
Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# How Do Delay Announcements Shape Customer Behavior? An Empirical Study

Qiuping Yu, Gad Allon, Achal Bassamboo

To cite this article:
Qiuping Yu, Gad Allon, Achal Bassamboo (2017) How Do Delay Announcements Shape Customer Behavior? An Empirical Study. Management Science 63(1):1-20. http://dx.doi.org/10.1287/mnsc.2015.2335

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 1, January 2017, pp. 1–20
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

# How Do Delay Announcements Shape Customer Behavior? An Empirical Study

**Qiuping Yu,**[a] **Gad Allon,**[b] **Achal Bassamboo**[b]

[a] Kelley School of Business, Indiana University, Bloomington, Indiana 47405; [b] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208

**Contact:** qiupyu@indiana.edu (QY); g-allon@kellogg.northwestern.edu (GA); a-bassamboo@kellogg.northwestern.edu (AB)

**Received:** April 16, 2013
**Accepted:** October 3, 2014
**Published Online in Articles in Advance:** March 8, 2016

**https://doi.org/10.1287/mnsc.2015.2335**



**Abstract.** In this paper, we explore the impact of delay announcements using an empirical approach by analyzing the data from a medium-sized call center. We first explore the question of whether delay announcements impact customers' behavior using a nonparametric approach. The answer to this question appears to be ambiguous. We thus turn to investigate the fundamental mechanism by which delay announcements impact customer behavior, by constructing a dynamic structural model. In contrast to the implicit assumption made in the literature that announcements do not directly impact customers' waiting costs, our key insights show that delay announcements not only impact customers' beliefs about the system but also directly impact customers' waiting costs. In particular, customers' per-unit waiting cost decreases with the offered waiting times associated with the announcements. The results of our counterfactual analysis show that it may not be necessary to provide announcements with very fine granularity.

**History:** Accepted by Yossi Aviv, operations management.
**Supplemental Material:** The online appendix is available at https://doi.org/10.1287/mnsc.2015.2335.

**Keywords:** delay announcement • patience time • customer abandonment • structural estimation

## 1. Introduction

Delay announcements that inform customers about anticipated delays are prevalent in service systems. Call centers often use recorded announcements to inform callers of the congestion in the system and encourage them to wait for an available agent. Whereas some of these announcements do not provide much information, such as the common message, "Due to a high volume of calls, we are unable to answer your call immediately," some call centers go as far as providing their customers with estimates of their waiting time or places in the queue. In many service systems, where the real state of the system is not visible to customers, delay announcements may affect customer behavior and may, in turn, have significant impact on system performance. Consequently, it is important for service providers to understand how delay announcements impact customer behavior and use delay announcements as a tool to induce the appropriate customer behavior, steering the system so as to maximize profits.

As these delay announcements are more and more widely used to inform customers about anticipated delay, two major questions arise. The first question is whether delay announcements indeed impact customer behavior. In thinking about customer behavior, we follow the common practice in the call center industry to use customer abandonment as a proxy for customers' discontent with long waits. Hence, to address the first question, we study the impact of delay announcements on customers' reneging behavior. The answer to the first question turns out to be somewhat ambiguous. We thus turn to a more fundamental question, where we investigate the mechanism by which delay announcements impact customer abandonment behavior. To address these questions, we use an empirical approach by analyzing the data of a call center in the financial industry. The data include detailed information about each individual call at a medium-sized call center of an Israeli bank, where delay announcements with estimates of anticipated delay are provided to the customers. (We will describe the data in more detail in §3.1.)

The relationship between delay announcements and customer behavior has been explored in the operations management literature. Most of the research in this literature is theoretical in nature. The key assumption made in this literature is that customers use delay announcements to update their beliefs about the congestion level experienced by the system. Following this assumption, we build our base structural model, in which customers use announcements to update their beliefs on the waiting time in the system. The goal of this work is to test whether the models based on this assumption can indeed explain customer behavior.

We study the first question using a nonparametric approach, and we investigate the second question of interest using a structural estimation approach.

We start with the question of whether delay announcements impact customers' abandonment behavior. Specifically, we compare the patience time induced by different announcements. Our results show that in most cases, the patience times of customers receiving different announcements are not statistically different. (We will describe the approach and its results in more detail in §3.3.)

To uncover the fundamental mechanism by which announcements impact customer behavior, we turn to the second question and propose a dynamic structural model. We model customer abandonment behavior as a dynamic decision process from the moment the first announcement is given. In particular, customers make abandonment decisions by trading off between the waiting cost and the reward that they receive from the service. We assume customers are forward-looking and heterogeneous in their cost–reward ratio. The model also allows customers to make decisions based on external factors that are modeled through idiosyncratic shocks. The external factors can be either customers' lack of adherence to pure rational decision making or customers' private information. An important feature of our structural model is that it accounts for the delay announcements received by customers. We start with a base model, which is built on the common assumption made in the theoretical literature that customers update their beliefs about the offered waiting time based on the announcements received. Using the censored maximum likelihood method, we estimate the cost–reward ratio and the variance of the idiosyncratic shock. If we believe this model, the estimation results imply that the phenomenon of customers abandoning at different times is driven by both the heterogeneity of their cost–reward ratios and the idiosyncratic shocks.

As mentioned above, the base model follows the theoretical literature in assuming that the announcements only impact customers' beliefs on the offered waiting time. To test this assumption, we construct a model that assumes that the announcements not only impact customers' beliefs on the offered waiting time but also directly impact their waiting costs. Our estimation results show that the model where customers not only update their beliefs about the system with the announcements but also update their waiting costs best explains the observed customers' abandonment behavior at this call center. Moreover, we find that the cost–reward ratio decreases with the offered waiting time associated with the announcements, whereas the variance of the idiosyncratic shock increases. Such an impact of announcements on customers' per-unit waiting costs can be explained by the economic theory that shows that the value of time equals the opportunity cost. This, combined with the observation that customers receiving longer estimated delays anticipate longer average waiting times, explains the result that the announcements impact customers' patience time in a nontrivial manner or sometimes do not impact customers' patience time at all. (We elaborate on this interpretation further in §6.3.)

To the best of our knowledge, this is the first work that relaxes the implicit assumption made in the theoretical literature that delay announcements only impact customers' beliefs about the offered waiting time. By estimating a structural model where this assumption is not imposed, we show that delay announcements not only impact customers' beliefs about the offered waiting time but also directly impact customers' per-unit waiting costs. Specifically, we find that customers' per-unit waiting cost decreases with the offered waiting time associated with the delay announcements, whereas the variance of the idiosyncratic shock increases. These results should provide call centers with new insights on how to measure the impact of delay announcements on their customers and how to inform their customers about the anticipated delay. In particular, we show that the survival analysis used by most call centers is not sufficient and that a richer analysis is needed to unravel the main effects of delay announcements. We also show that providing delay announcements improves the overall surplus of customers compared with the case when there is no information provided. The surplus of a customer equals the reward obtained minus her actual total waiting costs. Furthermore, we find that a simple announcement system that is made up of three messages—i.e., low, medium, and high—could result in better performance of the system in terms of customers' surplus, compared with the announcement systems with more refined granularity.

## 2. Literature Review

In this paper, we study the impact of delay announcements on customer abandonment behavior, using a structural estimation approach. We next outline the four branches of the literature that this work is related to: delay announcements, customer abandonment behavior, structural estimation, and the recent empirical studies of waiting in queues.

### 2.1. Delay Announcements

One of the first papers to discuss the question of whether to reveal the queue length information to customers is Hassin (1986), which studies the problem of whether a price-setting and revenue-maximizing service provider should provide the queue length information to arriving customers. Whitt (1999) brings the concept of information revelation to the specific setting of call centers, where call centers communicate with their customers about the anticipated delay by providing delay announcements. The author shows that the average waiting time can be reduced when accurate

announcements are provided. Guo and Zipkin (2007) extend the model above by studying the impact of delay announcements with different information accuracy. They show that accurate delay information might improve or hurt the system performance.

Although all these papers assume customers do not abandon the system once they join the queue, Armony et al. (2009) relax this assumption. The authors study the performance impact of making delay announcements to arriving customers in a many-server queue setting with customer abandonment. Customers are provided with either the delay of the last customer to enter service or an appropriate average delay upon arrival. The authors show that within the fluid-model framework, under certain conditions, the actual delay coincides with the announced delay. Motivated by this type of delay announcement, Ibrahim and Whitt (2009) explore the performance of different real-time delay estimators based on a recent delay experienced by customers, allowing for customer abandonment.

All the aforementioned works assume that the information is credible and is treated as such by customers, which is relaxed in Allon et al. (2011). It examines the problem of information communication by considering a model in which both the firm and the customers act strategically: the firm in choosing its delay announcement while anticipating customer response and the customers in interpreting these announcements and in making the decision of whether to join the queue.

The impact of delay announcements on customers' behavior has also been investigated experimentally. Hui and Tse (1996) show that both waiting duration information and the queue information have a significant impact on customers' reaction in the intermediate or long wait condition, but not in the short wait condition through an experimental study. Munichor and Rafaeli (2007) extend the work above by studying the impact of providing a sense of progress on customer behavior. Specifically, the authors examine the effect of time perception and sense of progress in telephone queues on caller reactions to three waiting time fillers: music, apologies, and information about location in the queue. The authors show that call evaluations are most positive with information about their location in the queue as the time filler, whereas the sense of progress in the queue rather than perceived waiting time mediated the relationship between telephone waiting time filler and caller reactions.

### 2.2. Customer Abandonment Behavior

Customers' abandonment behavior associated with announcements is one of the key components of our model. Several papers have proposed to use a rational decision model to capture customers' abandonment behavior (e.g., Haviv and Ritov 2001, Mandelbaum and Shimkin 2000), where customers make abandonment decisions by weighing service utility against expected waiting costs. In particular, our base model is inspired by Mandelbaum and Shimkin (2000), in the sense that customers have beliefs about the offered waiting time and act as utility maximizers by trading off between service reward and expected waiting costs. However, our structural model differs from the model in Mandelbaum and Shimkin (2000) in two important aspects: (1) our model accounts for the delay announcements, whereas Mandelbaum and Shimkin consider a setting where no announcements are provided and customers form beliefs about the system based on their experiences; and (2) although in Mandelbaum and Shimkin different customers' behavior can be explained through the differences in the cost–reward ratios, as part of embedding their model in an empirical study, we also allow for idiosyncratic shocks. For a more detailed review of customer abandonment behavior, we refer the reader to Hassin and Haviv (2003).

### 2.3. Structural Estimation

As we mentioned above, we characterize customers' utility not only through the reward of being served and the waiting costs but also through idiosyncratic shocks that are unobservable to the researchers. This type of random utility model is commonly seen in the structural estimation literature. In that literature, two of the most relevant papers to our work are Rust (1987) and Nair (2007). Rust studies a structural regenerative optimal stopping model for bus engine replacement, where at each time period, the superintendent of maintenance chooses whether to replace the engine of a bus or not by trading off between the cost of overhaul and the cost of unexpected failure. Nair develops a framework to empirically investigate the optimal pricing over time of a firm selling a durable-good product to forward-looking consumers, with an application to the market for video games in the United States. In particular, the author proposes a dynamic discrete choice model to capture customers' forward-looking behavior, where customers make decisions about whether to buy the product in the current period or wait until the next time period by trading off between the utility from the current purchase and the value of buying the product at a lower price in the future. Akşin et al. (2012) adopt the dynamic discrete choice model in Nair to study callers' sensitivities to waiting in call centers. Specifically, the authors model callers' decision-making processes as an optimal stopping time problem. At each time period, customers decide whether to abandon or to continue to wait by trading off between the reward from being served and the cost of waiting. The key differences between Akşin et al. and our work are that our model accounts for the announcements and that the data we use contain detailed and accurate information about the delay announcements.

### 2.4. Empirical Studies of Waiting in Queue

Our paper studies how providing information to customers on anticipated delays impacts customer behavior while waiting in a queue. One very relevant paper is Lu et al. (2013). The authors study empirically how waiting in a queue in the context of a retail store affects customer purchasing behavior where the queue is visible. The authors find that waiting in the queue impacts purchase incidence in a nonlinear manner and that customers appear to focus mostly on the queue length, without adjusting enough for the speed at which the queue moves. They also find that customer sensitivity to waiting is heterogeneous and has a negative correlation with price sensitivity. Another paper, Buell and Norton (2011), shows that perception of the service providers' effort induces customers' satisfaction through an experimental study. We refer readers to the survey in Mandelbaum and Zeltyn (2013) for a detailed review of other papers that deal with impatient customers.

## 3. Data and Preliminary Results

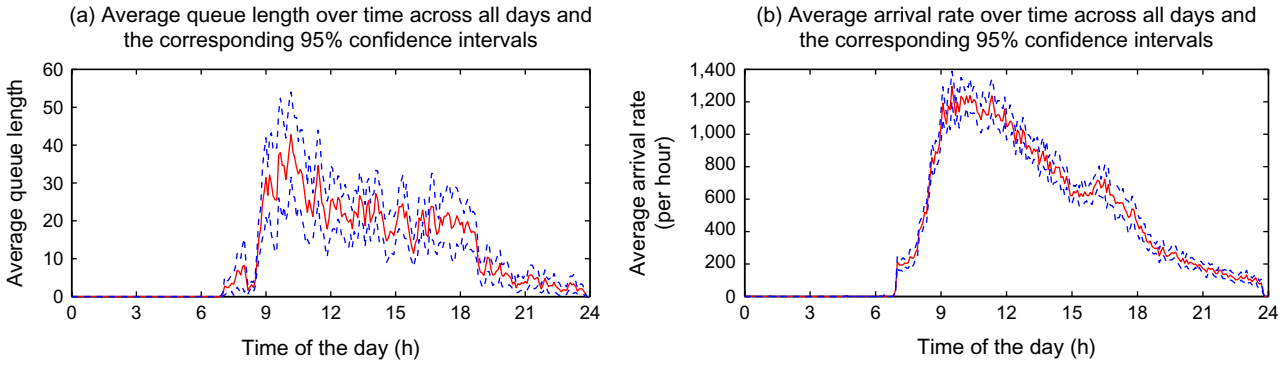
### 3.1. Data

The data are originally obtained from a medium-sized call center of a bank in Israel. (They are generously made available to us by the SEE Center at the Technion in Israel.) The data have been recorded from March 1, 2010 to March 31, 2010. The data contain detailed information about each individual call. Each call at the call center has three components: the interactive voice response (IVR) component where customers interact with the automated system, the queue component where customers wait for service, and the service component where customers are served by live agents. The data capture the starting time, the ending time, and the outcome (e.g., abandoned or not) of each component for each call. The timeline of a customer at the call center is as follows: When a customer makes a call to the call center, she first goes through the IVR component. If the problem is addressed by the IVR, the call is terminated there. Otherwise, the customer waits in the queue component until a live agent becomes available (conditional on the customer having the appropriate priority). While a customer waits in the queue component, she might be provided with delay announcements. The data do not have information about the exact announcements given to the customers. However, the call center provides us with two important pieces of information that were used to recover the exact announcements given to the customers: (1) the algorithm the call center uses to generate the estimate of the anticipated delay and (2) the exact time when the information is provided. At this call center, customers are not provided with delay announcements immediately after they enter the queue. In fact, 60 seconds after the customer enters the queue, she receives her first announcement. The customers are updated every 90 seconds after they receive the first announcement about the state of the system. In terms of the algorithm, customers are given estimates of the anticipated delay, e.g., a customer might receive the announcement, "Your expected delay is about 69 seconds."

At this call center, customers are sorted to different classes. The classification is based on customers' demographics, the service they seek, and their business importance to the bank. There are 268 classes in total at this call center. In this paper, we focus on the two biggest customer classes, which contain 58% of all customers who receive delay announcements. One class contains the VIP customers looking for checking account–related service,[1] which we refer to as the VIP class. The other class includes regular customers looking for checking account–related service, which we refer to as the regular class. From an operational perspective, customers from the VIP class have higher priority than those from the regular class. Furthermore, we observe that the VIP customers call the call center more frequently compared with the regular customers on average during the time when our data were collected. In the VIP class, there are 9,997 customers receiving at least one announcement while they wait in the queue. Among those customers who receive at least one announcement, 38.6% receive at least two announcements, 6.9% receive at least three announcements, and only 1.2% receive four or more announcements. In the regular class, there are 16,905 customers receiving at least one announcement. Among those customers who receive at least one announcement, 48.6% receive at least two announcements, 10.4% receive at least three announcements, and only 2.3% receive four or more announcements.

To provide some context and demonstrate the congestion being experienced at this call center, we compute the average queue length and the average arrival rate at this call center. We observe that the average queue length is 11.0 customers, and the average arrival rate is 418.2 customers per hour. To illustrate the variation over time, we plot the average queue length and the average arrival rate over time across all days along with the corresponding 95% confidence intervals; see Figures 1(a) and 1(b). We observe that the call center is fairly busy from 9 A.M. to 6 P.M., with significant variation in the congestion experienced by customers. It is worth mentioning that the working days in Israel are Sunday through Thursday. Note that our data show that the average queue length is significantly longer toward the end of day on Thursday compared with the ones on the middle days of the week (i.e., Monday, Tuesday, or Wednesday). Furthermore, we observe that the average queue length over time on Sundays seems to be slightly shorter than that on Monday, Tuesday,

**Figure 1.** (Color online) Average Arrival Rate and Average Queue Length at the Call Center Across All Days




and Wednesday. Because of the potential "day-of-the-week" effect and other considerations mentioned above, we focus our analysis on the data from 9 A.M. to 6 P.M. on Monday, Tuesday, and Wednesday.

### 3.2. Informative and Influential Announcement System

To study the impact of delay announcements on customers' abandonment behavior, let us clarify the concepts that are instrumental in both the nonparametric approach (in §3.3) and the structural estimation approach (in §4). Specifically, we start with the concepts of offered waiting time in a queue and customers' patience time. The *offered waiting time* of a given customer is defined as the time that she has to wait before she can receive the service from a live agent, if she never abandons the system. As for the patience time, we use it to characterize customers' abandonment behavior in the nonparametric approach, where we take a more traditional view compared with the one that we take in the structural estimation approach. In particular, the *patience time* is defined as the maximum time a customer can endure before she abandons the system.

We now turn to the definitions that are used to characterize the announcements and their impact. We use the term *announcement system* to describe the set of announcements that the firm provides, including their contents and timing. We first confirm that the announcements provide information about the state of the system, and we then study whether and how they influence customers' abandonment behavior in the two approaches that we mentioned above. To that end, let us introduce the following definitions of an announcement system being "informative" and "influential."

**Definition 1.** We say that an announcement system is *informative* if the offered waiting time of a customer receiving an announcement with longer estimated delay has first-order stochastic dominance over that of a customer receiving an announcement with shorter estimated delay.

**Definition 2.** We say that an announcement system is influential if there are at least two different announcements such that the patience time of a customer receiving one announcement has first-order stochastic dominance over the patience time of a customer receiving the other one.

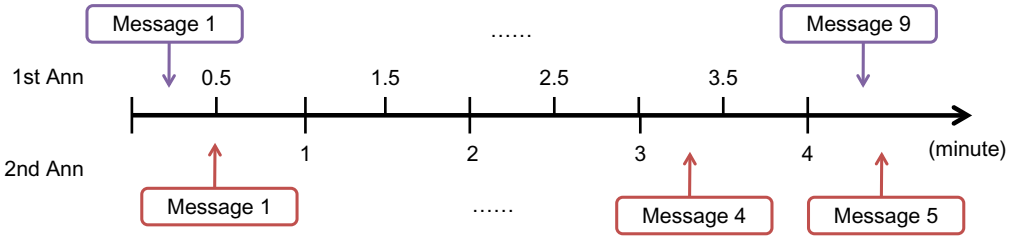
To facilitate the preliminary analysis in §3.3 and the estimation of the structural models in §5, we divide the announcements into intervals according to the announced time so that there are enough customers in each interval.[2] As we pointed out earlier, customers are given estimates of the anticipated delay up to the second. However, not many customers receive a given announcement, e.g., the announcement "Your waiting time is about 69 seconds" is received by only 0.5% of customers in the VIP class. To balance the number of announcement intervals and the number of customers belonging to each announcement interval, we construct the announcement system for the regular customers as follows: We divide the first announcements and the second announcements received by customers from the regular class into intervals as illustrated in Figure 2; similarly, we divide the third announcements into two intervals, with the first interval from zero to four minutes and the second interval equal to or longer than four minutes. We construct the announcement system for the VIP customers using similar criteria. The detailed announcement system for the VIP class is illustrated in Figure 4 of Online Appendix A (online appendices available as supplemental material at https://doi.org/10.1287/mnsc.2015.2335).

### 3.3. Preliminary Results

In this section, we address our first research question of whether delay announcements impact customers' behavior. Following the traditional approach that call centers take to analyze customers' abandonment behavior,[3] we start by verifying whether or not the announcement system is informative, and then we study whether it is also influential.

To verify whether the announcement system is informative and influential, we need to compare the offered

**Figure 2.** (Color online) Announcement System for the Regular Class




waiting time and the patience time across different announcements. Note that the offered waiting time is right censored[4] for abandoned customers, because they must abandon the system before they would have been served. Similarly, customers' patience time is right censored for customers who are served by live agents, because they must get served before they would have abandoned the system. To address the censoring issue, we conduct two nonparametric statistical tests constructed specifically for comparing survival functions of right-censored data. The two tests that we use are the Gehan's generalized Wilcoxon test (Gehan 1965) and the log-rank test (Peto and Peto 1972).[5]

The results of the tests[6] show that the first and third announcements are both informative[7] for both the regular class and the VIP class, but not the second announcement.[8] Thus we confirm that the first and third announcements indeed inform customers about the state of the system but not the second announcement. If estimates of the anticipated delay in the first and third announcements are large, then congestion in the system is high. Yet this is not the case for the second announcement.

We now turn to the question of whether the announcement system is influential. Let $T_k^j$ represent the patience time of customers receiving the $j$th announcement from the $k$th interval, e.g., $T_1^2$ is the patience time of customers receiving the second announcement from the first interval, which is between 0 and 30 seconds. We find that, for the VIP class, the first announcement influences customers' patience time, whereas the second and third announcements do not. In particular, we have $T_5^1 <_{\text{st}} T_3^1$[9] for the VIP class. For the regular class, the first and third announcements do not influence customers' patience time, whereas the second announcement does. In particular, we have $T_1^2 <_{\text{st}} T_5^2$ for the regular class.

It appears that there are no particular patterns in which we can tell how announcements impact customers' patience time by analyzing customers' ultimate responses to the announcements in terms of their patience time. The result that delay announcements do not impact customers' patience times does not mean that these announcements have no impact whatsoever on customers. To uncover the fundamental mechanism by which announcements may impact customer behavior, we turn to the second question and propose a dynamic structural decision model in §4.

## 4. The Structural Model

To explain how customers react to explicit and real-time information on anticipated delays, we present a dynamic structural model where customers decide whether and when to abandon the system. We model customers' abandonment behavior as a dynamic decision process from the moment the first announcement is given. The time horizon is divided into small time periods of equal length. At the beginning of each time period, a customer decides whether to abandon the system or wait until the next time period by trading off between the waiting cost and the reward of the service. We assume customers are forward-looking. An important feature of our structural model is that it accounts for the delay announcements received by the customers over time. In our model, we focus on the first three announcements and disregard the other announcements received by customers during any given call, since only 1.2% of the VIP customers and 2.3% of the regular customers receive four or more announcements. To better describe these announcements, let $M = \{m_1, m_2, \ldots, m_n\}$ be the set of possible messages. Let $\tau_j$ be the time period, at the beginning of which the $j$th announcement is provided, for $j \in \{1,2,3\}$. Moreover, it is important to note that the same customer may make multiple calls to the call center over time. To this end, we let $A_{ih}^j$ denote the $j$th announcement customer $i$ receives during the $h$th call that she makes to the call center, for $j \in \{1,2,3\}$. We then have $\tau_1 = 0$, given that the model is constructed from the moment the first announcement is given.[10] We define $A_{ih}(t)$ as the last announcement that customer $i$ receives by time period $t$ during the $h$th call that she makes. This is given by

$$A_{ih}(t) = \begin{cases} A_{ih}^1 & \text{for } \tau_1 = 0 \le t < \tau_2, \\ A_{ih}^2 & \text{for } \tau_2 \le t < \tau_3, \\ A_{ih}^3 & \text{for } t \ge \tau_3. \end{cases}$$

Now we are ready to construct our base model where customers take the first three announcements into account. To make decisions about whether to wait or to abandon, customer $i$ forms beliefs about the

offered waiting time based on the last announcement $A_{ih}(t)$ she received during the $h$th call that she makes.[11] The cumulative distribution function of the offered waiting time of customers receiving the announcement $A_{ih}(t)$ is denoted by $F^{A_{ih}(t)}(\cdot)$, where the offered waiting time is counted from the moment the first announcement is provided for any given call. The customer also has beliefs about the announcement evolution over time. We match customers' beliefs about the offered waiting time and the announcement evolution to the actual offered waiting time of the customers receiving the corresponding announcements and the actual announcement evolution, respectively.[12] Let us define $\kappa_t^{A_{ih}(t)}$ as the probability of being served at time period $t$, conditional on the fact that customer $i$ is still in the queue at time period $t$ during the $h$th call that she makes. For $t \geq 0$, $\kappa_t^{A_{ih}(t)}$ is given by

$$\kappa_t^{A_{ih}(t)} = \frac{F^{A_{ih}(t)}(t+1) - F^{A_{ih}(t)}(t)}{1 - F^{A_{ih}(t)}(t)}. \tag{1}$$

The value-to-go of customer $i$ at time period $t$ during the $h$th call that she makes to the call center is given by

$$u_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \varepsilon_{iht}(a_{iht}), a_{iht}) = \Psi_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta, a_{iht}) + \varepsilon_{iht}(a_{iht}), \tag{2}$$

where $r_{ih}$ is the reward customer $i$ gets for the service during her $h$th call, $c_{ih}$ is customer $i$'s cost of waiting per time period during her $h$th call, and $a_{iht}$ is customer $i$'s decision about whether to abandon at time period $t$ or wait until time period $t+1$ during her $h$th call. The function $\Psi_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta, a_{iht})$ is the nominal value-to-go of customer $i$ whose last announcement received until time period $t$ during her $h$th call is $A_{ih}(t)$. It is important to note that $\psi_{iht}^{A_{ih}(t)}(\cdot)$ includes not only the nominal utility at time period $t$ but also the nominal utility at the time periods beyond period $t$. The last term $\varepsilon_{iht}(a_{iht})$ is the idiosyncratic shock that is observed only by customer $i$ but not the researchers. The idiosyncratic shock captures customers' lack of adherence to pure rational decision making or customers' private information. Note that all these factors may vary with customers' decisions on whether to abandon the system. Thus, we let the idiosyncratic shock be a function of customers' action $a_{iht}$. Further, we assume the idiosyncratic shock follows a Gumbel distribution[13] with $\varepsilon_{iht}(a_{iht}) \sim \text{Gumbel}(\alpha, \theta)$, where $\alpha$ is the location parameter and $\theta$ is the scale parameter. We normalize the mean of $\varepsilon_{iht}(a_{iht})$ to be zero, and as a result, $\alpha = -\gamma\theta$, where the Euler–Mascheroni constant is given by $\gamma \approx 0.577$. We assume that $\varepsilon_{iht}(a_{iht})$ are independent across different time periods, different calls, and different customers' decisions and that they are also independent from the nominal value-to-go $\Psi_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta, a_{iht})$. It is important to note that $\psi_{iht}^{A_{ih}(t)}(\cdot)$ is the expected utility of time period $t$ and beyond, over the idiosyncratic shocks. Thus, $\psi_{iht}^{A_{ih}(t)}(\cdot)$ is a function of the idiosyncratic parameter denoted by $\theta$ but not a function of any realization of the idiosyncratic shocks.

Customer $i$ makes a decision whether to abandon the system or wait until the next time period at each time period $t$ during the $h$th call that she makes, to maximize her utility. In particular, the optimal action $a_{iht}$ is given by

$$a_{iht} = \arg\max u_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \varepsilon_{iht}(a_{iht}), a_{iht}).$$

If customer $i$ decides to abandon, she would do so immediately at the beginning of time period $t$ and will get the nominal utility, which we normalize to be zero. If customer $i$ chooses to wait, she will stay in the queue until the next time period. As she waits, she may get served, in which case she incurs the cost of waiting for time period $t$ and gets the reward from the service. Otherwise, she incurs the cost of waiting for time period $t$, and then she decides again whether to abandon or continue to wait as she enters the next time period. Thus, for $t \geq 0$, the nominal value-to-go at time period $t$ is given by

$$\Psi_{it}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta, a_{iht}) = \begin{cases} -c_{ih} + \kappa_t^{A_{ih}(t)} r_{ih} + (1 - \kappa_t^{A_{ih}(t)}) V_{it}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta) \\ \quad \text{if } a_{iht} = \text{wait}, \\ 0 \quad \text{if } a_{iht} = \text{abandon}, \end{cases} \tag{3}$$

where $V_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta)$ denotes the expected total future utility of customer $i$ beyond time period $t$ conditional on the fact that she decides to wait until time period $t+1$ during her $h$th call. We refer to it as the aggregated future utility function. It is given by

$$V_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta) = \mathbb{E}\left[ \max_{a \in \{\text{abandon, wait}\}} \left( u_{i,h,t+1}^{A_{ih}(t+1)}(r_{ih}, c_{ih}, \varepsilon_{i,h,t+1}(a), a) \right) \middle| A_{ih}(t) \right]. \tag{4}$$

The following proposition shows that we can write $V_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta)$ in a closed form. The proofs of all results including this proposition are relegated to Online Appendix D.

**Proposition 1.** *Let the idiosyncratic shock $\varepsilon_{iht}(a_{iht})$ be Gumbel distributed. We aslo assume it is independent and identically distributed (i.i.d.). Given (2), (4) can be written as*

$$V_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta) = \mathbb{E}\left[ \theta \log\left( 1 + \exp\left( \frac{\Psi_{i,h,t+1}^{A_{ih}(t+1)}(r_{ih}, c_{ih}, \theta, wait)}{\theta} \right) \right) \middle| A_{ih}(t) \right]. \tag{5}$$

It is worth mentioning that if no new announcement is provided at time period $t+1$, i.e., $t+1 \neq \tau_j$ with $j \in \{2,3\}$, $A_{ih}(t+1)$ is identical to $A_{ih}(t)$. By contrast, if a new announcement is given at time period $t+1$, i.e., $t+1=\tau_j$ with $j \in \{2,3\}$, $A_{ih}(t+1)$ is the new announcement. In particular, for time period $t$, where $t+1=\tau_j$ with $j \in \{2,3\}$, (5) can be written as

$$V_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta) = \sum_{m \in M} \mathbb{P}(A_{ih}(t+1) = m \mid A_{ih}(t)) \cdot \left(\theta \log\left(1 + \exp\left(\frac{\Psi_{i,h,t+1}^{m}(r_{ih}, c_{ih}, \theta, \text{wait})}{\theta}\right)\right)\right).$$

### 4.1. Terminal Condition

Let $\eta(A_{ih}(\infty))$ be the maximum time period we observe in the data among the customers who receive the announcement $A_{ih}(\infty)$, with $A_{ih}(\infty) = \lim_{t\to\infty} A_{ih}(t)$. We assume the expected utility at the time periods beyond $\eta(A_{ih}(\infty))$ is zero. That is, for $t > \eta(A_{ih}(\infty))$,

$$V_{iht}^{A_{ih}(\infty)}(r_{ih}, c_{ih}, \theta) = 0. \tag{6}$$

In constructing these models, one of the goals is to study which announcements impact customer behavior. To that end, we also consider two models that are identical to the base model with the following modifications: (i) a model where customers disregard the third announcements and (ii) a model where customers disregard the second and third announcements. Thus, the first modified model is mathematically equivalent to the base model by setting $\tau_3 = \infty$ and $\lim_{t\to\infty} A_{ih}(t) = A_{ih}^2$; the second modified model is mathematically equivalent to the base model by setting $\tau_2 = \infty$ and $\lim_{t\to\infty} A_{ih}(t) = A_{ih}^1$.

The three models described above all follow the theoretical literature, assuming that announcements only impact customers' beliefs on the offered waiting time. To test this assumption, we construct models that assume that the announcements not only impact customers' beliefs on the offered waiting time but also directly impact their waiting costs. We denote $W_lC_n$ as the model where customers keep updating their beliefs about the system as they receive the first $l$ announcements, and their waiting costs are influenced by the first $n$ announcements for any given call. In this nomenclature, if customers do not update their waiting costs based on the announcements, $n$ would be equal to 0. Thus, with this nomenclature, the base model corresponds to $W_3C_0$, and the two modified models mentioned above correspond to $W_2C_0$ and $W_1C_0$.

To describe how customers' per-unit waiting costs are influenced by the announcements received, let us define $A_{ih}^c(t)$ as the last announcement received by time period $t$ during customer $i$'s $h$th call that impacts her per-unit waiting cost. Note this is analogous to $A_{ih}(t)$, which was defined above as the last announcement received (regardless of whether or not the announcement impacts customers' waiting costs). Thus, the model $W_lC_n$, for $l,n \in \{1,2,3\}$, is identical to the model $W_lC_0$, with the modification that the reward, the cost, and the idiosyncratic shock depend on the announcement $A_{ih}^c(t)$. To show this dependence, we use the superscript and denote the reward, the cost, and the idiosyncratic shock as $r_{ih}^{A_{ih}^c(t)}$, $c_{ih}^{A_{ih}^c(t)}$, and $\varepsilon_{iht}^{A_{ih}^c(t)}(a_{iht})$, respectively. In particular, the utility in the $W_lC_n$ model is given by

$$u_{iht}^{A_{ih}(t)}\left(r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}, \varepsilon_{iht}^{A_{ih}^c(t)}(a_{iht}), a_{iht}\right) = \Psi_{iht}^{A_{ih}(t)}\left(r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}, \theta^{A_{ih}^c(t)}, a_{iht}\right) + \varepsilon_{iht}^{A_{ih}^c(t)}(a_{iht}) \tag{7}$$

instead of (2) for the $W_lC_0$ models with $l \in \{1,2,3\}$. Similarly, the nominal utility in the model $W_lC_n$ is given by

$$\Psi_{iht}^{A_{ih}(t)}(r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}, \theta^{A_{ih}^c(t)}, a_{iht}) = \begin{cases} -c_{ih}^{A_{ih}^c(t)} + \kappa_t^{A_{ih}(t)} r_{ih}^{A_{ih}^c(t)} + (1-\kappa_t^{A_{ih}(t)}) V_{iht}^{A_{ih}(t)} \\ \quad \cdot (r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}, \theta^{A_{ih}^c(t)}), & \text{if } a_{iht} = \text{wait}, \\ 0, & \text{if } a_{iht} = \text{abandon}, \end{cases} \tag{8}$$

where $V_{iht}^{A_{ih}(t)}(r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}, \theta^{A_{ih}^c(t)})$ denotes the expected total future utility of customer $i$ beyond time period $t$ conditional on the fact that she decides to wait until time period $t+1$ during her $h$th call. Based on Proposition 1, this future utility is given by

$$V_{iht}^{A_{ih}(t)}(r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}, \theta^{A_{ih}^c(t)}) = \mathbb{E}\left[\theta^{A_{ih}^c(t+1)} \log\left(1 + \exp\left(\frac{\Psi_{i,h,t+1}^{A_{ih}(t+1)}(r_{ih}^{A_{ih}^c(t+1)}, c_{ih}^{A_{ih}^c(t+1)}, \theta^{A_{ih}^c(t+1)}, \text{wait})}{\theta^{A_{ih}^c(t+1)}}\right)\right) \,\middle|\, A_{ih}(t)\right]. \tag{9}$$

It is worth pointing out that models $W_lC_n$, for $l = n \in \{1,2,3\}$, implicitly assume that if customers update their beliefs about the system with an announcement, they would update their waiting costs simultaneously with the same announcement, whereas the models $W_lC_n$, for $n < l$ and $l,n \in \{1,2,3\}$, assume that not all the announcements that impact customers' beliefs about the system also impact their costs. In addition, we assume that if an announcement does not impact customers' beliefs about the system, it does not impact their waiting costs either. Thus, we do not consider the models with $l < n$ and $l,n \in \{1,2,3\}$.

At this point, it is important to note that we can categorize the models constructed above into two groups. The first group includes the models $W_lC_0$, for $l \in \{1,2,3\}$, where customers only update their beliefs

about the offered waiting time. We refer to these models as *W-updating models*. The other group contains the models $W_lC_n$, for $l, n \in \{1,2,3\}$ and $l \geq n$, where customers update not only their beliefs about the offered waiting time but also their per-unit waiting costs. We refer to them as *WC-updating models*.

### 4.2. Modeling Customers' Heterogeneity

Above, we described customers' decision making. We next describe how we model heterogeneity among customers. Note that customers in the same class may differ in their per-unit costs and their rewards. As it is explained in Mandelbaum and Shimkin (2000), a customer's optimal choice decision depends only on her cost–reward ratio. Thus, instead of modeling customers' heterogeneity through the cost and the reward separately, we do so through the cost–reward ratio. In particular, we assume customers' cost–reward ratio follows a folded normal distribution[14] for each customer class. The parameters of the distribution do not depend on the announcements for the W-updating models, although they do for the WC-updating models. Moreover, it is important to note that the same customer may call the call center multiple times over time. To this end, we let the cost–reward ratios associated with all the calls made by the same customer be the same in the W-updating models. In particular, the cost–reward ratio in the W-updating models is given by $c_{ih}/r_{ih} = |\sigma z_i + \mu|$, where $z_i$ is drawn independently from a standard normal distribution for each customer $i$. The parameters $\mu$ and $\sigma$ are the mean and standard deviation, respectively, of the underlying normal distribution. In contrast to the W-updating models, we have the cost–reward ratios associated with all the calls made by the same customer who receives the same announcements during these calls to be the same in the WC-updating models. In particular, the cost–reward ratio in the WC-updating models is given by $c_{ih}^{A_{ih}^c(t)}/r_{ih}^{A_{ih}^c(t)} = |\sigma^{A_{ih}^c(t)} z_i^{A_{ih}^c(t)} + \mu^{A_{ih}^c(t)}|$, where $z_i^{A_{ih}^c(t)}$ is drawn independently from a standard normal distribution for any given customer $i$ who receives the announcement!$A_{ih}^c(t)$. The parameters $\mu^{A_i^c(t)}$ and $\sigma^{A_i^c(t)}$ are the mean and standard deviation, respectively, of the underlying normal distribution for the customers who receive the announcement $A_i^c(t)$.

We now turn to estimate all the different models we construct to explain the customers' abandonment behavior and its relationship to the delay announcements.

## 5. Estimation Strategy

In this section, we go over the estimation strategy for all the models constructed above. Specifically, we use the maximum likelihood estimation method to estimate the parameters in our dynamic structural models. The key idea is to first compute the likelihood of observing a sequence of actions (i.e., wait/abandon) for each customer and then to compute the overall likelihood of all customers. We estimate the model parameters by optimizing the overall likelihood function, which best explains the observed behavior.

To facilitate the estimation process, we use the announcement system constructed in §3.3. We assume that announcements from the same interval induce the same belief about the offered waiting time and impact customers' costs in an identical manner, (e.g., announcements that report anticipated delay from 30 to 60 seconds induce the same belief about the offered waiting time and impact customers' costs in an identical manner). Now we describe the estimation strategy used for the models we constructed in §4. For all the models, we first estimate the service probability $\kappa_t^{A_{ih}(t)}$ directly from the data. To this end, we assume the residual offered waiting time at time period $\tau_j$ when a customer receives an announcement, for $j \in \{1,2,3\}$, is Gamma distributed[15] with shape and scale parameters that depend on the announcements. (The residual offered waiting time at time period $\tau_j$, for $j \in \{1,2,3\}$, is simply equal to the offered waiting time minus $\tau_j$.) We use the censored maximum likelihood method to estimate these parameters. Using these estimates for $\kappa_t^{A_{ih}(t)}$, we next estimate the parameters of the model. We begin by estimating the models that are in line with the theoretical literature, i.e., the W-updating models, in §5.1. Recall that these models assume that announcements impact only customers' beliefs about their offered waiting time. In §5.2, we continue with the estimations for the models that assume that announcements not only impact customers' beliefs about the system but also their waiting costs, i.e., WC-updating models. We will then describe the identification strategy for both the W-updating and the WC-updating models in Online Appendix F.

### 5.1. Estimation of the W-Updating Models

To construct the likelihood function, we first characterize the probabilities of customers' choice decisions about whether to wait or abandon at each time period $t$. We define $P_{iht}^{A_{ih}(t)}(a_{iht}; r_{ih}, c_{ih}, \theta)$ as the probability that customer $i$, whose last announcement received is $A_{ih}(t)$ by time period $t$ during her $h$th call, chooses action $a_{iht}$ at time period $t$. Using the corresponding choice probability at each time period, we obtain the probability of observing the sequence of choices over time for each customer. The following proposition characterizes the probabilities of customers' choice decisions. The detailed proof is outlined in Ben-Akiva and Lerman (1985, §5.2).

**Proposition 2.** *Let the idiosyncratic shock $\varepsilon_{iht}(a_{iht})$ be i.i.d. Gumbel distributed. With the utility function given by (2), we have*

$$P_{iht}^{A_{ih}(t)}(a_{iht}; r_{ih}, c_{ih}, \theta) = \frac{\exp(\Psi_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta, a_{iht})/\theta)}{1+\exp(\Psi_{iht}^{A_{ih}(t)}(r_{ih}, c_{ih}, \theta, wait)/\theta)}, \quad \forall\, a_{iht} \in \{wait, abandon\}. \tag{10}$$

As we mentioned in §4, the time horizon is divided into small time periods of equal length. In our estimation, we choose the length of each time period, denoted by $t^l$, to be 5 seconds.[16] We denote customer $i$'s waiting time during her $h$th call in the queue, as observed in the data, by $t_{ih}^d$. Then the total number of time periods that customer $i$ waits in the queue during her $h$th call, denoted by $N_{ih}$, is given by $N_{ih} = \lfloor t_{ih}^d/t^l \rfloor + 1$. As we mentioned at the beginning of this section, the total log likelihood of observing the data, denoted by $L_W$, is equal to the sum of the log likelihood of observing the sequence of customer $i$'s actions $a_{iht}$ during her $h$th call for $h = 1, 2, \ldots, H_i$ over $i = 1, 2, \ldots, K$, where $H_i$ is the total number of calls that customer $i$ made to the call center during the time when we collected our data and $K$ is the total number of calls in the data. Thus, the total log likelihood $L_W$ is given by $L_W(\mu, \sigma, \theta) = \sum_{i=1}^{K} \ln \mathbb{E}[\prod_{h=1}^{h=H_i} \prod_{t=0}^{t=N_i} P_{iht}^{A_{ih}(t)}(a_{iht}; r_{ih}, c_{ih}(Z), \theta)]$, where $c_{ih}(Z) = |\mu + \sigma Z|$ and $Z \sim \mathrm{N}(0,1)$. It is important to note that when we multiply the parameters $r_{ih}$, $c_{ih}$, and $\theta$ by any given constant, we will have the same likelihood value for the observed customers' abandonment decisions. Thus, to resolve the identification issue, we normalize the reward $r_{ih}$ to be 1 for all customers (see more details in Online Appendix F). To estimate the parameters $\omega = (\mu, \sigma, \theta) \in \Omega = \{(\mu, \sigma, \theta)\colon\ \mu \in R, \sigma \in R_+, \theta \in R_+\}$, where $\Omega$ is the feasible set of the parameters, is equivalent to solving the optimization problem as follows:

$$\max_{\omega \in \Omega} L_W(\omega) = \sum_{i=1}^{K} \ln \int \prod_{h=1}^{h=H_i} \prod_{t=1}^{t=N_{ih}} P_{iht}^{A_{ih}(t)}(a_{iht}; r_{ih}, c_{ih}(y), \theta)\phi(y)\, \mathrm{d}y, \tag{11}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution.

We solve the optimization problem above using the nonlinear optimization solver Knitro in MATLAB. We use 30 randomly generated initial points to get the best estimates. To estimate the integral in (11), we use Gauss–Hermite quadrature with 10 nodes outlined in §7.2 of Judd (1998). According to Akşin et al. (2012), the estimation method can recover the true parameters very well. To construct the confidence interval of each parameter, we use the *nonreplacement subsampling*; see Horowitz (2001). In particular, we divide the data into subgroups based on the first announcements that customers receive. We randomly sample 95% of the data without replacement from each subgroup. Then we aggregate the sampled data from each subgroup together to form a new data set. We generate 100 data sets in total using the procedure we just described. We estimate the parameters for each data set with the estimation strategy described above and compute the corresponding standard deviations. Using the estimated standard deviations, we construct 95% confidence interval for each parameter assuming that the estimates from the 100 data sets follow a normal distribution.

### 5.2. Estimation of WC-Updating Models

Above, we described the estimation strategy used for the W-updating models. Now, we turn to estimate the WC-updating models. In these models, given that we allow both the waiting cost and the idiosyncratic error to depend on the announcements, the optimization problem suffers from high dimensionality. Thus, it is computationally challenging to solve the one-shot optimization problem directly in a manner similar to the W-updating models. To reduce the computational complexity, we decompose the optimization problem, accounting for the various announcements customers receive over time. For concreteness, throughout this section, we focus on the models where only the first announcements impact customers' per unit waiting cost, i.e., $W_l C_1$ with $l \in \{1, 2, 3\}$. It is worth mentioning that the estimation strategy for the other models where both the first announcements and the subsequent announcements may impact customers' per unit waiting cost, i.e., $W_2C_2$, $W_3C_2$, and $W_3C_3$, is slightly different. We refer the reader to Online Appendix E for the details on the estimation strategy for these models.

Following Proposition 2 and yet adjusting the utility and nominal utility to account for the fact that the waiting cost and the idiosyncratic error depend on the announcements (see (7) and (8)), the customer's choice probability $P_{iht}^{A_{ih}(t)}(a_{iht}; r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}(Z), \theta^{A_{ih}^c(t)})$ is given by

$$P_{iht}^{A_{ih}(t)}(a_{iht}; r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}(Z), \theta^{A_{ih}^c(t)}) = \frac{\exp(\Psi_{iht}^{A_{ih}(t)}(r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}(Z), \theta^{A_{ih}^c(t)}, a_{iht})/\theta^{A_{ih}^c(t)})}{1+\exp(\Psi_{iht}^{A_{ih}(t)}(r_{ih}^{A_{ih}^c(t)}, c_{ih}^{A_{ih}^c(t)}(Z), \theta^{A_{ih}^c(t)}, \mathrm{wait})/\theta^{A_{ih}^c(t)})}, \quad \forall\, a_{iht} \in \{\mathrm{wait}, \mathrm{abandon}\}. \tag{12}$$

In addition, the corresponding cost is given by $c_{ih}^{A_{ih}^c(t)}(Z) = |\mu^{A_{ih}^c(t)} + \sigma^{A_{ih}^c(t)} Z|$, with $Z \sim \mathrm{N}(0,1)$. As in the previous section, we normalize the reward $r_i$ to be 1 for all customers.

Given that we allow both the waiting cost and the idiosyncratic error to be impacted by the announcements, we denote the parameters associated with a

given announcement $m \in M$ as $\omega^m = (\mu^m, \sigma^m, \theta^m)$. Our goal is to estimate the parameters $\omega^m \in \Omega^m = \{(\mu^m, \sigma^m, \theta^m)\colon \mu^m \in R, \sigma^m \in R_+, \theta^m \in R_+\}$ for all $m \in M$, where $\Omega^m$ is the feasible set of the parameters associated with the announcement $m \in M$. Given the large number of parameters involved, it is computationally challenging to estimate them together. As part of our estimation strategy, we decompose the total log likelihood function into different components based on the last announcement that the customer receives that impacts her waiting costs. To describe the decomposition, we denote the total log likelihood function of observing the sequence of actions for all customers by $L_{WC}(\omega^{A^c_{ih}(t)}\colon A^c_{ih}(t) \in M)$. Meanwhile, we let $L^m_{WC}(\omega^m)$ be the log likelihood of observing customers' choices at all the time periods, for which the last announcement (that impacts customers' costs) is fixed at $m$. Furthermore, we assume that customers' cost–reward ratios associated with different announcements are independent. The following proposition formalizes the decomposition.

**Proposition 3.** *The total log likelihood function* $L_{WC}(\omega^{A^c_{ih}(t)}\colon A^c_{ih}(t) \in M)$ *can be decomposed as follows*:

$$L_{WC}(\omega^{A^c_{ih}(t)}\colon A^c_{ih}(t) \in M) = \sum_{m \in M} L^m_{WC}(\omega^m).$$

Using the same arguments as in §5.1, the log likelihood function $L^m_{WC}(\omega^m)$ is expressed by

$$L^m_{WC}(\omega^m) = \sum_{i=1}^{K} \ln \int \prod_{h=1}^{h=H_i} \prod_{\{1 \le t \le N_{ih} \mid A^c_{ih}(t)=m\}} \cdot P^m_{iht}(a_{iht}; r^m_{ih}, c^m_{ih}(y), \theta^m)\phi(y)\,\mathrm{d}y. \quad (13)$$

Following the decomposition shown above, we can estimate $\omega^m \in \Omega^m$ independently, for each given $m \in M$, by maximizing the log likelihood function (13). For each announcement, we solve the optimization problems and estimate the confidence intervals for the parameters in the WC-updating models with the same approach that we used for the W-updating models. When estimating these models, for simplicity, we assume that within each class customers are homogeneous. We will test the robustness of our results to this assumption in §7.

## 6. Results

Above, we explained our estimation strategy; we next present our estimation results. We first report the results for the W-updating models and then the results for the WC-updating models. Having multiple models that explain the observed customers' abandonment behavior, we discuss the concept of model selection in §6.3. In particular, we use an internal consistency criterion and the Akaike information criterion (Burnham and Anderson 2002) to select the model that best explains customers' abandonment behavior.

**Table 1.** Results of the W-Updating Models $W_3C_0$, $W_2C_0$, and $W_1C_0$ and the Model $W_0C_0$ for the Regular Class

| | Cost–reward ratio (per min) | | |
|---|---|---|---|
| Model | Mean | SD | $\theta$ |
| $W_0C_0$ | 0.077 | 0.058 | 0.165 |
| $W_1C_0$ | 0.021 | 0.016 | 0.182 |
| $W_2C_0$ | 0.019 | 0.014 | 0.183 |
| $W_3C_0$ | 0.020 | 0.015 | 0.183 |

Our key findings are that announcements not only impact customers' beliefs about the system but also directly impact their per-unit waiting costs. Specifically, we show that customers' per-unit waiting cost decreases with the offered waiting time associated with the announcement.

### 6.1. Results for W-Updating Models

Table 1 shows the estimation results of the W-updating models for the regular class. Our results show that customers from the regular class are heterogeneous in their cost–reward ratios. In other words, the results imply that the different abandonment times that we observe are not only due to the idiosyncratic shocks but also due to the different waiting costs of different customers. We obtain similar results for the VIP class; see the details in Online Appendix H, Table 13. To test the assumption that delay announcements do not directly impact customers' per-unit waiting cost, we turn to the results of the WC-updating models.

### 6.2. Results for WC-Updating Models

In this section, we report the results for the WC-updating models. Our key results for these models are that the announcements not only impact customers' beliefs about the system but also have significant impact on their waiting costs. Furthermore, customers' per-unit waiting cost decreases with the offered waiting time associated with the announcement.

To test the consistency of the assumption made in the W-updating models that announcements only impact customers beliefs about the system, we start with reporting the estimation results for the models $W_lC_1$, with $l \in \{1,2,3\}$. Recall that both models $W_lC_1$ and $W_lC_0$, for a given $l \in \{1,2,3\}$, assume that customers keep updating their beliefs about the system based on the first $l$ announcements over time. The only difference between the two models is that the $W_lC_1$ model assumes that the first announcements impact customers' waiting costs, whereas model $W_lC_0$ assumes that the waiting cost is impacted neither by the initial announcement nor the subsequent ones. However, the estimation results for the models $W_lC_1$, with $l \in \{1,2,3\}$, given in Table 2, show that the cost–

**Table 2.** Estimation Results of the Models $W_3C_1$, $W_2C_1$, and $W_1C_1$ for the Regular Class

| | Models | | | | | |
|---|---|---|---|---|---|---|
| | $W_3C_1$ | | $W_2C_1$ | | $W_1C_1$ | |
| First announcement | Cost–reward ratio (per min) | $\theta$ | Cost–reward ratio (per min) | $\theta$ | Cost–reward ratio (per min) | $\theta$ |
| [0, 30s) | 1.270 | 0.028 | 1.270 | 0.028 | 1.270 | 0.028 |
| [30, 60s) | 0.965 | 0.074 | 0.965 | 0.074 | 0.965 | 0.074 |
| [60, 90s) | 0.374 | 0.136 | 0.340 | 0.141 | 0.427 | 0.125 |
| [90, 120s) | 0.249 | 0.147 | 0.240 | 0.148 | 0.309 | 0.133 |
| [120, 150s) | 0.025 | 0.184 | 0.024 | 0.184 | 0.037 | 0.181 |
| [150, 180s) | 0.029 | 0.181 | 0.030 | 0.180 | 0.033 | 0.179 |
| [180, 210s) | 0.021 | 0.180 | 0.020 | 0.181 | 0.022 | 0.180 |
| [210, 240s) | 0.017 | 0.181 | 0.011 | 0.183 | 0.018 | 0.181 |
| [240, 360s) | 0.008 | 0.187 | 0.007 | 0.188 | 0.002 | 0.192 |

reward ratio decreases with the offered waiting time associated with the first announcement, whereas the variance of the idiosyncratic shock[17] increases. This contradicts the assumption made in the corresponding $W_lC_0$ model, for $l \in \{1,2,3\}$, that announcements do not impact customers' waiting costs. Thus, we claim that the W-updating models are not internally consistent.

At this point, it is important to note that the models $W_lC_1$ with $l \in \{1,2,3\}$ shown above all assume that only the first announcements impact customers' waiting costs. To investigate the extent to which the subsequent announcements impact customers' waiting costs, we explore all the possible WC-updating models, i.e., $W_lC_n$ with $l \geq n$ and $l, n \in \{1,2,3\}$. We present the estimation results for the $W_3C_3$ model here. Table 3 shows that the cost–reward ratio decreases with the offered waiting time associated with the first announcement, whereas the variance of the idiosyncratic shock increases in the $W_3C_3$ model. These results are consistent with the results observed in the models $W_lC_1$, with $l \in \{1,2,3\}$. In addition, the same trend exists for the cost–reward ratio and the variance of the idiosyncratic shock associated with the second and third announcements. We also obtain similar results for the $W_2C_2$ and $W_3C_2$ models (see Online Appendix G, Tables 10 and 11) as well as for the VIP class (see Online Appendix H, Tables 14–17). We will report the confidence intervals for the model that best explains customers' abandonment behavior in Table 6 (see §7).

### 6.3. Model Comparisons and Discussion

One can use the internal consistency criterion to favor the WC-updating models over the W-updating models

**Table 3.** Estimation Results of the Model $W_3C_3$ for the Regular Class

| | First announcement | | Second announcement | | Third announcement | |
|---|---|---|---|---|---|---|
| Announcement[a] | Cost–reward ratio (per min) | $\theta$ | Cost–reward ratio (per min) | $\theta$ | Cost–reward ratio (per min) | $\theta$ |
| [0, 30s) | 1.270 | 0.028 | 0.065 | 0.175 | 0.024 | 0.187 |
| [30, 60s) | 0.965 | 0.074 | | | | |
| [60, 90s) | 0.034 | 0.180 | 0.065 | 0.175 | | |
| [90, 120s) | 0.034 | 0.180 | | | | |
| [120, 150s) | 0.034 | 0.180 | 0.065 | 0.175 | | |
| [150, 180s) | 0.034 | 0.180 | | | | |
| [180, 210s) | 0.021 | 0.177 | 0.035 | 0.178 | | |
| [210, 240s) | 0.021 | 0.177 | | | | |
| [240, 360s) | 0.021 | 0.177 | 0.035 | 0.178 | 0.024 | 0.187 |

*Note.* We assume that customers' cost–reward ratios and the parameters $\theta$ induced by the first announcements [60, 90*s*), [90, 120*s*), [120, 150*s*), and [150, 180*s*) are the same; customers' per unit waiting costs and the parameters $\theta$ induced by the first announcements [180, 210*s*), [210, 240*s*) and [240, 360*s*) are the same; customers' per unit waiting costs and the parameters $\theta$ induced by the second announcements [0, 60*s*), [60, 120*s*) and [120, 180*s*) are the same; customers' per unit waiting costs and the parameters $\theta$ induced by the second announcements [180, 240*s*) and [240, 360*s*) are the same; and customers' per unit waiting costs and the parameters $\theta$ induced by all the third announcements are the same, while customers continue to form beliefs about the system based on the default announcement system.

[a]The intervals for the second and the third announcements are different from those for the first announcements to balance the number of intervals and the number of customers in each interval as we outlined in §3.3.

**Table 4.** Relative AIC Scores of All Models for the Regular Class (a) and the VIP Class (b)

| | $C_0$ | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|---|
| $W_0$ | −37.4 | | | |
| $W_1$ | −25.6 | −26.3 | | |
| $W_2$ | −15.4 | −12.1 | −10.6 | |
| $W_3$ | 0.0 | −3.5 | −0.9 | −0.4 |

(a)

| | $C_0$ | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|---|
| $W_0$ | −23.1 | | | |
| $W_1$ | −20.3 | −25.5 | | |
| $W_2$ | −17.8 | −13.9 | −16.4 | |
| $W_3$ | 0.0 | −13.8 | −0.9 | 5.3 |

(b)

*Notes.* The AIC scores of the base model $W_3C_0$ for both classes are normalized to be 0. Further, the row index corresponds to the number of announcements that customers keep updating the beliefs with over time, while the column index corresponds to the number of announcements that impact customers' waiting costs over time in the corresponding model.

as we discussed above. In this section, we use a statistical approach based on the Akaike information criterion (AIC) to select the model that best explains customer behavior. Given that AIC is a relative measure for the quality of a statistical model by trading off between the goodness of fit and the complexity of the model, we compare the AIC score of each model to the base model $W_3C_0$. The AIC score of model $W_3C_0$ is normalized to be zero for convenience. Tables 4(a) and 4(b) report the relative AIC score of each model for the regular class and the VIP class, respectively. According to Burnham and Anderson (2002), the smaller the relative AIC score is, the less information loss the model has compared with the base model $W_3C_0$. Thus, according to Tables 4(a) and 4(b), model $W_1C_0$ best explains the customer behavior among the W-updating models, whereas model $W_1C_1$ best explains the customers' behavior among the WC-updating models. Moreover, $W_1C_1$ explains the data significantly better than $W_1C_0$ for the VIP class, whereas $W_1C_1$ and $W_1C_0$ are not significantly different in explaining the data for the regular class. Hence, model $W_1C_1$ dominates $W_1C_0$ using both criteria, i.e., the internal consistency criterion and also the AIC.

Note that $W_1C_1$ has a relatively large number of parameters (i.e., a marginal cost and $\theta$ for announcements from each interval). One may envision that a model with fewer parameters will perform better in terms of the AIC score. We observe (see Table 2) that customers' per-unit waiting costs are identical over some consecutive announcement intervals in the $W_1C_1$ model; e.g., the per-unit waiting cost associated with the announcements $[180, 210s)$ and the one associated with the announcements $[210, 240s)$ are almost identical to each other. By assuming that the per-unit waiting costs induced by these announcements are the same, we obtain a revised model, referred to as the $W_1C_1^*$ model. The details about the announcement intervals, which induce the same per-unit waiting cost, are described in Online Appendix G, Table 12. The results show that the AIC scores of the revised $W_1C_1^*$ model for the regular class and the VIP class are −6.4 and −8.9 relative to $W_1C_0$, respectively, which implies that $W_1C_1^*$ is significantly better than $W_1C_0$ based on the AIC score for both customer classes. We report the detailed results of the $W_1C_1^*$ model for the regular class and the VIP class in Table 12 of Online Appendix G and Table 21 of Online Appendix H, respectively.

Note that the models we discussed above all assume that at least the first announcements impact customers' beliefs about the system. To verify this assumption, we investigate a model that assumes that none of the announcements impacts customers' beliefs about the system or their waiting costs. Along the nomenclature outlined in §4, we denote this model as $W_0C_0$. We report the estimation results of $W_0C_0$ for the regular class and the VIP class in Table 1 and Online Appendix H, Table 13, respectively.

Before we compare the $W_0C_0$ model to the W-updating and WC-updating models mentioned above, we first verify whether $W_0C_0$ is internally consistent. It is important to note that there are two important assumptions under the $W_0C_0$ model: (1) delay announcements do not impact customers' beliefs about the offered waiting time, and (2) delay announcements do not impact customers' per-unit waiting costs. To verify the second assumption, we now estimate the cost–reward ratios of customers receiving different announcements separately, while still assuming delay announcements do not impact customers' belief about the offered waiting time. We report the detailed estimation results for regular customers in Table 5. (Following the nomenclature outlined in §4, we refer to this model as $W_0C_1$.) If the assumption that delay announcements do not impact customers' per-unit waiting cost holds in the $W_0C_0$ model, we would expect the estimated cost–reward ratios for customers receiving different announcements to be not significantly different. However, on the basis of the results in Table 5, we see that the cost–reward ratios for regular customers receiving different announcements follow different distributions. In particular, the cost–reward ratio of the regular customers decreases with the offered waiting time associated with the delay announcements in terms of the mean overall. Thus, we can claim that the $W_0C_0$ model is not

**Table 5.** Results of $W_0C_1$ and $W_1C_1$ with Heterogeneous Customers for the Regular Class

| | Models | | | | | |
|---|---|---|---|---|---|---|
| | $W_0C_1$ | | | $W_1C_1$ | | |
| | Cost–reward ratio (per min) | | | Cost–reward ratio (per min) | | |
| Announcement | Mean | SD | $\theta$ | Mean | SD | $\theta$ |
| [0, 30s) | 0 | 0 | 0.185 | 1.359 | 0.044 | 0.015 |
| [30, 60s) | 0.255 | 0 | 0.115 | 0.965 | 0 | 0.074 |
| [60, 90s) | 0.204 | 0 | 0.128 | 0.427 | 0 | 0.125 |
| [90, 120s) | 0.156 | 0.094 | 0.130 | 0.184 | 0.138 | 0.145 |
| [120, 150s) | 0.071 | 0.053 | 0.167 | 0.046 | 0.034 | 0.179 |
| [150, 180s) | 0.089 | 0.066 | 0.159 | 0.038 | 0.028 | 0.178 |
| [180, 210s) | 0.090 | 0.068 | 0.158 | 0.026 | 0.019 | 0.179 |
| [210, 240s) | 0.082 | 0 | 0.165 | 0.017 | 0 | 0.181 |
| [240, 360s) | 0 | 0 | 0.191 | 0.002 | 0 | 0.192 |

internally consistent for regular customers. We obtain similar results for the VIP customers; see the detailed results in Online Appendix H, Table 18.

Other than the internal consistency criterion that we used above, we now use the AIC to compare the $W_0C_0$ model to the $W_1C_1^*$ model, which is the model that best explains the data among all the W-updating and WC-updating models. Our results show that $W_1C_1^*$ is significantly better than $W_0C_0$ along the line of the AIC for the VIP class, whereas $W_0C_0$ is significantly better than $W_1C_1^*$ for the regular class. In particular, the AIC scores of the $W_0C_0$ model for the VIP class and the regular class are 6.2 and −5.4 relative to $W_1C_1^*$, respectively. To this end, it is important to note that the AIC reflects the trade-off between the goodness of fit of the model measured by the maximized log-likelihood value and the number of parameters in the model. Thus, although the $W_0C_0$ model outperforms the $W_1C_1^*$ model in terms of the AIC, the goodness of fit of the $W_1C_1^*$ model is still significantly better than the $W_0C_0$ model. One possible explanation is that VIP customers are more experienced with the call center compared with regular customers. Therefore, VIP customers are more capable of making abandonment decisions based on the announcements received, whereas the regular customers may have difficulty accounting for the announcements. Thus, although the model $W_0C_0$, which assumes that customers do not take delay announcements into account, is not internally consistent, it appears to be more efficient in explaining the observed abandonment behavior of the regular customers based on the AIC score compared with the $W_1C_1^*$ model.

The analysis above shows that, for VIP customers, delay announcements impact not only customers' beliefs about the offered waiting time but also the cost–reward ratio and the variance of the idiosyncratic shock. In particular, we find that the cost–reward ratio of the VIP customers decreases with the offered waiting time associated with the announcement, whereas the variance of the idiosyncratic shock increases. In addition, we find that the model assuming that only the first announcements impact customers' beliefs about the system and their waiting costs explains VIP customers' behavior better than any other models we tested based on both the internal consistency criterion and the AIC scores. One possible explanation for the result that longer estimated delays induce lower per-unit waiting cost is as follows: when the estimated delay is long, customers anticipate longer waiting; given that waiting time is long, customers can choose an activity to engage in from a larger set of activities while waiting, compared with the case when the waiting time is short. This effectively reduces their opportunity cost of waiting and results in a smaller per-unit waiting cost. For example, when the waiting time is anticipated to be 5 minutes, one can respond to emails while waiting, yet an anticipated wait of 30 seconds may require the customer to stay idle while waiting in the queue. This explanation is in line with the theory in Becker (1965) that the value of the time is equal to the opportunity cost.

As for the results that the variance of the idiosyncratic shock increases with the offered waiting times associated with the delay announcements, one possible explanation is that customers' adherence to full rationality diminishes as the anticipated delay increases. The result that customers do not update their beliefs about the system or their waiting costs with the subsequent announcements may not be generalizable, since the call center provides information about the state of the system rather then the progress.

Considering only the models that are internally consistent, similar results hold for the regular customers. Recall that in terms of the amount of information needed to explain how regular customers behave, it might be sufficient to assume that regular customers disregard the announcements.

These results from the structural model can be used to shed light on the link between delay announcements and our preliminary results about patience time. Customers receiving longer announcements are subject to longer total waiting time yet smaller per-unit waiting cost. The combination of the two opposite effects results in the phenomena we observe in §3.3, that the announcements impact customers' patience in a nontrivial manner or sometimes do not have an impact on customers' patience time at all.

## 7. Robustness Study

We now study the robustness of our results that customers' per-unit waiting cost decreases with the offered waiting times associated with the announcements. Note that, so far, the paper has made the following two assumptions:[18] (1) we have ignored the "time-of-the-day" effect, and (2) we have assumed that customers make abandonment decisions based on the actual offered waiting time associated with the announcements instead of the literal meaning of the announcements. In this section, we first explore a model in which we control for the time of the day when the call was made. The results show that our key insight continues to hold even when we control for the time-of-the-day effect. We next investigate an alternative model, where customers make decisions based on the explicit estimated delays provided in the announcements. Although our key results continue to hold, our analysis shows that this model has a lower predictive power compared with the $W_1C_1$ model. Finally, we study a model where we allow the cost–reward ratio to be heterogeneous among customers. Again, although our main results continue to hold, the model under customer heterogeneity has significantly lower predictive power compared with the ones with homogeneous customers. Before we provide the details of these extensions, we first confirm the statistical significance of our key results by reporting the confidence intervals of the point estimates.

### 7.1. Statistical Significance

In Table 2, we presented the point estimates for the $W_1C_1$ model. We observed that the customers' per-unit waiting cost decreases with the offered waiting time associated with the announcement, whereas the variance of the idiosyncratic error increases. To check the statistical significance of these results, we report the 95% confidence intervals for the corresponding parameters in $W_1C_1$ model for the regular class in Table 6.[19] As one can see, the trends for both the per-unit waiting cost and the variance of the idiosyncratic error hold with 95% confidence. We observe similar results for the VIP class (see the details in Online Appendix H, Table 20).

**Table 6.** Estimation Results of $W_1C_1$ Model for the Regular Class with 95% Confidence Intervals in Square Brackets for the Corresponding Parameters

| First announcement | Cost–reward ratio (per min) | $\theta$ |
|---|---|---|
| [0, 30s) | 1.161 [0.487, 1.835] | 0.041 [0, 0.120] |
| [30, 60s) | 0.965 [0.897, 1.034] | 0.074 [0.066, 0.083] |
| [60, 90s) | 0.421 [0.340, 0.502] | 0.126 [0.114, 0.137] |
| [90, 120s) | 0.309 [0.276, 0.342] | 0.133 [0.127, 0.138] |
| [120, 150s) | 0.037 [0.014, 0.061] | 0.181 [0.176, 0.186] |
| [150, 180s) | 0.034 [0.022, 0.046] | 0.179 [0.176, 0.182] |
| [180, 210s) | 0.022 [0.013, 0.032] | 0.180 [0.178, 0.183] |
| [210, 240s) | 0.018 [0.008, 0.029] | 0.181 [0.177, 0.185] |
| [240, 360s) | 0.003 [0, 0.008] | 0.192 [0.189, 0.195] |

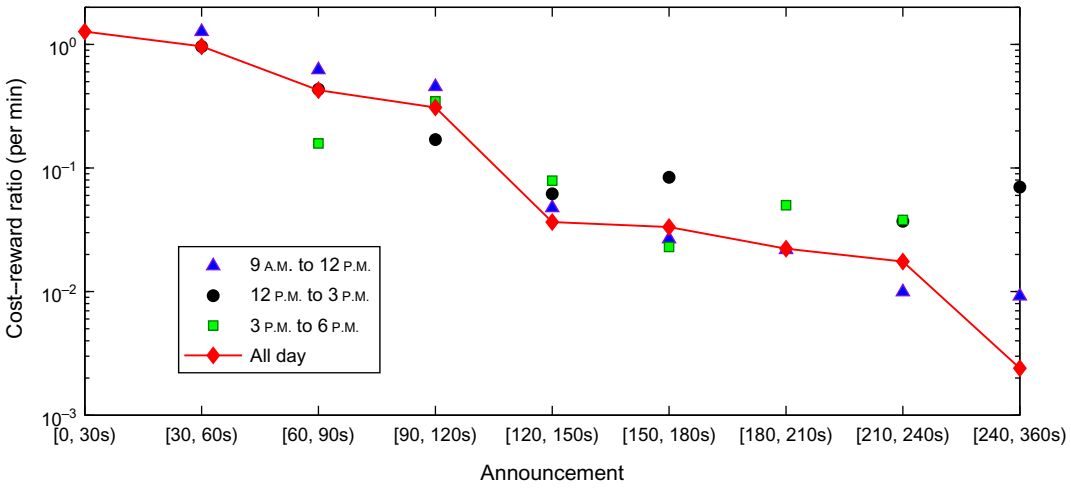
### 7.2. Time-of-the-Day Effect

We now turn to investigate the time-of-the-day effect and its impact on our results. In particular, our goal is to test whether the impact of the announcements on waiting cost persists regardless of the time of the day, in terms of both the order of magnitude and the trend. To do so, we first divide each day into three time segments, i.e., 9 A.M. to 12 P.M., 12 P.M. to 3 P.M., and 3 P.M. to 6 P.M.[20] We then estimate the model for each time segment separately. Figure 3 shows the estimated cost–reward ratios of the $W_1C_1$ model for all three time segments, juxtaposed with the cost–reward ratio of the $W_1C_1$ model reported in Table 2. We observe that the trend of the cost–reward ratio is preserved for all three time segments. Moreover, the estimated cost–reward ratio for the three time segments are all fairly close to the corresponding cost–reward ratio estimated with the whole data set. Similar results are obtained for the variance of the idiosyncratic error. Thus, the implication of these results is that our key insights continue to hold even when we control for the time-of-the-day effect.

### 7.3. Beliefs About the Offered Waiting Time

It is important to note that we have made two assumptions on customers' beliefs about the offered waiting time so far: (1) we assume that customers make abandonment decisions based on the actual offered waiting time associated with the announcement, and (2) we assume that customers form beliefs about the offered waiting time based on the last announcements that they receive. To test the robustness of our results to the first assumption, we explore an alternative model that is identical to the $W_1C_1$ model with the modification that customers make decisions based on the explicit delay estimates provided in the announcements. We refer to this alternative model as the Naïve-$W_1C_1$ model. Specifically, in the Naïve-$W_1C_1$ model, customers believe that the mean of the offered waiting time equals the delay estimates given in the announcements. We compare the Naïve-$W_1C_1$ model to the $W_1C_1$

**Figure 3.** (Color online) Comparison Among the Cost–Reward Ratios of $W_1C_1$ Model for the Three Data Subsets and the Cost–Reward Ratio of the $W_1C_1$ Model Estimated Using the Whole Data Set for the Regular Class




model based on the AIC score. Our results show that the AIC scores of the Naïve-$W_1C_1$ model increase by 3.6 and 3.7 relative to the $W_1C_1$ model for the regular and VIP customers, respectively. This implies that the $W_1C_1$ model, where customers make decisions based on the actual offered waiting time associated with the announcements, explains the data significantly better than the Naïve-$W_1C_1$ model where customers make decisions based on the literal meaning of the corresponding announcement. To study the robustness of our results to the second assumption, we now investigate an alternative model that is identical to the $W_3C_1$ model, with the modification that customers form beliefs about the offered waiting time based on the entire sequence of the announcements that they receive in the current transaction. Although our key insight still holds in this alternative model, the AIC score shows that the $W_1C_1$ model, which is the model that best explains the data under the assumption that customers form beliefs about the offered waiting time based on the last announcements that they receive, explains the observed customers' abandonment behavior significantly better than this alternative model.

### 7.4. Heterogeneity of Customers

As we pointed out earlier, we estimated all the aforementioned WC-updating models under the assumption that within the same class, the cost–reward ratios are the same for customers who receive the same announcement; i.e., $\sigma^m = 0$ for all $m \in M$. We now study a model where we allow customers' cost–reward ratios to be heterogeneous. To operationalize such heterogeneity, we allow the standard deviation of the folded normal distribution to be strictly positive in the WC-updating models. As a result, the models accounting for heterogeneous customers have additional parameters (specifically, the standard deviation $\sigma^m$, $\forall\, m \in M$) compared with the models with homogeneous customers. We report the estimated results of the $W_1C_1$ model with heterogeneous customers for the regular class in Table 5. We find that our key insights continue to hold under customer heterogeneity. In particular, the same trends still hold for the per-unit waiting cost (in terms of the mean) and the variance of the idiosyncratic shock under heterogeneous customers for the regular class. Because of the increased number of parameters, however, the AIC score of the $W_1C_1$ model with heterogeneous customers increases by 17.1 relative to $W_1C_1$ model with homogeneous customers for the regular class. This implies that the $W_1C_1$ model with homogeneous customers explains the data significantly better than the $W_1C_1$ model with heterogeneous customers for the regular class. Similar results are obtained for the VIP customers (see the details in Online Appendix H, Table 19).

## 8. Counterfactual Analysis

We next demonstrate how one can use our methodology to help firms make better announcement policies by conducting a counterfactual analysis in a simulation approach. We first explore the impact of modifying the granularity of the announcements by evaluating the impact of different announcement systems on customers' surplus. Once we understand how firms can potentially maximize customers' surplus using delay announcements, we then turn to a more basic question of whether firms should provide announcements at all.

Before we address these two questions, let us briefly describe the essence of our call center model. While customers wait in the queue, the call center informs them about the anticipated delay based on the offered waiting time of the system. Customers then form beliefs about the offered waiting time based on the announcements received. Based on these beliefs, customers make

abandonment decisions, which in turn impact the waiting times experienced by other customers. For a given announcement system, we are interested in characterizing the equilibrium where the offered waiting times associated with the announcements equal the ones experienced by customers. We create an iterative procedure, which we use to complete the system equilibrium emerged under a given announcement system. We refer readers to Yu (2014, §2.8) for further details of the simulation model and the iterative procedure. In our counterfactual study, we focus on customers from the regular class and the VIP class who call the bank between 9 A.M and 6 P.M.

Let us start from the question regarding the impact of modifying the granularity of the announcements. We first construct a benchmark system that mimics the observed dynamics at the call center. In this system, the benchmark announcement system, as described in Table 7, is used. We consider three alternative announcement systems other than the benchmark system: announcement systems $\mathscr{A}_{\mathrm{I}}$, $\mathscr{A}_{\mathrm{II}}$, and $\mathscr{A}_{\mathrm{III}}$; see Table 7.[21] The simulation models used to study the three announcement systems are the same as the one used for the benchmark system, but with different announcement systems. In particular, regardless of whether we use the benchmark announcement system or an alternative announcement system, for each one, we use the $W_1C_1$ model, outlined in §4, to characterize customers' abandonment behavior. Thus, for our study, we provide delay announcements only once. The estimates of the anticipated delays provided in these announcements are computed using the same algorithm that the call center used. The iteration procedures converge for all these announcement systems. To compare the performances of different announcement systems, we use the average customers' surplus in the state of equilibrium as the performance measure for all announcement systems. The surplus of a customer equals the reward obtained minus her actual total waiting costs. We compute the average surplus of each customer class for each announcement system and the weighted average surplus that accounts for the relative number of customers in each class. For simplicity, we normalize the reward of being served to be one dollar for both customer classes. We report the results for all these announcement systems in Table 8. The results show that the overall weighted average surplus under announcement system $\mathscr{A}_{\mathrm{I}}$ is slightly better than the ones under other announcement systems. In particular, compared with the benchmark system, the average surplus for the VIP customers under announcement system $\mathscr{A}_{\mathrm{I}}$ increases by 7%, and the average surplus for the regular customers decreases by 2%, whereas the overall weighted average surplus increases by 4%.

**Table 7.** Benchmark Announcement System and the Three Alternative Announcement Systems

| Time intervals | Benchmark system | $\mathscr{A}_{\mathrm{I}}$ | $\mathscr{A}_{\mathrm{II}}$ | $\mathscr{A}_{\mathrm{III}}$ |
|---|---|---|---|---|
| [0, 30s) | Message 1 | Message 1 | Message 1 | Message 1 |
| [30, 60s) | Message 2 | | | |
| [60, 90s) | Message 3 | Message 2 | | |
| [90, 120s) | Message 4 | | | |
| [120, 150s) | Message 5 | | Message 2 | |
| [150, 180s) | Message 6 | | | |
| [180, 210s) | Message 7 | | | Message 2 |
| [210, 240s) | Message 8 | | | |
| [240, ∞) | | Message 3 | Message 3 | Message 3 |

**Table 8.** Average Customer Surplus Under Different Announcement Systems for Both Customer Classes

| Announcement system | Average surplus (per customer per reward) | | Weighted average surplus |
|---|---|---|---|
| | Regular class | VIP class | |
| $\mathscr{A}_{\mathrm{I}}$ | 0.323 | 0.736 | 0.498 |
| $\mathscr{A}_{\mathrm{II}}$ | 0.326 | 0.720 | 0.493 |
| $\mathscr{A}_{\mathrm{III}}$ | 0.323 | 0.722 | 0.492 |
| Benchmark | 0.329 | 0.687 | 0.481 |
| No announcements | 0.135 | 0.775 | 0.405 |

It is worth mentioning that more VIP customers who receive announcements with long estimated delay stay for service under announcement system $\mathscr{A}_{\mathrm{I}}$ compared with the benchmark announcement system. Yet there is no such significant difference for regular customers or customers who receive announcements with short estimated delay. Note that the surplus of each served VIP customer who receives announcements with long estimated delay is positive in expectation. Thus, we observe that the overall surplus of the VIP customers increases under the alternative announcement system, whereas the overall surplus of the regular customers seems unchanged; see Table 8. Therefore, the above explains the improvement in terms of the overall weighted customers' surplus under the alternative announcement system compared with the benchmark announcement system.

To address the second question, we study a system where no announcements are provided. Our iterative procedure also converges for the system with no announcements. We report the results in Table 8. Compared with the benchmark system, the average surplus of the regular customers in the system with no announcements decreases by 59%, and the average surplus of the VIP customers increases by 13%, whereas the overall weighted average surplus decreases by 16%. The results imply that informing customers about the anticipated delay improves customers' surplus overall. Moreover, regular customers can make much better decisions in terms of their average surplus when there are delay announcements provided, whereas the impact of delay announcements on the VIP customers

are not as significant as the regular customers in terms of their average surplus.

Our counterfactual analysis above suggests that this bank should provide delay announcements in order to improve the overall customers' surplus. Furthermore, it may not be necessary for the delay announcements to have very fine granularity. A simple announcement system that comprises three messages, i.e., low, medium, and high, could result in better performance of the system in terms of customers' surplus, compared with an announcement system with more refined granularity. We observe that, under the alternative announcement system, more VIP customers who receive long estimated delay stay for service, which may lead to improvement of the overall customer surplus under the alternative announcement system compared with the benchmark announcement system.[22]

## 9. Concluding Remarks

In this paper, we study how customers react to explicit and real-time information on anticipated delays at a call center of a medium-sized bank in Israel using an empirical approach. In contrast to the implicit assumption made in the literature that delay announcements only impact customers' beliefs about the system by informing them about the anticipated delay, our key results show that delay announcements not only impact customers' beliefs but also impact customers' waiting costs. In particular, the per-unit waiting cost decreases with the offered waiting time associated with the announcements, whereas the variance of the idiosyncratic shocks increases.

The evidence that we provide about the impact of delay announcements on the per-unit waiting costs should inform the practice on how delay announcements should be used to shape customer behavior. Most call centers perform survival analysis that is similar to the one that we conducted using the nonparametric approach in §3.3. These call centers may be led to believe that the delay announcements do not impact customers' behavior. Yet we show that such an approach is insufficient, and a richer analysis is needed to disentangle the main effects of delay announcements. In particular, our results show that delay announcements do impact customers' behavior. Moreover, one of the key managerial insights from our counterfactual study is that firms should inform customers about the anticipated delay to improve customers' surplus. The results also suggest that the call centers need not provide announcements with granularity as fine as the current system and that a simple announcement system including as few as three messages (e.g., low, medium, and high) may result in better performance in terms of customers' surplus.

Finally, our study has certain limitations that should be explored in future research. In this paper, we study the impact of informing customers on the anticipated delay in a structural estimation approach. To relax the partial rationality assumption that we imposed on how customers make abandonment decisions in the structural model, we have been investigating the impact of delay announcements in a reduced-form approach. This alternative framework can account for more factors, (e.g., sunk cost of waiting and the impact of the music being played while customers are waiting in the queue) that cannot be incorporated in our structural model. One can also extend this work by exploring the impact of other types of delay announcements, e.g., queuing position information. Moreover, the impact of better informing customers about their progress in the system is worth exploring. As we pointed out earlier, Munichor and Rafaeli (2007) show that providing a sense of progress can improve customers' satisfaction in an experimental approach. This question could not be addressed in the call center that we study, because the repeated announcements provided do not inform customers about the progress. It is worth exploring the impact of providing multiple announcements, where the subsequent announcements indeed provide information about the progress for customers, on customers' abandonment behavior. Moreover, recall that we study how announcements with different granularity impact customers' abandonment behavior differently in our counterfactual analysis. To this end, it is important to note that the level of anxiety and uncertainty experienced by customers under announcement systems with different granularity may be different. However, we cannot account for such potential psychological effect in our counterfactual study because we are confined by a specific announcement system provided by the call center when we collected the data. Thus, our results about the impact of announcement granularity in the counterfactual analysis may be confounded by this potential psychological effect. As a limitation to any empirical work, we are constrained by the context in which we obtained the data. However, we believe that the main insights of this paper are applicable to settings where calls are informational in nature as it was in the call center where we obtained the data. To study the applicability and generalizability of the key insights of this paper, as well as to address other questions mentioned above, we have been conducting a field study at a different call center of the same bank.

From a theoretical modeling point of view, one can explore theoretical models that account for the results—that announcements not only impact customers' beliefs about the system but also directly impact customers' per-unit waiting costs—to help firms make better operations decisions (e.g., staffing) for a given announcement system.

## Endnotes

[1] It is important to note that the bank differentiates the service types of each call in a very fine manner. In particular, the bank distinguishes the checking account–related services from other service types such as securities, sales, pension funds, investments, insurance, dues, credit, etc. Thus, we believe that customers within the same class are homogeneous in terms of their service types. Moreover, based on our conversation with the manager of the bank, we believe that most of the calls from the two classes that we focus on are informational in nature.

[2] We find that 0.21% of customers receive announcements longer than six minutes for the regular class, whereas there are 0.49% of customers receiving announcements longer than five minutes for the VIP class. Thus, we only consider customers receiving delay announcements equal to or shorter than six minutes for the regular class. For the VIP class, we only consider customers receiving delay announcements equal to or shorter than five minutes.

[3] Feigin (2006) gives an example of the traditional approach that call centers take to analyze their customers' abandonment behavior and the impact of delay announcements.

[4] To be more precise, the offered waiting time is not directly observable. We could compute the offered waiting time if we knew all the information required. However, it is not feasible with the data we have. Thus, being not directly observable is computationally equivalent to being right censored in our case.

[5] We further review the applicability of the Gehan's generalized Wilcoxon test and the log-rank test in Online Appendix C.

[6] We conduct the tests for the offered waiting time associated with each pair of announcements. The significance level for each test is 0.05; i.e., $\alpha = 0.05$.

[7] We elaborate on the connection between the estimated delay provided in the announcements and the offered waiting time in Online Appendix B.

[8] It is important to point out that, in our call center, the subsequent announcements do not inform customers about their estimated delay but the estimated delay of the customers who just arrived at the system. Thus, we expect the second announcements to be noninformative.

[9] The inequality $T_5^1 <_{\text{st}} T_3^1$ means that $T_3^1$ has first-order stochastic dominance over $T_5^1$.

[10] As we will mention in §5.1, we choose the length of the small time periods that we divide the time horizon into to be five seconds. Thus, we have $\tau_1 = 18$ (which corresponds to 90 seconds when the second announcement is provided) and $\tau_2 = 36$ (which corresponds to 180 seconds when the third announcement is provided).

[11] To test the robustness of our results to the assumption that customers form beliefs about the offered waiting time on the last announcement that they receive for any given call, we study an alternative model. In this model, customers form beliefs about the offered waiting time based on the entire sequence of the announcements that they receive in the current transaction.

[12] Our data show a customer calls about 14 times a year on average to this call center. It is reasonable to assume that such customers are capable of forming the correct beliefs about the system.

[13] We assume that the idiosyncratic shock follows a Gumbel distribution in order to use the analytical result of the classical logit mode. Such an assumption is commonly used in the economic literature.

[14] We assume that the cost–reward ratio follows a folded normal distribution for technical convenience. In particular, the Gaussian–Hermite approximation that we use to efficiently solve the estimation problem can only be applied if the underlying stochasticity emerges from a normal distribution. Moreover, the cost–reward ratios have to be nonnegative. Thus, we choose a folded normal distribution for it.

[15] The Gamma distribution is commonly used to model waiting time in econometrics and other application areas (e.g., clinical trials); see §3.3 of Hogg et al. (2005). Furthermore, based on our data analysis, the Gamma distribution fits our data the best compared with various other distributions that we have explored.

[16] The length of the time periods does not have to be exactly five seconds. In fact, on the basis of our sensitivity analysis, we find that our estimation results are robust to the choice of the length for the time period. However, the length of the time period should not be too long. Otherwise, we can not capture customer dynamic abandonment decisions as finely.

[17] The variance of the idiosyncratic shock is equal to $(\pi^2/6)\theta^2$.

[18] Other than the two assumptions that we mention at the beginning of §7, we have also ignored the time that customers spend in the IVR component. To this end, in Yu (2014), we have explored a model that accounts for the time that customers spend in the IVR component of the call. Our results confirm that the key insight still holds even when we account for the "time-in-IVR" effect.

[19] The manager at the call center told us the low cost–reward ratios for customers receiving longer announcements were totally expected.

[20] We specifically choose the three time segments, i.e., 9 A.M. to 12 P.M., 12 P.M. to 3 P.M., and 3 P.M. to 6 P.M., because they roughly correspond to three different congestion levels; see Figure 1. Thus, by studying the customers arriving during these three time intervals separately, we can test whether customers arriving at times of different congestion levels behave differently. We call it the *time-of-the-day effect*.

[21] For each message in Table 7, the corresponding range of the estimated delay is announced to customers. For example, for message 1 of the alternative announcement system $\mathcal{A}_1$, the corresponding range of the estimated delay, i.e., 0 to 60 seconds, is announced to customers.

[22] Finally, we have also explored the problem of when to provide the first announcement. Our results show that it may improve the performance of the call center by providing the first announcement earlier. Yet such improvement diminishes as the ratio between the VIP customers' service reward and the regular customers' service reward increases. For a more detailed description about this counterfactual study, see §2.8 of Yu (2014).

## References

Akşin Z, Ata B, Emadi S, Su C-L (2012) Structural estimation of callers' delay sensitivity in call centers. *Management Sci.* 59(12):2727–2746.

Allon G, Bassamboo A, Gurvich I (2011) "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.

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

Becker GS (1965) A theory of the allocation of time. *Econom. J.* 75(299):493–517.

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

Buell RW, Norton MI (2011) The labor illusion: How operational transparency increases perceived value. *Management Sci.* 57(9):1564–1579.

Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. (Springer, New York).

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

Gehan EA (1965) A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52(1–2):203–223.

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Sci.* 53(6): 962–970.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54(4):1185–1195.

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

Haviv M, Ritov Y (2001) Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Systems* 38(4):495–508.

Hogg RV, McKean JW, Craig A (2005) *Introduction to Mathematical Statistics*, 6th ed. (Pearson Education, Boston).

Horowitz JL (2001) The bootstrap. Heckman JJ, Leamer E, eds. *Handbook of Econometrics*, Vol. 5 (Elsevier, Amsterdam), 3159–3228.

Hui MK, Tse DK (1996) What to tell consumers in waits of different lengths: An integrative model of service evaluation. *J. Marketing* 60(2):81–90.

Ibrahim R, Whitt W (2009) Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Sci.* 55(10):1729–1742.

Judd KL (1998) *Numerical Methods in Economics* (MIT Press, Cambridge, MA).

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

Mandelbaum A, Shimkin N (2000) A model for rational abandonments from invisible queues. *Queueing Systems* 36(1):141–173.

Mandelbaum A, Zeltyn S (2013) Data-stories about (im)patient customers in tele-queues. *Queueing Systems* 75(2–4):115–146.

Munichor N, Rafaeli A (2007) Numbers or apologies? Customer reactions to telephone waiting time fillers. *J. Appl. Psych.* 92(2): 511–518.

Nair H (2007) Intertemporal price discrimination with forward-looking consumers: Application to the U.S. market for console video-games. *Quant. Marketing Econom.* 5(3):239–292.

Peto R, Peto J (1972) Asymptotically efficient rank invariant test procedures. *J. Roy. Statist. Soc. Ser. A* 135(2):185–207.

Rust J (1987) Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica* 55(5):999–1033.

Whitt W (1999) Improving service by informing customers about anticipated delays. *Management Sci.* 45(2):192–207.

Yu Q (2014) Delay announcements in services: Theory and empirics. Ph.D. thesis, Northwestern University, Evanston, IL.