of the high priority callers' advancement τ with $\tau + \Delta\tau$, where $\Delta\tau$ is a fudge factor. Given that we only want to adjust the solutions of the state space collapse approximation, we do not change the amount of the high priority callers' advancement in other parts of the Markovian approximation.

Denote by $\widehat{\Delta\tau}$ the fudge factor for which $\mathrm{E}(n_h/(n_h+n_l)|n_h+n_l>0)$ in the simulation and approximation match closely.[19] For the call center that we have its data, policies with $\tau=-1800$ and $\tau=1800$ resemble the reversed strict and strict priority policies, respectively. Consequently, to find $\widehat{\Delta\tau}$, we start from $\widehat{\Delta\tau}=-1800$ and increase it until $\mathrm{E}(n_h/(n_h+n_l)|n_h+n_l>0)$ in the approximation matches that of the simulation closely. Using this procedure, we find the values of $\widehat{\Delta\tau}$ for Case 1 and 2 that are -60 and -90, respectively. Table 13 shows the comparison of the approximation and simulation results for Case 1 and 2 after adjusting the calculation in the state space collapse approximation results. The average of the relative errors across all statistics in Table 13 for Case 1 and Case 2 are 3.96% and 3.25%, respectively, which are less than the averages of the relative errors in Table 11.

Case 1	High priority				Low priority			
	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Approximation results	8.55%	47.82	45.85	68.93	24.22%	54.80	48.33	75.06
Simulation results	9.10%	47.13	44.49	73.54	23.84%	55.27	46.74	82.53
Relative error	6.03%	1.46%	3.05%	6.27%	1.58%	0.85%	3.40%	9.05%
Case 2	High priority				Low priority			
	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Approximation results	1.53%	402.17	401.90	419.98	3.91%	455.19	456.22	430.06
Simulation results	1.60%	406.85	406.18	449.06	4.08%	465.05	465.64	451.00
Relative error	4.53%	1.15%	1.05%	6.48%	3.97%	2.12%	2.02%	4.64%

Table 13: Comparison of the average waiting times and abandonment rates for the approximation and the simulation for Case 1 and Case 2 after adjusting the state space collapse approximation results by setting $\widehat{\Delta\tau}$ at -60 and -90, respectively.

Note that in Cases 1 and 2, callers' abandonment time distributions are exogenously given. However, in our counterfactual analyses in Section 6, we find the system equilibrium, where callers make their abandonment decisions based on an optimal stopping model. To find $\widehat{\Delta\tau}$ for the cases in Section 6, we run a one dimensional search using the following procedure:

Step 0. Initialization: set $\Delta\tau=-1800$.

Step1. State space collapse approximation: for $n\geq 1$, set $n_h=\max(0,[(n-\lambda_l(\tau+\Delta\tau))/(\lambda_l/\lambda_h+1)])$ and $n_l=n-n_h$.

Step 2. Markovian approximation:

Step 2.1. Solve the equilibrium equations given the state space collapse results in the previous step, and that the high priority callers' advancement is τ.

[19] That is, the difference between $\mathrm{E}(n_h/(n_h+n_l)|n_h+n_l>0)$ in the simulation and approximation is less than 0.001.

Step 2.2. Calculate the expected ratio of the high priority queue length to the total queue length given that the queue is not empty resulted from the approximation denoted by $\mathrm{E}_{app.}(n_h/(n_h+n_l)|n_h+n_l>0)$, and callers anticipated waiting time distributions for different announcement partitions given by $F^h(t;A^h_{j_h})$ and $F^l(t;A^l_{j_l})$.

Step 3. Simulation:

Step 3.1. Simulate the system considering the same setting as that in the approximation (i.e. the same callers' parameters, arrival rates, service rates, number of agents, announcement partitions and the high priority callers' advancement τ) and assuming that callers' expectations about their waiting time distribution are $F^h(t;A^h_j)$ and $F^l(t;A^h_j)$. In the simulation, callers' abandonment decisions are modeled using the optimal stopping model in Section 4.

Step 3.2. Calculate the expected ratio of the high priority queue length to the total queue length given that the queue is not empty resulted from the simulation denoted by $\mathrm{E}_{simul.}(n_h/(n_h+n_l)|n_h+n_l>0)$.

Step 4. If $|\ \mathrm{E}_{app.}(n_h/(n_h+n_l)|n_h+n_l>0)$- $\mathrm{E}_{simul.}(n_h/(n_h+n_l)|n_h+n_l>0)\ |<0.001$ then set $\widehat{\Delta\tau}=\Delta\tau$ and go to Step 5. Otherwise, replace $\Delta\tau$ with $\Delta\tau+1$ and go to Step 1.[20]

Step 5. Report callers' average waiting times and abandonment rates derived using the approximation.

E Comparison of the Prediction Performance of the Optimal Stopping Model with a Simpler Myopic Model

In this section we compare the prediction performance of our stopping time model with a simpler myopic model for callers' abandonment behavior, referred to as the straw-man model. In this model callers do not solve a dynamic program, but simply compare the following utilities while making decision in period t:

- Utility from waiting given by $r - c\ E[W \mid \textit{Announcement message history}] + \epsilon_0$
- Utility from abandonment given by $0 + \epsilon_1$

where ϵ_0 and ϵ_1 are type-I extreme value shocks and $E[W|\textit{Announcement message history}]$ is the expected delay given the announcement history. If there were no shocks added to the utilities, a caller would abandon if r/c is less than the expected delay $E[W|\textit{Announcement message history}]$.

[20] Given that the solutions of the state space collapse approximation are integer numbers, the change in $\mathrm{E}_{app.}(n_h/(n_h+n_l)|n_h+n_l>0)$ across subsequent iterations is not a continuous function of $\Delta\tau$, and consequently, there is the possibility of not achieving the stopping tolerance (in our extensive analyses for different cases, we always could achieve the tolerance. But theoretically there could be a case that we could not achieve it). For these cases, we find $|\ \mathrm{E}_{app.}(n_h/(n_h+n_l)|n_h+n_l>0)$- $\mathrm{E}_{simul.}(n_h/(n_h+n_l)|n_h+n_l>0)|$ for all values of $\Delta\tau$ between -1800 and 1800 and choose the one that corresponds to the lowest absolute difference as the fudge factor.

We assume that callers are heterogeneous in their parameters. Consequently, callers have different reward and cost ratios. Under the assumption that the random shocks have type-I extreme value distributions, the abandonment probability in this model takes the form of Logit probabilities similar to the probabilities in our optimal stopping model. However, the fundamental difference is that in the straw-man model callers do not take it into account that they may abandon in later periods. In contrast, our stopping time model includes the expected future utility that stems from all future (optimal) decisions.

To test which model (our stopping time model, or the straw-man model) can predict callers abandonment behavior better, we performed out of sample tests on calls between 9 am and 12 pm, and calls between 12 pm and 4 pm. The summary statistics of the data subsets are shown in Table 14.

Subset	Average waiting time (sec.)	Abandonment rate
High priority (9 am-12 pm)	88.40	18.42 %
High priority (12 pm-4 pm)	94.96	17.44 %
Low priority (9 am-12 pm)	103.45	33.18 %
Low priority (12 pm-4 pm)	104.21	33.16 %

Table 14: Average waiting times and abandonment rates of the callers in the data subsets used in the out of sample tests.

As can be seen in Table 14, the difference between callers abandonment behavior in the subsets of the high priority data is higher than the difference between the subsets of the low priority data. So we anticipate the predictions for the low priority group to be easier, and hence, more accurate.

The relative and absolute errors in predicting the abandonment rates for different combinations of the training and test data for the high and low priority classes are shown in the following tables.

Training set	Test set	Optimal Stopping Model	Straw-man Model
High priority (9 am-12 pm)	High priority (12 pm-4 pm)	2.43 %	3.97 %
High priority (12 pm-4 pm)	High priority (9 am-12 pm)	0.68 %	1.32 %
Average across all tests		1.56 %	2.65 %

Table 15: Comparison of the absolute errors in predicting the abandonment rates for the optimal stopping model and the straw-man model for the subsets of the high priority callers.

Training set	Test set	Optimal Stopping Model	Straw-man Model
High priority (9 am-12 pm)	High priority (12 pm-4 pm)	13.94 %	22.78 %
High priority (12 pm-4 pm)	High priority (9 am-12 pm)	3.68 %	7.17 %
Average across all tests		8.81 %	14.98 %

Table 16: Comparison of the relative errors in predicting the abandonment rates for the optimal stopping model and the straw-man model for the subsets of the high priority callers.

Training set	Test set	Optimal Stopping Model	Straw-man Model
Low priority (9 am-12 pm)	Low priority (12 pm-4 pm)	2.61 %	2.60 %
Low priority (12 pm-4 pm)	Low priority (9 am-12 pm)	3.79 %	4.06 %
Average across all tests		3.20 %	3.33 %

Table 17: Comparison of the absolute errors in predicting the abandonment rates for the optimal stopping model and the straw-man model for the subsets of the low priority callers.

Training set	Test set	Optimal Stopping Model	Straw-man Model
Low priority (9 am-12 pm)	Low priority (12 pm-4 pm)	7.89 %	7.83 %
Low priority (12 pm-4 pm)	Low priority (9 am-12 pm)	11.42 %	12.15 %
Average across all tests		9.65 %	9.99 %

Table 18: Comparison of the relative errors in predicting the abandonment rates for the optimal stopping model and the straw-man model for the subsets of the low priority callers.

The results demonstrated in Tables 15-18 suggests the following points:

- If the test and training sets are very similar in terms of callers abandonment behavior (such as in the low priority data subsets with very similar abandonment rates and average waiting times), our model does weakly better than the straw-man model. In this case the difference in prediction power between the two models is not high.
- If the difference between the abandonment behavior of the callers in the training and test sets is more significant (such as in the high priority data subsets) our model does a better job in predicting callers' abandonment behavior.

This analysis suggests that our stopping time model has better prediction power.

F Validating Equilibrium Computation

We validate different blocks of our proposed equilibrium computation approach in different sections of the paper (out-of-sample test to show the accuracy of our optimal stopping model in Section 4.2, and simulation studies in the online Appendix D to show the accuracy of the Markovian approximation). Consequently, the equilibrium we compute by solving the system of equations presented in Section 5.2 should not be far from the true equilibrium. We test this by performing a series of simulation studies.

Figure 3 in Section 5 shows the building blocks of our equilibrium computation. The system is in equilibrium state if waiting time distributions used in the derivation of the abandonment time distributions (based on the model of callers abandonment behavior) match the waiting time distributions derived from the Markovian approximation. We ensure this by imposing these conditions as equality constraints that need to be met in equilibrium.

Our approach to verify that our computation yields the equilibrium via simulation is based on the following observation: Our computational approach yields abandonment and waiting time distributions. Consider simulating a queueing system, where callers are endowed with exogenous

abandonment time distributions that are derived from our computation. The simulation yields waiting time distributions for each class. If our computation indeed yields the equilibrium, then the waiting time distribution and the corresponding summary statistics must be close to those from simulation. We verify this by performing a series of simulation studies with the following steps:

Step 1. Compute the system equilibrium using our approach and find average waiting times and abandonment rates.

Step 2. Simulate the call center assuming that callers abandon according to the abandonment time distributions calculated in the previous step. We run the simulation 20 times for 200,000 callers in each simulation trial.

Step 3. Finally, we find the confidence intervals (CIs) for abandonment rates and average waiting times resulted from 20 simulation trials. If abandonment rates and average waiting times found in Step 1 fall within the confidence intervals, we conclude that the equilibrium computed using our approach is close to the true equilibrium.

We performed this simulation study for the four announcement partitions in Tables 7 in Section 6: announcement partition with one subset (no information), announcement partition with two subsets, announcement partition with three subsets and announcement partition with fourteen subsets (full information). Tables 19-22 show abandonment rates and average waiting times found using our equilibrium computation method and the simulation method as explained above. The waiting times are in seconds.

Announcement partition with one subset (No inf.)	High priority				Low priority			
	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Computation	10.20%	49.99	48.85	60.06	27.74%	57.40	55.69	61.93
Simulation (mean)	10.29%	50.34	49.28	59.54	26.77%	55.52	53.43	61.34
Simulation (st. dev.)	0.33%	0.39	0.41	0.26	0.58%	1.37	1.44	1.22
Upper bound of 95% CI	10.98%	51.16	50.14	60.09	27.98%	58.38	56.44	63.89
Lower bound of 95% CI	9.60%	49.52	48.42	58.99	25.56%	52.66	50.42	58.79

Table 19: Comparison of the results of the equilibrium computation method and the simulation for the case with an announcement partition with one subset (no inf.).

Announcement partition with two subsets	High priority				Low priority			
	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Computation	10.18%	49.93	48.91	58.99	27.54%	57.48	55.97	61.74
Simulation (mean)	10.25%	50.11	49.23	57.98	26.68%	55.43	53.48	61.00
Simulation (st. dev.)	0.36%	0.38	0.40	0.55	0.49%	1.27	1.39	0.89
Upper bound of 95% CI	11.01%	50.91	50.07	59.14	27.71%	58.09	56.39	62.78
Lower bound of 95% CI	9.49%	49.31	48.40	56.83	25.66%	52.77	50.57	59.13

Table 20: Comparison of the results of the equilibrium computation method and the simulation for the case with an announcement partition with two subsets.

Announcement partition with three subsets	High priority				Low priority			
	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Computation	10.07%	48.48	48.03	52.49	26.66%	56.91	56.77	57.31
Simulation (mean)	10.14%	48.88	48.57	51.82	25.95%	54.83	54.32	56.32
Simulation (st. dev.)	0.35%	0.34	0.35	0.43	0.51%	1.15	1.32	0.94
Upper bound of 95% CI	10.88%	49.58	49.31	52.73	27.01%	57.24	57.09	58.28
Lower bound of 95% CI	9.40%	48.18	47.84	50.91	24.88%	52.42	51.56	54.35

Table 21: Comparison of the results of the equilibrium computation method and the simulation for the case with an announcement partition with three subsets.

Announcement partition with fourteen subsets (Full inf.)	High priority				Low priority			
	P(A)	E(W)	E(W|S)	E(W|A)	P(A)	E(W)	E(W|S)	E(W|A)
Computation	9.98%	47.37	47.09	49.84	26.10%	54.46	54.83	53.42
Simulation (mean)	10.00%	48.05	47.94	49.08	25.66%	52.60	52.77	52.13
Simulation (st. dev.)	0.35%	0.39	0.51	0.47	0.63%	1.11	1.25	0.97
Upper bound of 95% CI	10.74%	48.88	49.00	50.06	26.99%	54.93	55.39	54.15
Lower bound of 95% CI	9.26%	47.23	46.87	48.09	24.34%	50.27	50.14	50.11

Table 22: Comparison of the results of the equilibrium computation method and the simulation for the case with an announcement partition with fourteen subsets (full inf.).

As can be seen in Tables 19-22 for all cases abandonment rates and waiting times calculated using our equilibrium computation approach fall within the 95% confidence intervals. This suggests that our equilibrium computation method can estimate the true equilibrium with acceptable accuracy.