Impact of Callers' History on Abandonment: Model and Implications

Seyed Morteza Emadi, Jayashankar M. Swaminathan
Kenan-Flagler Business School, University of North Carolina at Chapel Hill

Abstract

Caller abandonment could depend on their past waiting experiences. Using Cox regressions we show that callers who abandoned or waited for a shorter time in the past abandon more in the future. However, Cox regression approach does not shed light on callers' prior belief about the duration of their delays. Moreover, Cox regressions cannot separate the impact of callers' parameters such as their waiting costs on their abandonment behavior from the impact of their beliefs about their delay durations, which are affected by their past waiting experiences. To tease out the impact of callers' waiting experiences on their abandonment behavior, we use a structural estimation approach in a Bayesian learning framework. We estimate the parameters of this model from a call center data set with multiple priority classes. We show that in this call center new callers who do not have any experience with the call center are optimistic about their delay in the system and underestimate its length irrespective of their priority class. We also show that our bayesian learning model not only has a better fit to the data set compared to the rational expectation model in Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017) but also outperforms the rational expectation model in out-of-sample tests. In addition, our bayesian framework does not lead to biased estimates, which would happen under the rational expectation assumption if callers' belief about their waiting durations does not match their actual waiting time distribution. Our bayesian framework has managerial implications at both tactical and operational levels such as managing customer expectation about their delays in the system, and implementation of patience-based priority policies such as Least-Patience-First and Most-Patience-First scheduling.

1 Introduction

The service industry has grown extensively and consists of more than 75% of the gross domestic product of the United States.[1] Call centers are a major channel for providing different types of services to customers employing nearly 3 million people (Czinkota and Ronkainen (2012)). Designing modern call centers requires understanding of callers' patience level and their abandonment behavior. This can be achieved by analyzing callers' contact data. Majority of call centers utilize sophisticated management software that records details of customers' contact information including: customer ID, requested service, waiting time, outcome of the call (abandonment/receiving service), service time, etc. Therefore, when a caller contacts the call center, the call center manager

[1] Source: http://data.worldbank.org/

can identify if the caller is a new caller or if she has contacted in the past, and if it is the latter what happened during the previous call or calls. In other words, the call center manager has access to callers' contact history information and their past interactions with the system.

Customers' past interaction data can be very useful as demonstrated in other industries. For example, retailers use customers' past purchase history to offer customized coupons and deals (Rossi et al. (1996)). In the call center literature researchers have studied how call center data can be used to estimate the distribution of callers' patience parameters such as their patience threshold and their waiting costs (Gans et al. (2003), Brown et al. (2005), Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017)).[2] Using the methods in this literature, one can say that a caller's patience parameter is a draw from a distribution, but one cannot identify the patience parameter for a specific caller. Moreover, in the above literature, callers' contact history information is disregarded in the sense that all contacts of a caller are treated the same irrespective of their sequence and whether the contact is the first one or the caller has contacted in the past. In this paper, we lay out a framework to understand callers' individual abandonment behavior based on their contact history data and changes in their behavior across their different contacts. This framework gives the call center manager information about each caller's patience level and her expectation about the system delay based on the caller's contact history.

To investigate if callers' abandonment behavior depends on their past waiting experiences, we use a series of Cox regressions (Hosmer et al. (2008)). The regression results show that callers' abandonment behavior is indeed history dependent, and that callers who abandoned or waited for a shorter time in the past are more likely to abandon in subsequent contacts. However, the Cox regression does not give us any information about callers' prior belief about their delay durations in the system. In other words, using the Cox regressions we do not know how new customers who do not have any experience with the system think about the service quality in terms of delay durations. In addition, the Cox regression results do not explain the underlying model for callers' abandonment behavior. In particular, it is not clear if callers who abandon more frequently compared to other callers, have intrinsically higher waiting costs, or if they frequent abandonments are driven by their beliefs about the waiting time durations, which may not be short based upon their past experiences. Consequently, it is not clear if the difference in callers' abandonment behavior is driven by their heterogeneity of preferences (e.g. different waiting costs) or by their different contact histories, which led to different abandonment behaviors. If we could separate the impact of callers' patience parameters on their abandonment behavior from the impact of their contact history then we will be able to find each caller's belief about the waiting time duration independent of their intrinsic parameters. This separation could help in counterfactuals on the priority policy because changing the priority policy in the call center would not affect callers' parameters but would change their future waiting experiences and how those experiences affect their abandonment behavior.

To disentangle the impact of caller parameters such as their waiting costs from the impact of

[2] A caller's patience threshold is the amount of time she is willing to wait in the queue. Consequently, if her actual waiting time exceeds her patience threshold, she abandons.

their contact history we use a structural estimation approach in a Bayesian learning setting. Similar to Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017), we use an optimal stopping model for callers' abandonment behavior. We assume that in each period of time callers compare the expected utility of waiting and the expected utility of abandonment. They wait if the expected utility of waiting is higher than the expected utility of abandonment and abandon otherwise. Callers' utilities depend on three factors: their waiting cost, their value for service and their expectation (belief) about the waiting time distribution. In contrast to the extant literature (Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017)), we do not adopt a rational expectation framework in which it is assumed callers know the actual waiting time distribution irrespective of their past experiences with the call center. Under the rational expectation framework it is assumed that even new callers who never contacted the call center before know the waiting time distribution. This is a strong assumption and may not be realistic.

In our setting, callers' beliefs about the waiting time distribution depend on their prior belief and their past waiting experiences through a Bayesian updating framework. In particular, we assume callers believe that the waiting time distribution is Weibull. Callers know the shape parameter of this Weibull distribution. However, they update their beliefs about its scale parameter while waiting in the queue. We choose the Weibull distribution as a special case of "newvendor distributions" identified by Braden and Freimer (1991) to make the bayesian updating in the presence of censoring (abandonment in our case) tractable. The class of "newvendor distributions" has been used extensively in the literature in the context of bayesian learning in the presence of censoring (see Mersereau (2015), Lariviere and Porteus (1999) and Heese and Swaminathan (2010)). To accommodate the conjugate prior scheme, we assume that callers' belief about the scale parameter of the Weibull distribution for the waiting time is an inverse gamma distribution. The parameters of this inverse gamma distribution change by callers' contact history in particular by how long they waited and whether they abandoned or received service in their previous contacts.

For identification purposes, we assume callers in the same priority group have the same prior beliefs. This is a standard assumption in the Bayesian learning literature in Marketing and Industrial Organization (Eckstein et al. (1988), Erdem and Keane (1996) and Ackerberg (2003)). However, to account for caller heterogeneity, we adopt the latent class model in Lazarsfeld et al. (1968) and Heckman and Singer (1982). To be more specific, we assume within each priority group there are two class of callers. Callers in the same class have the same reward and cost parameters. We not only estimate the reward and cost parameters for each class within each priority group but also estimate the probability of callers belonging to each class.

Using the observations in a bank call center with three priority classes (High, Medium and Low), we estimate the parameters of the model using a Maximum Likelihood Estimation approach. The parameters of the model are: callers' waiting cost and their reward from receiving service, and the parameters of callers' prior belief about the waiting time distribution. We find that callers are optimistic about their delays in the system and underestimate their delay duration. We also show that the difference between callers from different priority classes in terms of their prior belief about

the waiting time distribution is not as significant as their actual waiting times. To be more specific, new callers irrespective of their priority group believe that they will receive service in less than 15 seconds even though their actual average waiting times ranges from 40 seconds to 90 seconds. This agrees with the fact that in this call center callers are not aware of their priority class and in the first contact could have the same belief.

We compare our model with the rational expectation model in the extant literature (Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017)) in four aspects. First, we show that our model has a better fit to the data set in terms of statistical fit (AIC and BIC measures). Second, we show that our model has a better prediction power compared with the rational expectation model. To do so, we perform out of sample tests and show that the prediction error of our model is less than third of that of the rational expectation model. This shows that our bayesian learning model can be a better candidate for policy experiment and doing what-if analysis. Third, we show that the rational expectation model may lead to poor inference about callers' patience levels. To be more specific, we use a non-parametric Kaplan-Meier (Kaplan and Meier (1958)) approach to find the patience level ranking of different priority groups of callers. Then we compare the ranking resulted from the estimation results of the rational expectation model and our bayesian learning model. We show that in contrast to the rational expectation model the ranking resulted from our bayesian learning model matches that of the non-parametric Kaplan-Meier model. And finally, we show that our bayesian learning framework does not have the bias problem that the rational expectation framework is prone to. In particular, we show that if callers' belief about the waiting time distribution does not match the actual distribution, the estimated parameters of callers under the rational expectation equilibrium are biased, while our model and estimation procedure lead to unbiased estimates.

Our Bayesian learning framework with the structural estimation approach has managerial implications at both tactical and operational levels: At the tactical level, callers' belief about the waiting time distribution shows callers' expectation of their delay in the system, which can be considered as a measure for callers' overall evaluation of the service quality. Therefore, the call center manager can get a sense of callers' expectation about the service quality based on callers' contact history data, and can impact this expectation by providing delay information. For example, if callers are pessimistic about their delays, the call center manager can provide delay information, which shifts callers' expectation toward their actual delays. Moreover, if callers' are optimistic about their delays, the call center manager may prefer to not provide any delay information to avoid increasing callers' abandonment rates.

At the operational level, the call center manager can use each caller's contact history to find her individual belief about the waiting time distribution. The knowledge of callers' individual belief together with the optimal stopping model could enable the call center manager to compute each caller's abandonment time distribution and expected patience threshold as a proxy for their actual patience threshold, which is not observed in the data. Note that callers may have different expected patience thresholds because of having different contact histories. Having a proxy for each

individual caller's patience threshold makes possible the implementation of personalized patience-based policies such as Least-Patience-First policy (Mandelbaum and Momcilovic (2014)). In the extant literature it has been assumed that the call center manager knows the patience threshold of each caller. However, it is not explained how this knowledge is acquired. This paper provides a framework for acquiring this knowledge.

Our paper has four main contributions: First, to the best of our knowledge this is the first work that illustrates callers' history-dependent behavior. We do this using a series of Cox regressions and show that callers' past interaction data provides information about their chance of abandonment in their future contacts. Second, we separate the impact of callers' patience parameters such as their waiting costs from the impact of their contact history using a structural estimation approach in a Bayesian learning setting. This separation enables us to calculate each callers' evaluation of the system delay after controlling for their parameters. In addition, we show that our model has a better fit and also a better prediction power compared to the rational expectation model in the extant literature. Third, the paper provides novel insights about callers' expectation about their delays. We show that callers are optimistic about the length of delay irrespective of their priority classes. And finally, our framework provides practical tool to manage customer expectation about the system delay, and to implement patience-based priority policies.

In the remainder of the paper, Section 2 presents the literature review. Section 3 describes our data set and demonstrates the impact of callers' contact history on their abandonment behavior. Section 4 presents the model for callers' abandonment behavior with Bayesian learning. Section 5 lays out the estimation framework and results. Section 6 compares our bayesian learning model with the rational expectation model in different aspect including but not limited to the statistical fit and the prediction power. Section 7 discusses some applications of the introduced framework. Finally, Section 8 concludes the paper.

2 Literature Review

Incorporating the impact of customer abandonment is an integral part of designing call centers. Customer abandonment has been studied extensively in the literature ranging from the traditional way of assuming an exogenous and fixed distribution for callers' patience thresholds (Gans et al. (2003)) to utility-based approaches (Hassin and Haviv (1995), Mandelbaum and Shimkin (2000), Shimkin and Mandelbaum (2004), Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017)). In Mandelbaum and Shimkin (2000) and Shimkin and Mandelbaum (2004), the authors use a utility-based approach to model customers' abandonment behavior. Customers abandon the call center if their waiting time reaches their optimal abandonment time value. Aksin et al. (2013) provide a framework to estimate customers' parameters from call center data. The authors model customers' abandonment decisions as an optimal stopping model, where abandonment represents stopping. Aksin et al. (2017) and Yu et al. (2017) uses a similar optimal stopping model for callers' abandonment behavior under delay announcements. In all of these papers, the authors assume

that all customers (or customers who hear the same announcement message if delay information is provided) have the same belief about the waiting time distribution and this belief perfectly matches the actual waiting time distribution in the data. We use a similar optimal stopping model for modeling customers' abandonment behavior. However, we consider a Bayesian learning model where customers do not know the actual waiting time distribution but learn it over time based on their waiting time experiences. We next describe the Bayesian learning literature relevant to this work.

Our bayesian learning model builds upon the empirical dynamic discrete choice literature in the Marketing and Industrial organization fields (Eckstein et al. (1988), Erdem and Keane (1996) and Ackerberg (2003)). The focus of this literature is on how consumer learning about brand attributes affects purchasing behavior, using a dynamic programming approach. One of the most relevant papers to our work is Erdem and Keane (1996), which provides a Bayesian learning framework to study how brand choice probabilities depend on past usage experiences and advertising exposures. The authors form a likelihood function for their model and estimate its parameters using observations from a scanner data. Similarly, Ackerberg (2003) uses a dynamic learning model of consumer behavior on a frequently purchased packaged products to study the impact of past purchases and advertising. We use a similar Bayesian learning framework in our model. However, to the best of our knowledge this paper is the first empirical work on bayesian learning about service quality (in our setting waiting duration) in the service operations context.

Other than the marketing literature, bayesian learning has been used in the Operations literature in the context of demand learning. Wecker (1978), Nahmias (1994) and Agrawal and Smith (1996) study models to estimate demand parameter from sales data in the presence of stock outs, which leads to censoring of demand data. Harpaz et al. (1982) and Ding et al. (2002) use parametric bayesian models for learning demand from censored observations. Lariviere and Porteus (1999), Mersereau (2015) and Heese and Swaminathan (2010) use the "newsvendor distribution" framework of Braden and Freimer (1991) in their bayesian learning models. In this stream of literature, the firm learns about demand from a censored data set. However, in our setting, callers learn about the waiting time using their past, and possibly censored, waiting experience. Similar to most of the works in this literature, we also use the "newsvendor distributions" to facilitate our analysis.

The framework introduced in this paper can be used in implementation of patience-based priority policies by providing information about each caller's patience level. Bassamboo and Randhawa (2014) use the amount of time callers have been waiting in the system as a proxy for their patience level and provide a Time-In-Queue policy that significantly improves the performance measures of the call center such as the queue length and the offered wait times. Mandelbaum and Momcilovic (2014) compare the Least-Patience-First (LPF) policy with the First-Come-First-Served policy and show that the LPF policy can lead to lower abandonment rates in the call center.

3 Illustrating the Impact of Past Waiting Experiences

We first describe our data set. Then, we use a series of Cox proportional hazard regressions to illustrate the impact of callers' past interactions on their abandonment behavior.

3.1 Data Description

Our data set contains detailed call level data from a medium-sized bank call center with around 400 agents.[3] The data set spans a 27 month period from April of 2007 through June of 2009. The call center provides six types of services: Private, Securities, Internet, Other languages, Loans and Solutions. We focus on callers who requested the Private service type as this portion of the data contains more than 80% of the callers.

This call center serves four classes of customers: High-priority (VIP customers), Medium-priority, Low-priority and No-priority. The priorities are assigned based on callers' account information and sales data, which are not observed by us. However, we can observe callers' assigned priority classes in data. We use customer IDs to track customers and record their contact history. We can observe the customer ID for the high, medium and low-priority callers. The no-priority callers are unidentified customers, and their IDs are not observable. Therefore, we exclude the no-priority callers from our analysis. Other than the customer IDs, we can observe the following entries in the data: arrival time and day, waiting time in the queue, whether the caller abandoned the system or waited until talking to an agent, service time and ID of the agent who served the call.

Figure 1 shows the histogram of the time between customers' consecutive contacts. On average, the time between two consecutive contacts is 27.88 days, and it does not exceed 240 days with 99% probability. In our analysis, the order of customer contacts is an important factor. For example, we need to know if a customer's specific contact is the first contact or the caller has contacted before. Therefore, we focus on callers whose first recorded contact in the data does not occur in 2007 (April to December of 2007). In other words, all callers included in our analysis contacted for the first time between January of 2008 and June of 2009. Given that the time between contacts does not exceed 240 days with 99% probability and that the total number of days from April through December of 2007 is 270, the chance that the callers we included in our analyses contacted before 2008 is very low.

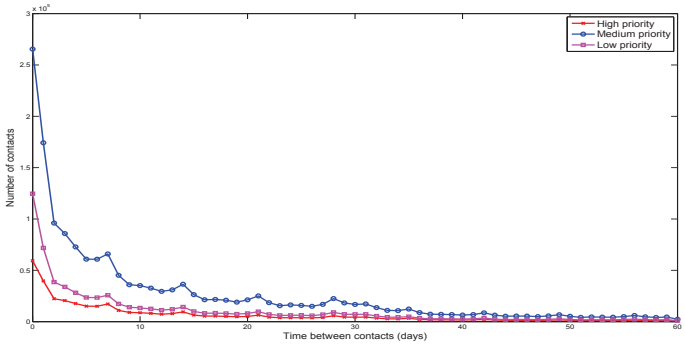

Figure 1: The histogram of the time between customers' consecutive contacts.

Figure 2 shows the histogram for the number of times customers contacted the call center. On average during the 18 month period from January 2008 through June of 2009 , customers contact the call center 5.65 times.

[3]Our data set was generously made available to us by the Service Enterprise Engineering (SEE) lab at the Technion (http://ie.technion.ac.il/Labs/Serveng/).

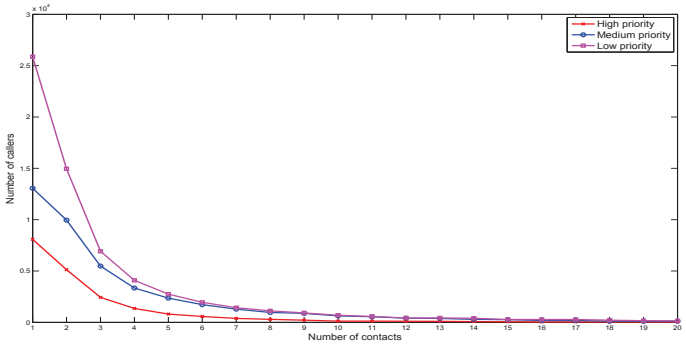

Figure 2: The histogram of the number of contacts.

Our data set exhibits a strong time of the day effect. Figure 3 shows the number of arrivals to the call center depending on time of the day. Furthermore, Figures 4 shows the average of callers' waiting times and abandonment rates during a day. As can be seen in Figure 3 the time between 9am and 4pm corresponds to the most congested time of the day (i.e. rush hours). In addition, Figure 4 shows that even though the time between 9am and 4pm corresponds to the most congested time of the day, because of the staffing policy in the call center and having a higher number of staff in this time period, the average waiting times and abandonment rates are relatively lower than the rest of the day.

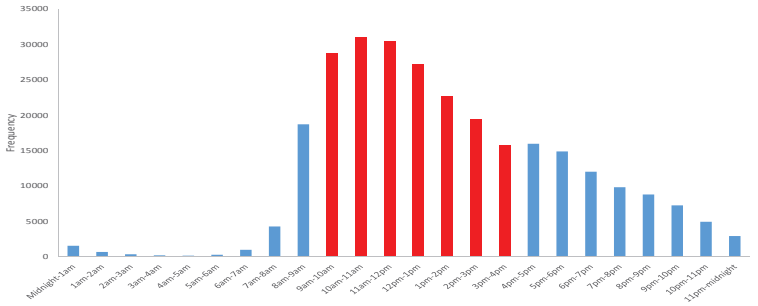

Figure 3: Number of arrivals during a day.

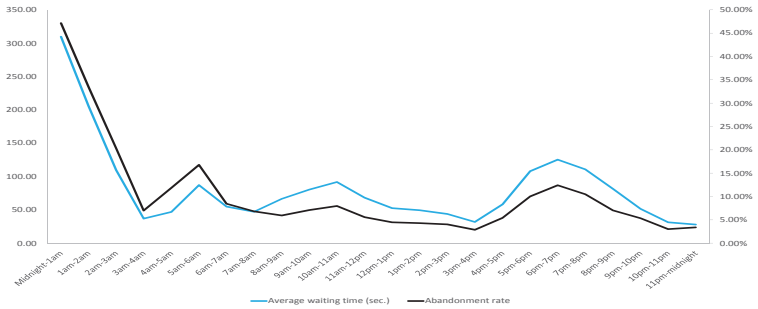

Figure 4: Average waiting times and abandonment rates during a day.

To account for time of the day effect in our future analysis, we divide a day to two intervals in our analyses: Rush-hours (9am to 4pm) and Non-rush-hours (before 9am and after 4pm). In summary our data sets includes all contacts of callers from the high, medium and low priority

groups whose first recorded contact in the data appears between January of 2008 and June of 2009, and requested the retail service type. The summary statistics for the portion of the data we used in our analysis is given in Table 1.

Priority class	Number of callers	Average number of contacts	Abandonment rate	Average waiting time (sec.)
High	14,148	4.53	3.23 %	43.93
Medium	28,209	5.93	5.19 %	61.53
Low	45,674	5.80	8.44 %	84.57

Table 1: Summary statistics for the portion of the data used in the analysis.

In the next section, we use survival analyses such as the Kaplan-Meier estimation and Cox proportional hazard regressions to illustrate the impact of callers' past contact history on their abandonment behavior in our data set.

3.2 Impact of Past Waiting Experiences

Denote by W_{-1} and O_{-1} the waiting times and outcomes of callers' previous contact, where the outcome variables is equal to 1 if the caller abandoned in the previous contact and is equal to 0 otherwise. We define customers' abandonment behavior as the distribution of customers' abandonment time (patience time) and its hazard function, which captures the probability of abandonment in each period if the caller has not received service or abandoned yet.[4]

We first use the Kaplan-Meier estimator (Kaplan and Meier (1958)) to find customers' survival function and show that it changes depending on call outcomes.[5] Then, we use Cox proportional hazard regressions (Cox (1972)) to see how customers' abandonment hazard rate depends on their waiting experience in their last contact. Figure 5 shows the survival functions for two groups of customers: customers who abandoned in their previous contact ($O_{-1} = 1$) and customers who waited until entering service in their previous contact ($O_{-1} = 0$).

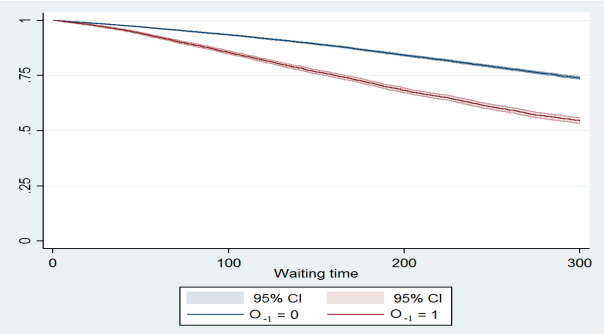

Figure 5: The survival function of customers depending on the outcome of their previous contact. The figure also shows the 95% confidence interval for the survival function estimate.

As can be seen in Figure 5, customers' survival functions are different depending on the outcome

[4] As the data is censored in the sense that we do not observe the actual abandonment times (patience times) of callers who receive service, we need to use survival analyses that take censoring into account.

[5] The survival function is 1-CDF of abandonment time distribution.

of their first contact.[6] Moreover, Figure 5 shows that customers who abandoned in their previous contact have a lower probability of survival in their current contact, i.e. abandon with a higher probability in the current contact.

To find the impact of callers' contact history on their abandonment behavior we use the stratified Cox regression (Hosmer et al. (2008)), where we define the high, medium and low-priority classes as the strata. We use the stratified Cox regression instead of using a simple Cox regression to accommodate the possibility that callers from different priority groups may have different baseline hazard functions. See Appendix A for more details about the stratified Cox regression.

The explanatory variables in the Cox regressions are: waiting time in the previous contact W_{-1}, the outcome of the previous contact O_{-1}, the interaction between outcome and wait duration $O_{-1} \times W_{-1}$, the indicator variables for time of the contact denoted by $D_{Rush-hour}$ (equal to 1 if the call occurred between 9am and 4pm), and the indicator variables for weekdays versus weekends. Table 2 illustrates the results.

Table 2: The results for the stratified Cox regression, where the priority groups are the strata.

Variable	**Coefficient**	**(Std. Err.)**
O_{-1}	0.7303**	(0.0285)
W_{-1}	-0.0015**	(0.0001)
$O_{-1} \times W_{-1}$	0.0007**	(0.0001)
$D_{Rush-hour}$	-0.0481**	(0.0172)
Weekdays	0.0508	(0.0281)
Contact number	0.0054**	(0.0006)

**Denotes statistically significant at 0.05.

As can be seen in Table 2, the coefficients for both O_{-1} and W_{-1} are significant. The coefficient for O_{-1} is positive and for W_{-1} is negative. This shows that callers who abandoned in the previous contact abandon with a higher probability in their current contact. Moreover, callers who waited for a longer time in the previous contact, which may indicate a higher patience level, have a lower chance of abandonment in the current contact. However, given that the coefficient for the interaction term $O_{-1} \times W_{-1}$ is negative, callers who waited for a longer time but abandoned in their previous contact would abandon with a higher probability in their current contact compared to customers who received service in their previous contact. Furthermore, the coefficient for the contact number is positive and significant, which shows more frequent callers are less patient. In other words, callers' patience level goes down by their experience with the call center. In addition, the coefficient for the indicator variable for Weekdays ("Weekdays") is not significant but that of the Rush-hour ("$D_{Rush-hour}$") is significant, which shows time of the day effect is more significant than day of the week effect in terms of callers' abandonment behavior.[7] To check the robustness of

[6]The Logrank test (Mantel (1966)) shows that the survival functions in Figure 5 are statistically different at a 0.05 significance level.

[7]We added an indicator variable for redial to the regression, which is equal to 1 if the second call is within 24 hours of the first call. Surprisingly, the coefficient for the redial indicator variable was not significant, suggesting that whether a call is a redial or not does not significantly impact callers' abandonment behavior. Moreover, we changed the definition of redial by assuming that a call is a redial if the caller contacts within 4, 6, 12 and 48 hours of the first contact and got the same insight.

our findings, we have repeated the same Cox regression analysis on each priority group in isolation and got the same insights. See Appendix B for the details of this analysis.

Discussion: Analyses in this section show that customers demonstrate history-dependent behavior. This history-dependent behavior is similar to customer inertia in the marketing literature (Dubé et al. (2010)). That is, customers who purchased a product in the past (similar to abandoning in our setting) have a higher probability of purchasing it in the future. However, as indicated by Heckman (1979) and Dubé et al. (2010), it is not clear whether this behavior is driven by the difference in customers' waiting costs or the difference in their learning processes.

In particular, a higher chance of abandonment in the current contact can be driven by two factors: 1) Customer learning: Customers who abandoned in the past believe that the waiting times are long, so the higher chance of abandonment in the current contact is driven by change in customers' expectation about the waiting time distribution, and 2) Customers' intrinsic high waiting cost: Customers who abandoned in the past or were not willing to wait for longer times are intrinsically less patient, therefore, they abandon with a higher probability in the current contact as they did in the past. Separating the impact of callers' preferences such as their waiting cost from the impact of their waiting experiences helps the call center manager find callers' evaluation of the system delay independent of her patience parameters. Moreover, it helps in performing priority policy counterfactual analyses since any change in the priority policy would impact callers' waiting experiences but would not affect their parameters. Consequently, these counterfactuals cannot be done using the Cox regressions.

Moreover, the Cox regression analysis does not give us information about callers' prior belief about their delay duration. The Cox regression shows that callers get less patient as they acquire more experience with the call center based on the positive coefficient for the contact number in Table 2. However, the Cox regression does not shed light on callers' prior belief about their delay durations and whether their belief is optimistic or pessimistic compared to the actual delays in the call center. To disentangle the effect of customer preferences from their learning process, and to get a sense of callers' prior belief about their delay duration, we use a structural model in the next section.

4 Model for Customer Learning and Abandonment Behavior

In this section we first lay out a Bayesian framework for customer learning about the waiting time distribution. Then, we model callers' abandonment decision as an optimal stopping time problem.

Preliminaries. Suppose that callers are indexed by $i \in \{1, ..., N\}$. Denote by n_i the number of times caller i contacts the call center in the data set. To account for time of the day effect, we divide a day to M intervals and let q_{in} denote the index of the time interval for customer i's n^{th} contact.[8] We denote by O_{int} the outcome of caller i's n^{th} contact after waiting for t periods, which is equal to 0 if the caller has entered service at time t, and is equal to 1 if the caller has abandoned at time t or has not entered the service stage yet. Finally, we denote the waiting time of caller i in

[8] In our estimation, we divide the day to two intervals ($M = 2$): Rush-hours (9am-4pm), and Non-rush-hours.

her n^{th} contact by w_{in} and the final outcome of the n^{th} contact by O_{in}. Note that $O_{in} = O_{inw_{in}}$.

4.1 A Bayesian Framework for Customer Learning

We assume that callers learn about the waiting time distribution in a Bayesian fashion. They have a prior belief about a component of the waiting time distribution and update it while waiting in the queue.

We consider a Weibull distribution for the actual waiting time distribution of the callers in the data, which is a priori unknown to callers. We have chosen the Weibull distribution for the following reasons: First, the Weibull distribution is perhaps the most widely used parametric survival distribution (Ibrahim et al. (2005)) and has the ability to assume the characteristics of different distribution types. Moreover, a Weibull distribution can show increasing, decreasing or constant hazard rate based on the value of its shape parameter. Second, given that customers may abandon in our data set, their observations of the waiting times are censored. Consequently, we have chosen the Weibull distribution which is in the family of "newsvendor" distributions. The class of "newsvendor" distributions has the conjugate prior property in the presence of censoring, which makes the analysis of Bayesian updating tractable; see Braden and Freimer (1991) for more details. This family of distributions have been used extensively to study bayesian learning under censoring (see Mersereau (2015) and Lariviere and Porteus (1999)). A newsvendor distribution is a continuous distribution with a cdf of the form $1 - \exp(-\zeta h(x))$ for $x \geq 0$. where the function $h()$ is positive, differentiable and increasing for x. These conditions are satisfied for the Weibull distribution as illustrated below.

The Weibull distribution for the actual waiting times of the day interval $m \in \{1, ..., M\}$ with the pdf (cdf) denoted by $f^m(t; k_0^m, \gamma_0^m)$ ($F^m(t; k_0^m, \gamma_0^m)$) is given by

$$f^m(t; k_0^m, \gamma_0^m) = \frac{k_0^m}{\gamma_0^m} t^{k_0^m - 1} e^{-\frac{t^{k_0^m}}{\gamma_0^m}} \text{ and } F^m(t; k_0^m, \gamma_0^m) = 1 - e^{-\frac{t^{k_0^m}}{\gamma_0^m}}, \tag{1}$$

where k_0^m is the shape parameter and γ_0^m is the scale parameter. Note that the Weibull distribution is a newsvendor distribution with $h(x) = x^{k_0^m}$. The shape parameter of the waiting time distribution k_0^m solely determines if its hazard rate is increasing, decreasing or constant. To be more specific, $k_0^m < 1$, $k_0^m = 1$ and $k_0^m > 1$ correspond to decreasing, constant and increasing chance of receiving service. We assume that customers know the shape parameters $k_0^m,\ m \in \{1, ..., M\}$. Hence, they know if their chances of receiving service goes up/down or stays constant by time. However, we assume that customers are uncertain about the scale parameter γ_0^m and learn about it based on their waiting time experiences.[9]

We assume that callers who do not have any contact experience with the call center believe

[9]Later on in this section, we introduce the conjugate prior scheme and the bayesian updating. To the best of our knowledge we have the conjugate prior property under censoring (abandonment in our case) if customers update only the scale parameter of the distribution. In addition, the updating process is analytically intractable if we assume callers update both the shape and scale parameter. Furthermore, it will lead to identification issues in the estimation procedure. As due to the higher degrees of freedom the Maximum Likelihood Estimation problem will have multiple solutions.

that γ_0^m is distributed according to an inverse gamma distribution with the shape parameter $\mu^{pr,m}$ and the scale parameters $\delta^{pr,m}$. We denote this distribution by $Inv-Gamma(\mu^{pr,m},\delta^{pr,m})$, which is callers' prior belief about the distribution of γ_0.[10] We choose the inverse gamma distribution because it can accommodate the conjugate prior scheme for the Weibull distribution. That is callers' posterior belief about the distribution of γ_0^m will be another inverse gamma distribution. Next, we describe callers' Bayesian updating process.

Callers update their belief about the distribution of γ_0^m only if they contact during day interval m. Denote by $(\mu_{in}^{pr,m},\delta_{in}^{pr,m})$ the parameters of caller i's prior belief about the distribution of γ_0^m right before her n^{th} contact. Note that even though callers have the same prior belief before having any contact experience with the call center, they may have different beliefs in their future contacts because of having different waiting experiences. Moreover, we assume callers update their belief while waiting and denote by $(\mu_{in}^{po,m}(t),\delta_{in}^{po,m}(t))$ the parameters of caller i's posterior belief about the distribution of γ_0^m at her n^{th} contact after waiting for t periods. Upon arrival at $t=0$ the caller has not acquired any new information about her waiting time. Consequently, we have $(\mu_{in}^{po,m}(0),\delta_{in,m}^{po}(0))=(\mu_{in}^{pr,m},\delta_{in}^{pr,m})$. Figure 6 shows the diagram for the updating process.

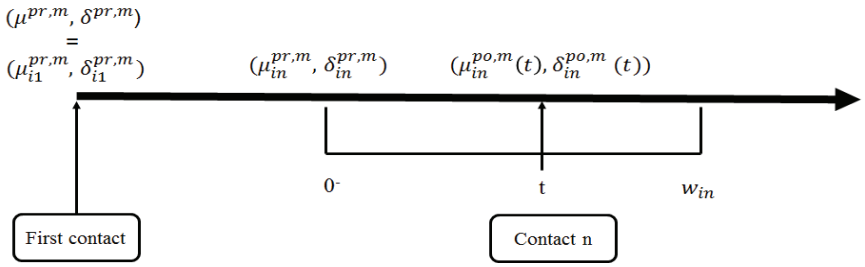

Figure 6: The diagram of the updating process.

Proposition 1 characterizes callers' Bayesian updating process; see Appendix C for its proof.

Proposition 1. *Suppose that caller i's belief about the distribution of γ_0^m prior to her n^{th} contact is an Inverse Gamma distribution with parameters $(\mu_{in}^{pr,m},\delta_{in}^{pr,m})$. Then, her posterior belief after waiting for t periods in her n^{th} contact has the following parameters:*

$$\mu_{in}^{po,m}(t)=\mu_{in}^{pr,m}+\mathbb{I}_{\{q_{in}=m\}}(1-O_{int}), \tag{2}$$

$$\delta_{in}^{po,m}(t)=\delta_{in}^{pr,m}+\mathbb{I}_{\{q_{in}=m\}}t^{k_0^m}, \tag{3}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. Moreover, caller i's posterior predictive distribution by time t in her n^{th} contact for the waiting time distribution of day interval m denoted by $F_{in}^{po,m}(t)$ is given by

$$F_{in}^{po,m}(t)=1-\left(\frac{\delta_{in}^{po,m}(t)}{\delta_{in}^{po,m}(t)+t^{k_0^m}}\right)^{\mu_{in}^{po,m}(t)}, \tag{4}$$

and the caller's belief about her chance of receiving service in the next period if she decides to wait

[10] The pdf of $Inv-Gamma(a,b)$ is given by $\exp(-bt)b^a/(\Gamma(a)t^{-(a+1)})$, where $\Gamma(\cdot)$ is the gamma function.

denoted by $\pi_{in}^{po,m}(t)$ *has the following form*

$$\pi_{in}^{po,m}(t) = \frac{F_{in}^{po,m}(t+1) - F_{in}^{po,m}(t)}{1 - F_{in}^{po,m}(t)} = 1 - \Big(\frac{\delta_{in}^{po,m}(t) + t^{k_0^m}}{\delta_{in}^{po,m}(t) + (t+1)^{k_0^m}}\Big)^{\mu_{in}^{po,m}(t)}. \tag{5}$$

Finally, suppose that the caller has waited for w_{in} *period in her* n^{th} *contact, then assuming callers do not forget what they learned in the past we have*

$$\mu_{in+1}^{pr,m} = \mu_{in}^{po,m}(w_{in}), \ \textit{and} \ \ \delta_{in+1}^{pr,m} = \delta_{in}^{po,m}(w_{in}). \tag{6}$$

Proposition 1 shows that callers' contact history information impacts their posterior beliefs about the scale parameter of the waiting time distribution in two ways: the outcome of their contacts (whether they received service or not) affects the shape parameter of their posterior belief and the duration of their waitings impacts the scale parameter of their posterior belief. To be more specific, the shape parameter of the inverse gamma distribution for callers' posterior belief ($\mu_{in}^{po,m}(t)$) changes only and only if callers receive services; i.e. $O_{int} \neq 1$. However, the scale parameter ($\delta_{in}^{po,m}(t)$) increases after each contact even if the caller does not receive service.

Given the updating process characterized in Proposition 1, we can show that callers' updating process is consistent in the sense that callers will eventually learn the scale parameter of the waiting time distribution γ_0^m even though because of abandonments callers' observations could be censored. To be more specific, callers' posterior belief distribution about the scale parameter of the waiting time distribution γ_0^m converges to a distribution with a variance equal to zero and a mean equal to γ_0^m. Proposition 2 provides more details about the convergence of callers' posterior belief distribution. Without loss of generality we suppress the subscript for the day interval m in Proposition 2; see Appendix C for its proof.

Proposition 2. *Suppose that the waiting time distribution is Weibull with the shape and scale parameters equal to* k_0 *and* γ_0*, respectively. Callers' prior belief about the* γ_0 *is an inverse gamma distribution with parameters* μ^{pr} *and* δ^{pr}*. Callers know* k_0*, and they update their belief about* γ_0 *according to the process described in Proposition 1. Callers abandon if their actual waiting time, which is a draw from the Weibull distribution of the waiting time, is greater than their patience time. We assume callers' patience time is a random draw from a distribution with pdf and cdf equal to* $g(\cdot)$ *and* $G(\cdot)$*, and is independent of the waiting time random variable. If callers' patience time distribution has some mass at non-zero values (i.e.* $G(0) < 1$*) then callers' posterior belief about* γ_0 *converges to a distribution with a variance equal to zero and a mean equal to* γ_0*.*

Callers make their abandonment decision based on their belief about their chances of receiving service and their preferences. In the next section, we lay out an optimal stopping model for callers abandonment decisions.

4.2 An Optimal Stopping Model for Callers' Abandonment Decisions

We use an optimal stopping model similar to the model introduced in Aksin et al. (2013) to characterize callers' abandonment decisions with one fundamental difference: In contrast to Aksin et al. (2013), we do not assume that the actual waiting time distribution in the system is common knowledge across all callers. But we assume callers learn about the actual waiting time distribution while waiting, which in turn will impact their abandonment decisions.

We assume that callers are forward looking. They abandon the queue if their expected utility of abandonment is higher than their expected utility of waiting. If they decide to wait, they choose between waiting and abandonment again in the next time period based on the expected utilities. Suppose that caller i's n^{th} contact happens during the day interval m. Caller i makes her abandonment decisions based on three factors: her waiting cost per unit of time denoted by c_i, her reward from receiving service denoted by r_i and her posterior belief about the waiting time distribution in the day interval m by period t denoted by $F_{in}^{po,m}(t)$ and its hazard rate denoted by $\pi_{in}^{po,m}(t)$. These factors impact caller i's decisions through her utility. Let $H_{in} = \Big([w_{ik}, O_{ik}, q_{ik}]_{k=1..n-1}\Big)$ denote the vector of her past waiting times, final call outcomes and indexes of day intervals of contacts. In each period of time the caller take the action $d \in \{0, 1\}$ that maximizes her utility; $d = 1$ corresponds to abandonment and $d = 0$ corresponds to waiting. Caller i's utility in period t since her arrival at her n^{th} contact is given by

$$u_{in}(t, d, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}, \epsilon_{int}(d)) = v_{in}(t, d, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) + \epsilon_{int}(d), \tag{7}$$

where $\epsilon_{int}(d)$ is the error term corresponding to action d that captures the impact of external shocks that may shift caller i's utility toward an action. We assume that the error terms are independent across actions, callers, contacts and time periods. The function $v_{in}(t, d, r, c, \mu^{pr,m}, \delta^{pr,m}; H_{in})$ is the nominal utility and is independent of the error term. We normalize the nominal utility of abandonment to zero. That is

$$v_{in}(t, 1, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = 0. \tag{8}$$

The nominal utility of waiting is given by

$$v_{in}(t, 0, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = -c_i + \pi_{in}^{po,m}(t) r_i + (1 - \pi_{in}^{po,m}(t)) V_{in}(t, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}), \tag{9}$$

where $V_{in}(t, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in})$ is the integrated value function, which is the expected maximum utility in the next period, where expectation is taken over the error terms in the next period, and has the following form

$$V_{in}(t, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = \int \int \max_{d \in \{0,1\}} \Big(u_{in}(t+1, d, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}, \epsilon_{int+1}(d))\Big) \tag{10}$$
$$d\epsilon_{int+1}(1) d\epsilon_{int+1}(0).$$

The three terms on the right-hand side of (9) are as follows: the cost of waiting, caller i's belief about the expected utility of receiving service in period t and callers i's belief about the expected utility of not receiving service in period t but making optimal decisions in the upcoming periods.

We assume that the error terms ϵ_{int} have a type-I extreme value distribution with the scale parameter σ_ϵ and the location parameter $-\sigma_\epsilon\gamma$, where γ is Euler's constant. This ensures that the mean of the extreme value distribution is zero. Then based on Aksin et al. (2013), the recursive formula for the integrated value function is given by

$$V_{in}(t, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = \sigma_\epsilon \log\Big[1 + \exp\Big(\frac{v_{in}(t+1, 0, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in})}{\sigma_\epsilon}\Big)\Big]. \tag{11}$$

Similar to Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017) We assume that the terminal value for the integrated value function is $V_{in}(T, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = 0$, where T is the maximum waiting time of callers. Following the distributional assumption for the error terms, the probability of choosing action $d \in \{0, 1\}$ in period t of caller i's n^{th} contact has the Logit form and is given by

$$P_{int}(d, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = \frac{\exp(\frac{v_{in}(t,d,r_i,c_i,\mu^{pr,m},\delta^{pr,m};H_{in})}{\sigma_\epsilon})}{1 + \exp(\frac{v_{in}(t,0,r_i,c_i,\mu^{pr,m},\delta^{pr,m};H_{in})}{\sigma_\epsilon})}. \tag{12}$$

4.3 Caller Heterogeneity and Structural Parameters of the Model

To account for caller heterogeneity, we adopt the latent class model in Lazarsfeld et al. (1968) and Heckman and Singer (1982). To be more specific, we assume that for each priority class callers reward and cost parameters (r_i, c_i) are equal to (r^1, c^1) with probability $0 \leq \eta^1 \leq 1$ and are equal to (r^2, c^2) with probability $\eta^2 = 1 - \eta^1$. In other words, we segment the population of callers in the same priority class to two groups and assume callers are in the first segment with probability η^1 and in the second segment with probability η^2.

Furthermore, following Erdem and Keane (1996) and Erdem et al. (2008), for identification purpose, we assume that callers in the same priority group have the same prior belief about the distribution of the scale parameter of the waiting time distribution (callers from different priority groups can have different priors). Note that even though callers from the same priority group have the same prior belief, given their different contact history and waiting experiences, they will be heterogeneous in their belief about the waiting time distribution in their future contacts. Hence, the heterogeneity of beliefs arises endogenously across time/contacts through their contact history.

Moreover, as we will explain in section 5.2.1, for identification we normalize σ_ϵ and fix it at 1. Following these assumptions, the set of structural parameters of the model for callers (from the same priority group) are $\Theta = \{(k_0^m, \gamma_0^m)_{\{m=1...M\}}, (r^l, c^l, \eta^l)_{\{l=1,2\}}, (\mu^{pr,m}, \delta^{pr,m})_{\{m=1...M\}}\}$. Note that callers from different priority groups have different sets of structural parameters. If we know these parameters and a caller's past waiting experiences (H_{in}), we can use equation (12) to find the caller's chance of abandoning in each contact and each period. In the next section, we provide the estimation strategy to recover the structural parameters of the model using the observations in our data set.

5 Estimation Strategy

In this section we provide the estimation strategy to recover the structural parameters of the model from the data set. We assume each priority group has its own $\Theta = \{(k_0^m, \gamma_0^m)_{\{m=1...M\}}, (r^l, c^l, \eta^l)_{\{l=1,2\}}, (\mu^{pr,m}, \delta^{pr,m})_{\{m=1...M\}}\}$ and we estimate it for each priority group separately.

In addition, to alleviate the complexity of solving the Maximum Likelihood Estimation problem, we decrease the number of decision periods of callers by assuming that callers make their abandonment decisions every 10 seconds. Since our data is more granular, we truncate the abandonment times downward and the service initiation times upward.[11] We provide the details of the estimation procedure in the remainder of this section.

Denote by d_{int} the action of caller i in period t of her n^{th} contact. If the caller decides to wait $d_{int} = 0$ and if she decides to abandon $d_{int} = 1$. To account for the time of the day effect, we divide a day to Rush-hour (R) and Non-Rush-hour (NR) intervals ($M = 2$), where R corresponds to 9am to 4pm and NR corresponds to the remainder of the day.

Our estimation procedure consists of two main steps below for each priority group:

- Step 1: Estimating the parameters of the actual waiting time distribution $(k_0^m, \gamma_0^m)_{\{m=1...M\}}$ for each time interval.
- Step 2: Estimating the distribution of callers reward and cost parameters $((r^l, c^l, \eta^l)_{\{l=1,2\}})$ and their prior beliefs about the waiting time distribution in each day interval $(\mu^{pr,m}, \delta^{pr,m})_{\{m=1...M\}}$.

5.1 Estimating the Shape Parameter of the Waiting Time Distribution

As mentioned in Section 4.1, we assume that the waiting time distribution for each day interval is a Weibull distribution. To estimate the parameters of the waiting time distributions in the day interval m, i.e. (k_0^m, γ_0^m), we use a Maximum Likelihood Estimation method.

Suppose that caller i contacts during the day interval m in her n^{th} contact ($q_{in} = m$). Recall that w_{in} is the waiting time of the caller in this contact. If the final outcome of this call is entering the service stage, the likelihood that the caller's waiting time is a draw from the actual waiting time distribution for the day interval m is $f^m(w_{in}; k_0^m, \gamma_0^m)$.[12] However, if the caller abandons in this contact, the likelihood of observing w_{in} is $1 - F^m(w_{in}; k_0^m, \gamma_0^m)$ as we know that the actual time of entering the service stage which is a draw from the actual waiting time distribution is larger than w_{in}. Consequently, given (1) the log-likelihood function of observation in the day interval m denoted by $LL(k_0^m, \gamma_0^m)$ has the following form:

$$LL(k_0^m, \gamma_0^m) = \sum_{i=1}^{N} \sum_{n=1}^{n_i} \mathbb{I}_{\{q_{in}=m\}} (1 - O_{in}) \log \left(\frac{k_0^m}{\gamma_0^m} {w_{in}}^{k_0^m - 1} e^{-\frac{w_{in}}{\gamma_0}^{k_0^m}} \right) + \mathbb{I}_{\{q_{in}=m\}} O_{in} \log \left(e^{-\frac{w_{in}}{\gamma_0^m}^{k_0^m}} \right).$$

[11] Assuming that callers make decision every 15 or 20 seconds changes the estimation results but does not affect the key results and the insights of the paper.

[12] Recall that $f^m(\cdot; k_0^m, \gamma_0^m)$ and $F^m(\cdot; k_0^m, \gamma_0^m)$ denote the p.d.f. and c.d.f. of the waiting time distribution for the day interval m.

To estimate k_0^m and γ_0^m we maximize $LL(k_0^m, \gamma_0^m)$ subject to $k_0^m \geq 0$ and $\gamma_0^m \geq 0$. Table 3 shows the estimation results for different priority groups and day intervals.

Priority group	Shape parameter (k_0)	Scale parameter (γ_0)
High-priority (NR)	0.726 (0.009)	4.371 (0.109)
High-priority (R)	0.741 (0.007)	3.698 (0.071)
Medium-priority (NR)	0.715 (0.006)	6.626 (0.101)
Medium-priority (R)	0.742 (0.004)	5.111 (0.0567)
Low-priority (NR)	0.704 (0.005)	9.717 (0.128)
Low-priority (R)	0.754 (0.004)	7.313 (0.069)

Table 3: The estimates for the shape and scale parameter of the Weibull distribution of the waiting times. The numbers in parenthesis show the standard errors of the estimates.

5.2 Estimating Callers' Reward and Cost Parameters, and Their Prior Beliefs

In this section, we lay out the procedure to estimate the distribution of callers' reward and cost parameter, and the parameters of their prior beliefs. We perform this estimation procedure on each priority group in isolation. We first discuss identification, and then explain the estimation procedure.

5.2.1 Identification

Using Equations (8) to (9) and Equatione (11) to (12), we can show that caller i's abandonment probability in period t of her n^{th} contact can be written as follows:

$$P_{int}(1, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = \Omega\Big(\Big\{ -\frac{c_i}{\sigma_\epsilon} + \frac{r_i}{\sigma_\epsilon}\pi_{in}^{po,m}(s)\Big\}_{s\geq t}\Big), \tag{13}$$

where $\Omega(\cdot)$ is a suitably defined function, and $-c_i/\sigma_\epsilon + (r_i/\sigma_\epsilon)\pi_{in}^{po,m}(s); s \geq t$ is the per period utility of waiting in period s. As can be seen in (13), multiplying r_i, c_i and σ_ϵ by a constant would not change the choice probabilities. Consequently, we need to normalize one of these parameters. Following the standard practice in the Industrial Organization literature (Nevo (2000)), we fix $\sigma_\epsilon = 1$ from now on. In addition, given Equation (12) for $d = 1$ we can write

$$P_{int+1}(1, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = \frac{1}{1 + \exp(\frac{v_{in}(t+1,0,r_i,c_i,\mu^{pr,m},\delta^{pr,m};H_{in})}{\sigma_\epsilon})}. \tag{14}$$

Given Equations (14) and (11) we have

$$\begin{aligned} V_{in}(t, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) &= \sigma_\epsilon \log(1/P_{int+1}(1, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in})), \\ &= -\log(P_{int+1}(1, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}). \end{aligned} \tag{15}$$

Using Equation (9) and (12), we have

$$
\begin{aligned}
&\log(1/P_{int}(1, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}) - 1) \qquad (16)\\
=&v_{in}(t, 0, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}),\\
=&-c_i + \pi_{in}^{po,m}(t) r_i - (1 - \pi_{in}^{po,m}(t)) \log(P_{int+1}(1, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in}).
\end{aligned}
$$

After rearranging Equation (16), and suppressing the arguments of the choice probabilities for simplification we can write

$$
\log(1/P_{int}) + (1 - \pi_{in}^{po,m}(t)) \log(P_{int+1}) = -c_i + \pi_{in}^{po,m}(t) r_i. \qquad (17)
$$

Now, Consider a subset of callers with sufficient prior experience with the call center (a high number of contacts). Such callers do not have uncertainty about the waiting time distribution and their belief converges to the actual distribution in the data (following Proposition 2). Consequently using (17) for experienced callers in their late contacts, we can write: $\log(1/P_{int}) + (1 - \pi(t)) \log(P_{int+1}) = -c_i + \pi(t) r_i$, where $\pi(t)$ is the hazard rate of the actual waiting time distribution. The hazard rate of the actual waiting time distribution $\pi(t)$ and the abandonment probabilities for experienced callers in their late contacts are identified from the data set. Consequently, given that all variables on the left hand side of this equation are observed (and identified), the variation in the left hand side would identify the reward and cost parameter (r_i,c_i) and their distribution. We briefly explain the intuition behind this. If we assume $y_t = \log(1/P_{int}) + (1 - \pi(t)) \log(P_{int+1})$ and $x_t = \pi(t)$, regressing y_t on x_t would give us the intercept (c_i) and the slope (r_i) and the distribution for them. This shows that, the distribution of the reward and cost parameters can be identified by change in abandonment behavior of experienced callers in their late contacts across different periods. Given the identification of the distribution of r_i and c_i and that we can observe callers' history H_{in}, the parameters of the prior belief distribution $(\mu^{pr,m}, \delta^{pr,m})$ are identified through the change in callers' abandonment behavior across their different contacts. Note that callers' abandonment probabilities depend on $\pi_{in}^{po,m}(s)$, which is a function of $(\mu^{pr,m}, \delta^{pr,m})$, the reward and cost parameter (r_i, c_i) and the history H_{in}. Hence, beyond the reward and cost parameters and caller history callers abandonment behavior across their different contacts is driven by the parameters of the prior belief $(\mu^{pr,m}, \delta^{pr,m})$. Given the identification of the reward and cost parameters and that we observe callers' history, we can identify the prior belief using the variation in callers' abandonment probabilities across their contacts.

5.2.2 Maximum Likelihood Estimation Problem

After estimating the shape parameter of the waiting time distribution k_0 in Section 5.1, we are ready to estimate the rest of the parameters denoted by $\Theta_0 = \{(r^l, c^l, \eta^l)_{\{l=1,2\}}, (\mu^{pr,m}, \delta^{pr,m})_{\{m=1...M\}}\}$. The likelihood of callers i's action in the data is given by

$$
\begin{aligned}
L_i(\Theta_0) = \sum_{l=1}^{2} \eta^l \prod_{n=1}^{n_i} \mathbb{I}_{\{q_{in}=m\}} \prod_{t=0}^{w_{in}} &\Big(P_{int}(1, r^l, c^l, \mu^{pr,m}, \delta^{pr,m}; H_{in})\Big)^{\mathbb{I}_{d_{in,t}=1}} \qquad (18)\\
&\times \Big(1 - P_{int}(1, r^l, c^l, \mu^{pr,m}, \delta^{pr,m}; H_{in})\Big)^{\mathbb{I}_{d_{in,t}=0}}.
\end{aligned}
$$

Note that given that we do not observe r_i and c_i we calculate the product of caller i's choice probabilities across her different contacts for (r^1, c^1) and (r^2, c^2) and then find the linear combincation based on η^1 and η^2. The log-likelihood of the entire population denoted by $\log L(\Theta_0)$ is given by $\log L(\Theta_0) = \sum_{i=1}^{N} \log(L_i(\Theta_0))$. The Maximum Likelihood Estimation problem to find $\Theta_0 = \{(r^l, c^l, \eta^l)_{\{l=1,2\}}, (\mu^{pr,m}, \delta^{pr,m})_{\{m=1...M\}}\}$ is as follows:

$$\underset{\Theta_0}{\text{maximize}} \log L(\Theta_0)$$

$$\begin{aligned}
\text{subject to} \quad & \text{for all } i = 1, ..., N,\ n = 1, ..., n_i,\ t = 0, ..., w_{in}, l = 1, 2: \\
& P_{int}(d_{int}, r_l, c_l, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = \frac{\exp(\frac{v_{in}(t,_{int}, r_l, c_l, \mu^{pr,m}, \delta^{pr,m}; H_{in})}{\sigma_\epsilon})}{1 + \exp(\frac{v_{in}(t,0, r_l, c_l, \mu^{pr,m}, \delta^{pr,m}; H_{in})}{\sigma_\epsilon})}, \\
& V_{in}(t, r_l, c_l, \mu^{pr,m}, \delta^{pr,m}; H_{in}) = \sigma_\epsilon \log\Big[1 + \exp\Big(\frac{v_{in}(t+1, 0, r_l, c_l, \mu^{pr,m}, \delta^{pr,m}; H_{in})}{\sigma_\epsilon}\Big)\Big], \\
& \pi_{in}^{po,m}(t) = 1 - \Big(\frac{\delta_{in}^{po,m}(t) + t^{k_0^m}}{\delta_{in}^{po,m}(t) + (t+1)^{k_0^m}}\Big)^{\mu_{in}^{po,m}(t)},\ \eta_1 + \eta_2 = 1, \\
& \mu_{in}^{po,m}(t) = \mu_{in}^{pr,m} + \mathbb{I}_{\{q_{in}=m\}}(1 - O_{int}),\ \delta_{in}^{po,m}(t) = \delta_{in}^{pr,m} + \mathbb{I}_{\{q_{in}=m\}} t^{k_0^m} \\
& \mu_{i1}^{pr,m} = \mu^{pr,m},\ \delta_{i1}^{pr,m} = \delta^{pr,m},\ \mu_{in+1}^{pr,m} = \mu_{in}^{po,m}(w_{in}),\ \delta_{in+1}^{pr,m} = \delta_{in}^{po,m}(w_{in}).
\end{aligned} \tag{19}$$

Solving this Maximum Likelihood Estimation problem is more challenging and time consuming than the estimation problem under the rational expectation equilibrium assumption as done in Aksin et al. (2013) and Yu et al. (2017). In Aksin et al. (2013) and Yu et al. (2017), the authors consider a rational expectation setting where all callers have the same belief about their chances of receiving service. That is the belief about the chance of entering service ($\pi_{in}^{po,m}(t)$) is the same as the actual hazard rate of the waiting time distribution for all callers in all of their contacts. Consequently, for each set of structural parameters only one calculation of the integrated value function using the recursive formula in (11) is needed. However, in our setting, given that $\pi_{in}^{po,m}(t)$ changes across callers, contacts and period, the recursive formula for the integrated value function in (11) should be calculated for $1 + \sum_{i=1}^{N}(n_i - 1)$ times, where 1 indicates one calculation for the first contact for all callers and $(n_i - 1)$ indicates the number of calculations for each caller for contact two to contact n_i.[13] To solve the Maximum Likelihood Estimation problem we use the non-linear optimization solver KNITRO (Byrd et al. (2006)) with AMPL interface. the integrations are calculated numerically using the Gauss-Hermite quadrature with 5 nodes (Judd (1998)). We solve the maximum likelihood estimation problem for 300 randomly generated starting points to make sure that we are finding the true solution of the maximization problem.[14]

[13] Note that following our assumption that all callers have the same prior belief, the calculation of the integrated value function at the first contact for all callers is the same.

[14] To tackle the computational complexity of the estimation problem we attempted using the Conditional Choice Probability approach (Hotz and Miller (1993), Arcidiacono and Miller (2011)). The intuition behind this approach is using Equation (15) to approximate the integrated value function from data instead of calculating it for each set of structural parameters. However, this approach showed serious identification problems because of a large set of state variables in our case resulted from different customer histories. Our Monte-Carlo studies shows that the CCP approach cannot recover the true parameters of callers. However, our "Brute Force" approach of calculating the integrated value functions can identify all parameters.

5.3 Estimation Results for the prior belief, reward and cost parameters

Using the MLE in Section 5.2.2, we estimate $\Theta_0 = \{(r^l, c^l, \eta^l)_{\{l=1,2\}}, (\mu^{pr,m}, \delta^{pr,m})_{\{m=1...M\}}\}$ for each priority group. Tables 4 and 5 shows the estimation results. The numbers in parenthesis in Tables 4 and 5 show the standard errors of the estimates. To calculate the standard errors we use the nonparametric bootstrap method by taking samples by replacements from the pool of callers (not the pool of observations). If a caller is chosen to be in a sample, all contacts of the caller will be included.

Priority group	η^1	c^1	r^1	η^2	c^2	r^2	Mean-c	Mean-r
High-priority	0.92 (0.07)	0.95 (0.12)	7.56 (0.73)	0.08 (0.07)	0.18 (0.06)	4.14 (0.52)	0.89	7.29
Medium-priority	0.81 (0.05)	0.96 (0.08)	7.45 (0.56)	0.19 (0.05)	1.11 (0.14)	6.02 (0.88)	0.99	7.18
Low-priority	0.85 (0.08)	1.01 (0.15)	6.74 (0.72)	0.16 (0.08)	1.21 (0.23)	5.68 (0.95)	1.04	6.88

Table 4: The estimates for callers' reward and cost parameter. Note that Mean-c=$\eta^1 c^1 + \eta^2 c^2$,and Mean-r=$\eta^1 r^1 + \eta^2 r^2$.

Priority group	$\mu^{pr,NR}$	$\delta^{pr,NR}$	$\mu^{pr,R}$	$\delta^{pr,R}$
High-priority	1497.84 (310.23)	880.65 (201.36)	1220.58 (110.78)	774.64 (214.58)
Medium-priority	5120.23 (540.36)	3024.81 (542.64)	1730.61 (254.68)	1092.98 (198.849)
Low-priority	3028.80 (132.56)	1598.45 (242.99)	1484.12 (150.77)	928.02 (235.01)

Table 5: The estimates for callers' prior belief parameters.

The log-likelihood value for the MLE problems of the high, medium and low priority groups are -6,800.89, -30,238.63 and -65,644.61, respectively. The estimation results in Tables 4 and 5 lead to the three insights as follows:

First insight: Callers are optimistic about their chances of receiving service irrespective of their priority group. Using the estimates in Table 5 we calculate callers' prior predictive distributions and their averages for the Rush-hour and Non-Rush-hour day intervals. We calculated these averages using the cdf of the actual and the prior predictive distributions. The cdf of the prior predictive distributions are computed using the estimates in Table 5 and Equation (4). We found the cdf of callers' actual waiting time distributions using the Kaplan-Meier estimator (Kaplan and Meier (1958)) from observation in our data set. Table 6 shows the comparison of the actual average waiting times and callers beliefs about it for all priority groups. As can be seen in Table 6 callers who do not have any experience with the call center irrespective of their priority class believe that their average waiting times will be less than 15 seconds. However, the actual average delays are

much longer. This agrees with the fact that in this call center callers are not aware of their priority class and in the first contact could have the same belief.

Priority group	Prior belief (sec.)	Actual value (sec.)
High-priority (NR)	12.83	48.81
High-priority (R)	12.54	40.94
Medium-priority (NR)	12.90	72.37
Medium-priority (R)	13.26	55.36
Low-priority (NR)	12.24	99.85
Low-priority (R)	13.14	75.32

Table 6: Comparison of callers' actual average waiting time versus their prior belief about it calculated using callers' prior predictive distribution.

Figures 7 to 9 show the comparison between the actual waiting time distribution and callers' prior predictive distributions for the high to low priority groups for the Rush-hour and Non-Rush-hour day intervals. As can be seen in these figures, callers' prior predictive is optimistic and callers believe they will receive service much earlier than what actually happened. We performed the one sided Kolmogorov-Smirnov test to compare the actual and prior predictive distributions, and for all cases, we could reject the null hypothesis that the actual and predictive distributions are the same with 99% confidence level.

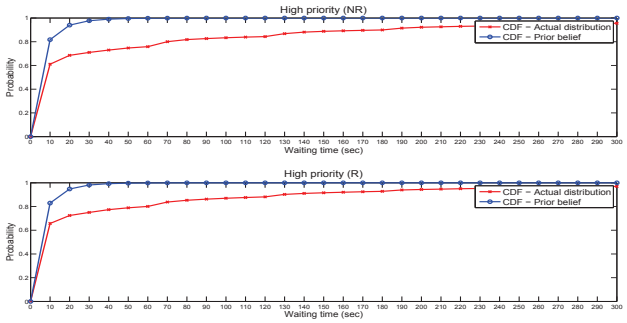

Figure 7: The comparison of the cdf of the actual distribution for callers' waiting times and their prior predictive for the waiting time for the high priority callers.

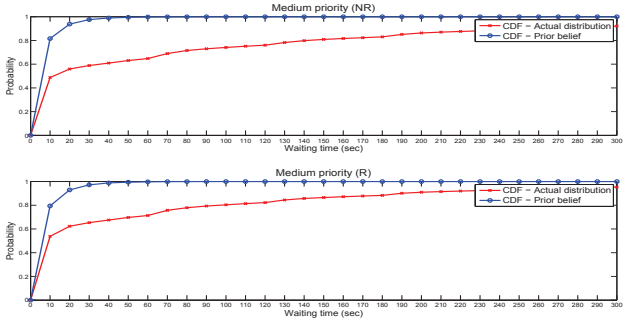

Figure 8: The comparison of the cdf of the actual distribution for callers' waiting times and their prior predictive for the waiting time for the medium priority callers.

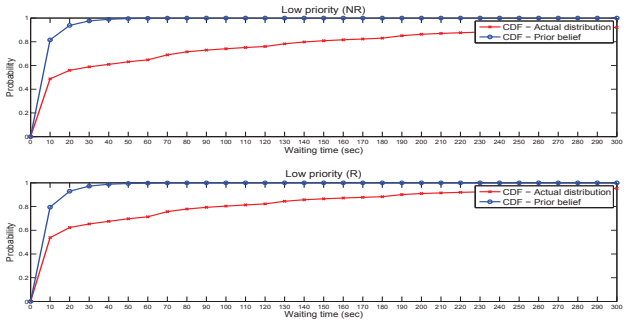

Figure 9: The comparison of the cdf of the actual distribution for callers' waiting times and their prior predictive for the waiting time for the low priority callers.

Second insight: Low priority callers are the least patient ones, and the High priority callers are the most patient ones. Based on the mean of the reward and cost parameters in Table 4, high priority callers have the highest value for service and the Low priority callers have the lowest value for service. Moreover, high priority callers have the lowest waiting cost and low priority callers have the highest waiting cost. The average of the ratios of the reward and cost parameter across the two caller segments for the high, medium and low priority callers ($\eta^1 r^1/c^1 + \eta^2 r^2/c^2$) are 7.37, 6.48 and 6.07 , respectively. The ratio of the cost and reward parameters can be considered as a proxy for callers' patience threshold. Consequently, the estimation results show that the ranking of the priority groups from the most patient to the least patient is High, Medium and Low.

To see if the ranking of the priority groups inferred from their reward and cost estimates agrees with what data tells us using a non parametric approach (and model free approach), we estimated the survival probabilities of the priority groups using the Kaplan-Meier estimator. Figure 10 shows the survival curves for the high, medium and low priority groups.

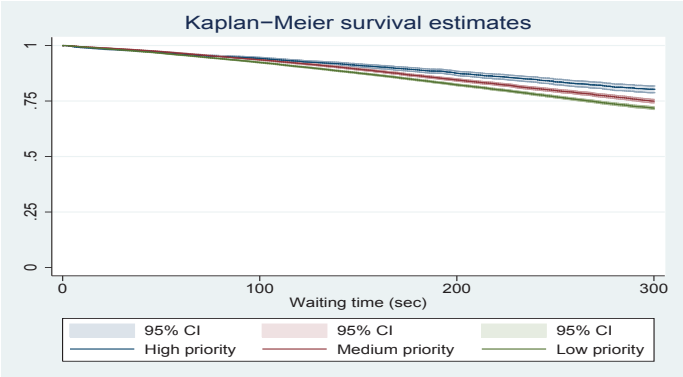

Figure 10: The survival curves for the high, medium and low priority groups estimated using the Kaplan-Meier approach.

As can be see in Figure 10, the high priority callers have the highest survival probabilities (lowest abandonment probabilities) and the low priority callers have the lowest survival probabilities. In other words, the ranking of priority groups in terms of callers' patience level inferred from our Bayesian estimates in Table 4 matches the ranking inferred from the non-parametric Kaplan-Meier estimator shown in Figure 10.

Third insight: The difference between callers' prior beliefs across different priority groups is not as significant as the actual difference in their waiting time distributions. As can be seen in Table 6 even though the difference between the actual waiting times across different priority groups is significant, the difference between callers' prior predictive beliefs about the average waiting times is small. In other words, callers from different priority groups who are first time callers are not significantly different in terms of their belief about the waiting time distribution. We can see this phenomenon more clearly in Figure 11, which shows the comparison of the actual and prior predictive distributions across different priority groups.

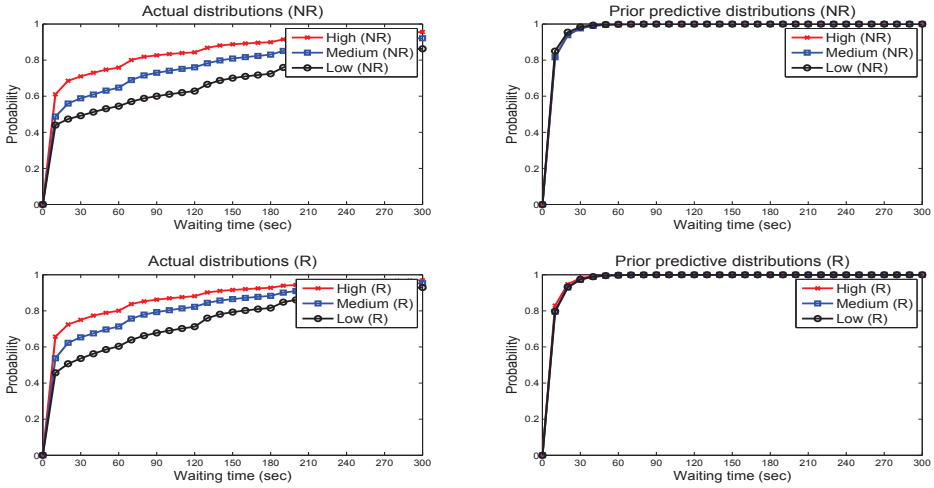

Figure 11: The comparison of the actual and prior predictive distributions across different priority groups.

6 Comparison Between the Bayesian Learning Model and the Rational Expectation Model

In this section we compare the bayesian learning model introduced in this paper with the rational expectation model in Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017), which is the state of the art in the literature. We first explain how the parameters of the model under the rational expectation assumption are estimated. Then we estimate the parameters of the rational expectation model. And finally, we compare the bayesian learning model and the rational expectation model in four aspects: fit to the data set, inference about patience levels, the prediction power and estimation bias.

Estimating the parameters under the rational expectation assumption. In the rational expectation model (Aksin et al. (2013), Aksin et al. (2017) and Yu et al. (2017)) the main assumption is that call callers irrespective of their contact history know the actual distribution in the call center. Consequently, the model for callers' abandonment behavior would be similar to the model presented in Section 4.2 with one difference: callers' belief about their chances of receiving service $\pi_{in}^{po,m}$ matches the actual hazard rate of the waiting time distribution for all callers across

all contacts. Consequently, we do not estimate a prior belief, and the structural parameters of the model will be only the parameters of the cost and reward denoted by $\Theta_0^{RE} = \{(\eta^l, r^l, c^l)_{\{l=1,2\}}\}$. The details of the estimation procedure for the rational expectation model is explained in Appendix D. Table 7 shows the estimation results under the rational expectation assumption.

Priority group	η^1	c^1	r^1	η^2	c^2	r^2	Mean-c	Mean-r
High-priority	0.98 (0.09)	0.07 (0.02)	6.06 (0.73)	0.02 (0.09)	0.00 (0.02)	1.91 (0.23)	0.07	5.97
Medium-priority	0.88 (0.07)	0.08 (0.03)	6.35 (0.87)	0.12 (0.07)	0.17 (0.06)	5.81 (0.79)	0.09	6.28
Low-priority	0.83 (0.06)	0.01 (0.01)	5.17 (0.48)	0.17 (0.06)	0.05 (0.01)	4.17 (0.66)	0.02	5.00

Table 7: The estimates for callers' reward and cost parameter.

The log-likelihood values for the MLE problems of the high, medium and low priority groups are -6,852.79, -30,682.90 and -66,550.72, respectively.

Comparison of the fit of the bayesian learning and the rational expectation models. Table 8 shows the comparison between the AIC and BIC values for the bayesian learning and rational expectation models. As can be seen in Table 8 for all priority groups the AIC and BIC values

Priority group	Bayesian L. (AIC)	Rational Exp. (AIC)	Bayesian L. (BIC)	Rational Exp. (BIC)
High-priority	13,613.79	13,717.60	13,664.66	13,726.55
Medium-priority	60,489.27	61,377.80	60,546.51	61,388.89
Low-priority	131,301.22	133,113.43	131,360.35	133,125.14

Table 8: The AIC and BIC values for the maximum likelihood estimation problems of the bayesian learning and rational expectation models.

for the bayesian learning model are less than those of the rational expectation model. This shows that the bayesian learning model has a better fit to the data set.

Comparison of the inference about callers' patience levels. Based on the estimation results of the rational expectation model in Table 7 the averages of the ratio of the reward and cost parameter across the two caller segments for the high, medium and low priority callers ($\eta^1 r^1/c^1 + \eta^2 r^2/c^2$) are 87.71, 67.12 and 403.72, respectively. In other words, the ranking of the priority groups in terms of their patience level using the rational expectation approach from the highest patience to the lowest patience is low, high and medium. This ranking does not agree with what the non-parametric Kaplan-Meier estimator tells us in Figure 10. Recall from the second insight in Section 5.3 that the bayesian learning model tells us that the ranking from the highest patience to the lowest patience is high, medium and low priority groups, which matches with the results of the Kaplan-Meier estimator. This shows that the rational expectation assumption may lead to poor inferences in terms of callers' patience levels.

Comparison of the prediction power. To compare the prediction power of the bayesian learn-

ing model with the rational expectation model, we perform out of sample test. To do so we divide the data sets across callers to two sets randomly: a training set and a test set. We use the observations in the training set to estimate the parameters of the model, which include $((\eta^l, r^l, c^l)_{\{l=1,2\}}$, $\mu^{pr,NR}$,$\mu^{pr,R}$,$\delta^{pr,NR}$,$\delta^{pr,R})$ for the bayesian learning model and $(\eta^l, r^l, c^l)_{\{l=1,2\}}$ for the rational expectation model. Then, we use the estimated parameters and the corresponding model to predict callers' abandonment behavior in the test set. In the bayesian learning model, we use the prior distribution in the training set to find callers' belief in each contact in the test set. However, in the rational expectation model, consistent with the assumption that callers know the actual waiting time distribution, we estimate the waiting time distribution for the rush-hour and non-rush-hours in the test set and assume these distributions are common knowledge across all callers. Table 9 shows the absolute and relative errors in predicting the abandonment rates for the bayesian learning model and the rational expectation model, respectively.

Priority group	Relative Error (Bayesian L.)	Absolute Error (Bayesian L.)	Relative Error (Rational Exp.)	Absolute Error (Rational Exp.)
High-priority	8.19 %	0.24 %	17.38 %	0.51 %
Medium-priority	4.30 %	0.22 %	28.86 %	1.45 %
Low-priority	4.29 %	0.36 %	13.04 %	1.10 %
Average across all tests	5.59 %	0.27 %	19.70 %	1.02 %

Table 9: The relative and absolute errors in predicting the abandonment rates for the bayesian learning model and the rational expectation model.

As can be seen in Table 9, the average of the errors for the bayesian learning model is less than third of that of the rational expectation model. We repeated this analysis for other randomly selected training and test sets and got the same results.

Comparison of the estimation bias. In this section using a Monte-Carlo simulation study, we show that if callers' belief about the waiting time distribution does not match the actual distribution, estimating the parameters of the optimal stopping time model under the rational expectation assumption would lead to biased estimates. While our bayesian learning framework can recover the true parameters without any bias. To do so, we create 100 simulated data sets in which callers' belief about the waiting time distribution is optimistic and does not match the actual waiting time distribution in the data. Then we estimate back the true parameters of callers using both the rational expectation model and the bayesian learning model and compare the results.

We created 100 simulated data sets with 20,000 callers in each data set. There are two day intervals and callers have the same probability of contacting in these intervals. The distribution of callers' frequency of contact is estimated empirically from the data set. We assume that the actual waiting time distribution of the first and the second day intervals are $Weibull(1.5, 40)$ and $Weibull(2, 40)$, respectively. Callers updating scheme and abandonment behavior are the same as those explained in Section 4. Callers are evenly split to type 1 and type 2; i.e. $\eta^1 = \eta^2 = 0.5$. The reward and cost parameters of type 1 and type 2 callers are $(r^1 = 8,\ c^1 = 0.3)$ and $(r^1 = 6,\ c^1 = 0.5)$, respectively. Furthermore, their prior belief about the scale parameters of the waiting time distri-

butions in the day interval one and two are $Inverse-gamma(5,20)$ and $Inverse-gamma(10,40)$, respectively. Note that the average of callers' prior expectation about the scale parameter of the waiting time distributions in day interval and two are 5 and 4.44, which are much smaller than the actual values (40). Consequently, callers are optimistic about the waiting time durations. Table 10 shows the true and estimated value for the reward and cost parameters under the rational expectation assumption.

Priority group	η^1	c^1	r^1	η^2	c^2	r^2
True parameters	0.50	0.30	8.00	0.50	0.50	6.00
Mean (Simulated data)	0.69	0.10	6.68	0.31	0.20	6.53
standard deviation (Simulated data)	0.06	0.02	0.40	0.06	0.01	0.20
Upper bound (95% CI) (Simulated data)	0.80	0.15	7.47	0.42	0.22	6.94
Lower bound (95% CI) (Simulated data)	0.58	0.06	5.89	0.20	0.17	6.13

Table 10: The results of the Monte-Carlo study for the rational expectation model.

As can be seen in Table 10 the mean of the estimates are not close to the actual values and the actual values are not in the confidence intervals of the estimates either. This shows that estimating the parameters using the rational expectation framework may lead to biased estimates in settings that callers' belief does not match the actual distribution in the data, which might be the case in most of the real world situations.

We estimated the parameters of the 100 simulated data sets using our bayesian learning framework as well. Table 11 shows the results of the Monte-Carlo simulation study for the bayesian learning model.

Priority group	η^1	c^1	r^1	η^2	c^2	r^2	$\mu^{pr,1}$	$\delta^{pr,1}$	$\mu^{pr,2}$	$\delta^{pr,2}$
True parameters	0.50	0.30	8.00	0.50	0.50	6.00	5.00	20.00	10.00	40.00
Mean (Simulated data)	0.47	0.29	8.21	0.51	0.46	5.92	4.86	20.96	11.24	41.23
standard deviation (Simulated data)	0.04	0.03	0.54	0.04	0.09	0.41	0.69	3.24	1.33	6.63
Upper bound (95% CI) (Simulated data)	0.55	0.35	9.27	0.59	0.64	6.72	6.21	27.31	13.85	54.22
Lower bound (95% CI) (Simulated data)	0.39	0.23	7.15	0.43	0.28	5.12	3.51	14.61	8.63	28.24

Table 11: The results of the Monte-Carlo study for the bayesian learning model.

As can be seen in Table 11, the mean of the estimated parameters of the simulated data sets are close to the true values and all true values are in the confidence intervals constructed from the estimates of the simulated data sets. In other words, the bayesian learning framework is capable of recovering the true parameters of the simulated dataset without any bias.

The analyses in this section show that the bayesian learning model not only has a better statistical fit compared to the rational expectation model but also it has a better prediction power, which is extremely important in performing what-if analyses and policy experiments. Moreover, in

contrast to the rational expectation framework the bayesian learning framework does not lead to biased estimates or poor predictions about caller patience levels.

7 Managerial Applications

The introduced Bayesian learning model and the estimation framework provides a deeper understanding of callers' expectation about their delays and the impact of their past interactions on their abandonment behavior. It also provides several opportunities for the call center managers to improve the performance measures of the call center and the firm-customer relationship. In what follows we discuss two applications of the framework introduced in this paper:

Managing customers' expectation about their delays: An important measure of the service quality in call centers is callers' delays. Therefore, the call center manager can get a better understanding of callers' expectations about the service quality by acquiring more information on customers' beliefs about their delays in the system.

Our framework can be used to estimate callers' prior beliefs about their delays. The estimation results show that in this specific call center callers are generally optimistic about their chances of receiving service.[15] In other words, callers overestimate the service quality in the call center. In this setting, providing delay information may actually increase callers' abandonments because it moves callers' prior belief toward the actual distribution which entails longer delays. Consequently, the manager may decide to not provide any delay information in this call center. However, in a setting that callers are pessimistic about their chances of receiving service providing delay information via delay announcements shifts callers' expectation and aligns it with the actual delay distribution in the call center. Given that callers' prior expectation is pessimistic, providing delay announcements lowers callers' abandonment probabilities. In addition, such information can be provided in a customized fashion for each caller.

To show the significance of impact of callers' prior belief on their abandonment behavior, we find the impact of changing the location parameters in callers' prior belief distribution ($\delta^{NR,pr}$ and $\delta^{R,pr}$) about the scale parameter of the waiting time distribution on their abandonment rates in the data. Note that given that the mean of the Inverse-gamma distribution is proportional to its location parameter, changing $\delta^{NR,pr}$ and $\delta^{R,pr}$ would proportionally impact callers belief about the mean of the location parameter of the actual waiting time distribution, which in turn would impact callers' belief about their waiting duration.

We change the location parameters in callers' prior belief by assuming $\delta^{NR,pr}_{new} = \beta \times \delta^{NR,pr}$ and $\delta^{R,pr}_{new} = \beta \times \delta^{R,pr}$ for $\beta = 0.9,\ 1$ and 1.1, and use our model to predict callers' abandonment probabilities and the total abandonment rate in the data. Table 12 shows the predicted abandonment rates in the data for different values of β.

As can be seen in Table 12 inflating or deflating β would increase and decrease the abandonment probabilities of callers, respectively. Note that increasing (decreasing) β would lead to a higher

[15] We cannot generalize this finding to all call centers, and consequently, cannot roll out the possibility of customers being pessimistic about the duration of their delays.

Predicted abandonment rates	$\beta = 1$	$\beta = 0.9$ (Relative change compared to $\beta = 1$)	$\beta = 1.1$ (Relative change compared to $\beta = 1$)
High-priority	3.15%	2.95% (-6.02%)	3.36% (6.73%)
Medium-priority	5.10%	4.71% (-7.56%)	5.58% (9.21%)
Low-priority	8.26%	7.41% (-10.31%)	8.89% (7.55%)

Table 12: The predicted abandonment rates for different values of β. The values in parenthesis show the change in the abandonment rates relative to the case with $\beta = 1$.

(lower) mean for the scale parameter of callers' belief about the waiting time distribution, which results in callers expecting to have a longer (shorter) duration of waiting. Furthermore, if callers believe the waiting duration will be longer (shorter) they abandon with a higher (lower) probability. Increasing β by 10% from 1 to 1.1 would increase the abandonment rates of the high, medium and low priority groups by 6.73%, 9.21% and 7.55%, respectively. Decreasing β by 10% from 1 to 0.9 would decrease the abandonment rates of the high, medium and low priority groups by 6.02%, 7.75% and 10.31%, respectively. Hence, callers' prior belief would impact their abandonment rates in the call center, and influencing this belief would change the performance measures in the call center.

Changing the scheduling policy based on callers' patience level: A growing body of literature in Operations Management has shown that modifying the scheduling policy of the call center based on callers' abandonment probabilities or their remaining patience threshold leads to a significant improvement in call center performance measures.

Mandelbaum and Momcilovic (2014) investigate the impact of the Least-Patience-First (LPF) policy on system performance and show that the LPF policy can provide significant improvements over the First-Come-First-Served (FCFS) policy by decreasing the abandonment rate. The authors assume that the call center manager knows the patience thresholds of all callers. Bassamboo and Randhawa (2014) propose a Time-In-Queue policy that prioritizes customers based on the amount of time they have been waiting in the system. The authors assume that customers' patience thresholds are drawn from the same distribution, and how long they have been waiting reveals some information about their individual patience threshold. The authors show that the new policy can significantly improve the performance measures in the system such as the average waiting time and the queue length. The framework in the current work complements the aforementioned literature as follows:

Unlike Mandelbaum and Momcilovic (2014) and Bassamboo and Randhawa (2014), we do not assume that callers' individual patience threshold (or its distribution) are perfectly known and are exogenously given. But we provide a framework that utilizes callers' contact history data to calculate two types of caller-specific information about their patience. These two types of

information may differ across callers depending on their contact history. The first type of caller-specific information acquired using our framework is the expected abandonment probability of the callers in period t of their n^{th} contact $\mathbb{E}[P_{int}(d, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in})]$ that can be calculated using the estimated distribution for callers reward and cost parameter. Consequently, for each caller, given her contact history, we can find her expected abandonment probability in each period of waiting. To be more specific, we have

$$\mathbb{E}[P_{int}(d, r_i, c_i, \mu^{pr,m}, \delta^{pr,m}; H_{in})] = \sum_{l=1}^{2} \eta^l P_{int}(d, r^l, c^l, \mu^{pr,m}, \delta^{pr,m}; H_{in}). \tag{20}$$

The second type of caller-specific information calculated using our framework is callers' expected patience threshold (abandonment time). Denote by $\mathbb{E}[G_{in}(t, r, c, \mu^{pr,m}, \delta^{pr,m}; H_{in})]$ the expected abandonment time distribution of caller i in her n^{th} contact that occurred in the day interval m, i.e. $q_{in} = m$. We have

$$\begin{aligned} &\mathbb{E}[G_{in}(t, r, c, \mu^{pr,m}, \delta^{pr,m}; H_{in})] = \mathbb{E}[G_{in}(t-1, r, c, \mu^{pr,m}, \delta^{pr,m}; H_{in})] \\ &+ \mathbb{E}\Big[(1 - G_{in}(t-1, r, c, \mu^{pr,m}, \delta^{pr,m}; H_{in})) \times (1 - \pi^m(t; k_0^m, \gamma_0^m)) P_{int}(1, r, c, \mu^{pr,m}, \delta^{pr,m}; H_{in})\Big]. \end{aligned} \tag{21}$$

where $\pi^m(t; k_0^m, \gamma_0^m)$ is the hazard rate of the actual waiting time distribution in day interval m ($F^m(0; k_0^m, \gamma_0^m)$ in (1)). The right-hand side of (21) is the sum of two terms: 1) the probability of abandoning by $t-1$, and 2) the probability of not abandoning by $t-1$, not receiving service at t and abandoning at t. Finding the mean of $\mathbb{E}(G_{in}(t, r, c, \mu^{pr,m}, \delta^{pr,m}; H_{in}))$ gives the manager a proxy for caller i's patience threshold that depends on her contact history H_{in}. This can enable the manager to implement patience-based priority policies, in which callers may get a higher or a lower priority based on their patience thresholds.

8 Concluding Remarks

Understanding callers' abandonment behavior using their contact history data is an opportunity for call center managers that has been overlooked in the Operations Management literature. In this paper we try to fill in this gap using a Bayesian learning framework. We first show that callers' past interactions with the system reveal information about callers' preferences and their chance of abandonment in the future. Then to separate the impact of callers' patience parameters (waiting cost and valuation for service) from the impact of their contact history, we use a structural estimation approach in a Bayesian learning setting.

In the extant literature about callers' behavior in call centers all contacts of a caller are treated the same, and it is assumed that callers know the exact waiting time distribution in the data. To the best of our knowledge this paper is the first work to relax that assumption and try to accommodate a situation where callers' expectation about their chances of receiving service does not necessarily match their actual values. We provide a Bayesian framework for callers' learning

where callers update their belief about the distribution of a waiting time distribution parameter based on their waiting experiences in the call center. We also provide an estimation approach to use the observation in a data set to estimate callers' parameters that shed light on callers' expectations and preferences. These parameters are callers' waiting cost and valuation for service, and callers' prior belief about their waiting time distribution. Our estimation results show that callers are optimistic about their delays in the system and underestimate their durations. We also show that new callers irrespective of their priority class believe that they will receive service in less than 15 seconds even though their actual average waiting time ranges between 40 and 90 seconds. The introduced framework can be used to manage customers' expectation about the service quality and to implement the patience-based scheduling policies. We compare the bayesian learning model introduced in this work with the rational expectation framework in the extant literature and show that our framework not only has a better fit to the data set bit also it has a better out of sample performance. In addition, in contrast to the rational expectation framework our framework does not produced biased estimated for callers' reward and cost parameters.

Our paper identifies several opportunities for future research. One interesting area of research is considering the impact of the service provided by the agent on callers' behavior. In our data set we cannot observe the quality of agents' work; consequently, we did not attempt to include learning about this part of callers' service encounter on their behavior. In addition, we show that changing callers' prior belief may have a significant effect on their abandonment behavior and total abandonment rates in the call center. One may study practical ways to impact new callers' belief about the waiting durations.

References

Ackerberg, D. A. (2003). Advertising, learning, and consumer choice in experience good markets: an empirical examination. *International Economic Review 44*(3), 1007–1040.

Agrawal, N. and S. A. Smith (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics (NRL) 43*(6), 839–861.

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

Aksin, Z., B. Ata, S. Emadi, and C. Su (2017). Impact of delay announcements in call centers: An empirical approach. *Operations Research 65*(1), 242–265.

Arcidiacono, P. and R. A. Miller (2011). Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica 79*(6), 1823–1867.

Bassamboo, A. and R. S. Randhawa (2014). Scheduling homogeneous impatient customers. *Working paper, Available at SSRN 2312643*.

Braden, D. J. and M. Freimer (1991). Informational dynamics of censored observations. *Management Science 37*(11), 1390–1404.

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

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

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187–220.

Czinkota, M. and I. Ronkainen (2012). *International marketing*. Cengage Learning.

Ding, X., M. L. Puterman, and A. Bisi (2002). The censored newsvendor and the optimal acquisition of information. *Operations Research 50*(3), 517–527.

Dubé, J.-P., G. J. Hitsch, and P. E. Rossi (2010). State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics 41*(3), 417–445.

Eckstein, Z., D. Horsky, and Y. Raban (1988). An empirical dynamic model of optimal brand choice. *Foerder Institute of Economic Research, Working Paper* (88).

Erdem, T. and M. P. Keane (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing science 15*(1), 1–20.

Erdem, T., M. P. Keane, and B. Sun (2008). A dynamic model of brand choice when price and advertising signal product quality. *Marketing Science 27*(6), 1111–1125.

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

Harpaz, G., W. Y. Lee, and R. L. Winkler (1982). Learning, experimentation, and the optimal output decisions of a competitive firm. *Management Science 28*(6), 589–603.

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

Heckman, J. J. (1979). *Statistical models for discrete panel data*. Department of Economics and Graduate School of Business, University of Chicago.

Heckman, J. J. and B. Singer (1982). The identification problem in econometric models for duration data. *Advances in econometrics 39*.

Heese, H. S. and J. M. Swaminathan (2010). Inventory and sales effort management under unobservable lost sales. *European Journal of Operational Research 207*(3), 1263–1268.

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

Hotz, V. J. and R. A. Miller (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies 60*(3), 497–529.

Ibrahim, J. G., M.-H. Chen, and D. Sinha (2005). *Bayesian survival analysis*. Wiley Online Library.

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

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

Lariviere, M. A. and E. L. Porteus (1999). Stalking information: Bayesian inventory management with unobserved lost sales. *Management Science 45*(3), 346–363.

Lazarsfeld, P. F., N. W. Henry, and T. W. Anderson (1968). *Latent structure analysis*, Volume 109. Houghton Mifflin Boston.

Mandelbaum, A. and P. Momcilovic (2014). Personalized queues: The customer view, via least-patient-first routing. *Submitted for Publication*.

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

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1 50*(3), 163–170.

Mersereau, A. J. (2015). Demand estimation from censored observations with inventory record inaccuracy. *Manufacturing & Service Operations Management 17*(3), 335–349.

Nahmias, S. (1994). Demand estimation in lost sales inventory systems. *Naval Research Logistics 41*(6), 739–758.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy 9*(4), 513–548.

Rossi, P. E., R. E. McCulloch, and G. M. Allenby (1996). The value of purchase history data in target marketing. *Marketing Science 15*(4), 321–340.

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

Wecker, W. E. (1978). Predicting demand from sales data in the presence of stockouts. *Management Science 24*(10), 1043–1054.

Yu, Q., G. Allon, and A. Bassamboo (2017). How do delay announcements shape customer behavior? an empirical study. *Management Science 63*(1), 1–20.

Online Appendix for "Impact of Callers' History on Abandonment: Model and Implications"

A Stratified Cox Regression to Study the Impact of Past History on Caller Abandonment

In this section, we provide more details about the stratified Cox regression that was used to study the impact of callers' history on their abandonment behavior.

Suppose that $h_{i,k}(t)$ is the hazard rate of the abandonment time distribution for caller i from stratum k with the set of independent variables X_i. Note that the stratum in our setting are the priority groups. The vector X_i includes variables that may impact the caller's abandonment behavior. In our setting X_i includes whether callers abandoned or not in their last contact, their waiting time in the last contact, the interaction term between call outcome and the waiting time variable, the time of the day and week day dummy variables, and the contact number. In the Cox regression analysis we assume that

$$h_{i,k}(t) = h_{0,k}(t)\ \exp(X_i\beta^T), \tag{22}$$

where $h_{0,k}(t)$ is the baseline hazard function for stratum k that can be any function of t as long as $h_{0,k}(t) > 0$. The function $h_{0,k}(t)$ does not have to be specified. The vector β is the vector of coefficients of the independent variables, which captures the impact of the independent variables on callers' abandonment hazard rates. Given that the exponent term in (22) does not involve a time variable, the ratio of hazard functions of two callers does not depend on time and $h_{0,k}(t)$. To be more specific, the ratio of hazard function of callers i and j is given by

$$\frac{h_{i,k}(t)}{h_{j,k}(t)} = \exp((X_i - X_j)\beta^T). \tag{23}$$

As can be seen in (23) in the Cox regression analysis it is assumed that the ratio of hazard functions of two callers from the same stratum only depend on the difference between their independent variables. This assumption is called the proportional hazard rate assumption.

B Additional Survival Analyses

For robustness check we perform the cox regression for each priority group in isolation. Note that in Section 3.2 we use a stratified cox regression, which assumes different baseline hazard function for different priority groups but would lead to the same coefficients for all priority groups (parameter β in (23)). If we run the regression for each priority group separately, we not only consider different baseline hazard functions but also we will get different parameters. Tables 13 to 15 show the cox regression results for each priority group separately.

Table 13: The results of the Cox regression for the high priority callers.

Variable	Coefficient	(Std. Err.)
O_{-1}	0.6272**	(0.1329)
W_{-1}	-0.0017**	(0.0003)
$O_{-1} \times W_{-1}$	0.0013**	(0.0005)
$D_{Rush-hour}$	-0.1571**	(0.0723)
Weekdays	-0.0618	(0.1261)
Contact number	0.0047**	(0.0013)

** Denotes statistically significant at 0.05.

Table 14: The results of the Cox regression for the medium priority callers.

Variable	Coefficient	(Std. Err.)
O_{-1}	0.7892**	(0.0527)
W_{-1}	-0.0013**	(0.0002)
$O_{-1} \times W_{-1}$	0.0006**	(0.0002)
$D_{Rush-hour}$	-0.0207	(0.0317)
Weekdays	-0.0132	(0.0533)
Contact number	0.0083**	(0.0011)

** Denotes statistically significant at 0.05.

As can be seen in Table 13 to 15, the sign and significance of the coefficients for the past history variables (O_{-1}, W_{-1} and $O_{-1} \times W_{-1}$) are the same as those in Table 2.

C Proofs

Proof of Proposition 1. Without loss of generality we suppress the superscript for the day interval m. Given that the distribution of the waiting time is $Weibull(k_0, \gamma_0)$ and caller i's posterior belief about γ_0 by time t is $Inv-Gamma(\mu_{in}^{po}(t), \delta_{in}^{po}(t))$, the cdf of caller i's prior predictive distribution denoted by $F_{in}^{pr}(t)$ is given by

$$F_{in}^{po}(t) = \int_0^\infty (1 - e^{-\frac{t^{k_0}}{\gamma}})(\frac{\delta_{in}^{po}(t)^{\mu_{in}^{po}(t)}}{\Gamma(\mu_{in}^{po}(t))} \frac{e^{-\frac{\delta_{in}^{po}(t)}{\gamma}}}{\gamma^{(\mu_{in}^{po}(t)+1)}})\ d\gamma, \tag{24}$$

where the first term inside the integral is the cdf of the waiting time distribution and the second term is the pdf of the inverse gamma distribution with the parameters $\mu_{in}^{po}(t)$ and $\delta_{in}^{po}(t)$. We can rewrite (24) as follows:

$$\begin{aligned} F_{in}^{po}(t) &= \int_0^\infty (\frac{\delta_{in}^{po}(t)^{\mu_{in}^{po}(t)}}{\Gamma(\mu_{in}^{po}(t))} \frac{e^{-\frac{\delta_{in}^{po}(t)}{\gamma}}}{\gamma^{(\mu_{in}^{po}(t)+1)}})\ d\gamma - \int_0^\infty e^{-\frac{t^{k_0}}{\gamma}} (\frac{\delta_{in}^{po}(t)^{\mu_{in}^{po}(t)}}{\Gamma(\mu_{in}^{po}(t))} \frac{e^{-\frac{\delta_{in}^{po}(t)}{\gamma}}}{\gamma^{(\mu_{in}^{po}(t)+1)}})\ d\gamma, \\ &= 1 - \frac{\delta_{in}^{po}(t)^{\mu_{in}^{po}(t)}}{(\delta_{in}^{po}(t) + t^{k_0})^{\mu_{in}^{po}(t)}} \int_0^\infty (\frac{(\delta_{in}^{po}(t) + t^{k_0})^{\mu_{in}^{po}(t)}}{\Gamma(\mu_{in}^{po}(t))} \frac{e^{-\frac{\delta_{in}^{po}(t)+t^{k_0}}{\gamma}}}{\gamma^{(\mu_{in}^{po}(t)+1)}})\ d\gamma. \end{aligned} \tag{25}$$

Table 15: The results of the Cox regression for the low priority callers.

Variable	**Coefficient**	**(Std. Err.)**
O_{-1}	0.7173**	(0.0354)
W_{-1}	-0.0016**	(0.0001)
$O_{-1} \times W_{-1}$	0.0006**	(0.0002)
$D_{Rush-hour}$	-0.0531**	(0.0213)
Weekdays	0.0833**	(0.0343)
Contact number	0.0044**	(0.0007)

** Denotes statistically significant at 0.05.

The term inside the integration in (25) is the pdf of an inverse gamma distribution with parameters $\mu_{in}^{po}(t)$ and $\delta_{in}^{po}(t) + t^{k_0}$; consequently, it is equal to 1. Therefore, we have

$$F_{in}^{po}(t) = 1 - (\frac{\delta_{in}^{po}(t)}{\delta_{in}^{po}(t) + t^{k_0}})^{\mu_{in}^{po}(t)}. \tag{26}$$

Given, (26) the calculation of $\pi_{in}^{po}(t)$ in (5) is trivial.

Next, we characterize the Bayesian updating. Denote by $I_{in}(t)$ the information caller i has acquired about the waiting time in her n^{th} contact by time t. Furthermore, denote by s_{in} the amount of time caller i has to wait to enter the service stage in her n^{th} contact, which is a draw from the actual waiting time distribution $Weibull(k_0, \gamma_0)$. Recall that the outcome of caller i's contact by time t is O_{int}. If $O_{int} = 1$ the caller has abandoned or has not received service yet. Consequently, she believes that the waiting time in her n^{th} contact is longer than t, which means $I_{in}(t) = \{s_{in} > t\}$. However, if $O_{int} = 0$ the caller has received service by time t. Consequently, we have $I_{in}(t) = \{s_{in} = t\}$.

Denote by $g_{in}^{pr}(\gamma)$ the prior belief of caller i about the distribution of the scale parameter before her n^{th} contact. Also, denote by $g_{in}^{po}(\gamma|I)$ the posterior belief about the distribution of the scale parameter if the caller has acquired information I about s_{in}. Given that $I_{in}(t) \subset I_{in}(t-1) \subset ... \subset I_{in}(0)$, we have $g_{in}^{po}(\gamma|I_{in}(t) \cap I_{in}(t-1) \cap ... \cap I_{in}(0)) = g_{in}^{po}(\gamma|I_{in}(t))$. In other words, considering updating while waiting in the queue, the only relevant information for making decision in period t is $I_{in}(t)$. Consequently, to find $\mu_{in}^{po}(t)$ and $\delta_{in}^{po}(t)$, we need to find $g_{in}^{po}(\gamma|I_{in}(t))$.

Suppose that the caller entered the service stage at $t = s_{in}$, i.e. $I_{in}(t) = \{s_{in} = t\}$. Hence, we

have $O_{int} = 0$. Then by Bayes' rule we can write

$$\begin{aligned} g_{in}^{po}(\gamma|I_{in}(t)) &= \frac{f(t;k_0,\gamma)g_{in}^{pr}(\gamma)}{\int f(t)g_{in}^{pr}(\gamma)d\gamma}, \\ &= \frac{(\frac{k_0}{\gamma}t^{k_0}e^{-\frac{t^{k_0}-1}{\gamma}})(\frac{\delta_{in}^{pr\mu_{in}^{pr}}}{\Gamma(\mu_{in}^{pr})}\frac{e^{-\frac{\delta_{in}^{pr}}{\gamma}}}{\gamma^{(\mu_{in}^{pr}+1)}})}{\int_0^\infty(\frac{k_0}{\gamma}t^{k_0}e^{-\frac{t^{k_0}-1}{\gamma}})(\frac{\delta_{in}^{pr\mu_{in}^{pr}}}{\Gamma(\mu_{in}^{pr})}\frac{e^{-\frac{\delta_{in}^{pr}}{\gamma}}}{\gamma^{(\mu_{in}^{pr}+1)}})\ d\gamma}, \\ &= \frac{\frac{(\gamma+t^{k_0})^{\mu_{in}^{pr}+1}}{\Gamma(\mu_{in}^{pr}+1)}\frac{e^{-\frac{\delta_{in}^{pr}+t^{k_0}}{\gamma}}}{\gamma^{\mu_{in}^{pr}+2}}}{\int\frac{(\gamma+t^{k_0})^{\mu_{in}^{pr}+1}}{\Gamma(\mu_{in}^{pr}+1)}\frac{e^{-\frac{\delta_{in}^{pr}+t^{k_0}}{\gamma}}}{\gamma^{\mu_{in}^{pr}+2}}\ d\gamma}. \end{aligned} \tag{27}$$

The numerator of (27) is the pdf of an inverse gamma distribution with parameters $\mu_{in}^{pr}+1$ and $\delta_{in}^{pr}+t^{k_0}$, and the denominator is its integral. Therefore, we have

$$g_{in}^{po}(\gamma|I_{in}(t)) = \frac{(\gamma+t^{k_0})^{\mu_{in}^{pr}+1}}{\Gamma(\mu_{in}^{pr}+1)}\frac{e^{-\frac{\delta_{in}^{pr}+t^{k_0}}{\gamma}}}{\gamma^{\mu_{in}^{pr}+2}}. \tag{28}$$

The right-hand side of (28) is the pdf of an inverse gamma distribution with parameters $\mu_{in}^{pr}+1$ and $\delta_{in}^{pr}+t^{k_0}$. By definition, $g_{in}^{po}(\gamma|I_{in}(t))$ is the pdf of caller i's posterior belief with parameters $\mu_{in}^{po}(t)$ and $\delta_{in}^{po}(t)$. Consequently, for the case in which the caller does not abandon and enters the service stage after t seconds, we have

$$\begin{aligned} \mu_{in}^{po}(t) &= \mu_{in}^{pr}+1 = \mu_{in}^{po}+1-O_{int}, \\ \delta_{in}^{po}(t) &= \delta_{in}^{pr}+t^{k_0}. \end{aligned}$$

For the case in which the caller does not receive service, or abandons, by time t ($O_{in}(t) = 1$, $I_{in}(t) = \{s_{in} > t\}$), using the Bayes' rule the pdf of caller i's posterior belief is given by

$$\begin{aligned} g_{in}^{po}(\gamma|I_{in}(t)) &= \frac{(1-F(t;k_0,\gamma))g_{in}^{pr}(\gamma)}{\int(1-F(t;k_0,\gamma))g_{in}^{pr}(\gamma)d\gamma}, \\ &= \frac{e^{-\frac{t^{k_0}}{\gamma}}(\frac{\delta_{in}^{pr\mu_{in}^{pr}}}{\Gamma(\mu_{in}^{pr})}\frac{e^{-\frac{\delta_{in}^{pr}}{\gamma}}}{\gamma^{(\mu_{in}^{pr}+1)}})}{\int_0^\infty e^{-\frac{t^{k_0}}{\gamma}}(\frac{\delta_{in}^{pr\mu_{in}^{pr}}}{\Gamma(\mu_{in}^{pr})}\frac{e^{-\frac{\delta_{in}^{pr}}{\gamma}}}{\gamma^{(\mu_{in}^{pr}+1)}})\ d\gamma}, \end{aligned}$$

$$= \frac{\frac{(\gamma+t^{k_0})^{\mu_{in}^{pr}}}{\Gamma(\mu_{in}^{pr})}\frac{e^{-\frac{\delta_{in}^{pr}+t^{k_0}}{\gamma}}}{\gamma^{\mu_{in}^{pr}+1}}}{\int \frac{(\gamma+t^{k_0})^{\mu_{in}^{pr}}}{\Gamma(\mu_{in}^{pr})}\frac{e^{-\frac{\delta_{in}^{pr}+t^{k_0}}{\gamma}}}{\gamma^{\mu_{in}^{pr}+1}}\ d\gamma},$$

$$= \frac{(\gamma+t^{k_0})^{\mu_{in}^{pr}}}{\Gamma(\mu_{in}^{pr})}\frac{e^{-\frac{\delta_{in}^{pr}+t^{k_0}}{\gamma}}}{\gamma^{\mu_{in}^{pr}+1}}. \tag{29}$$

The right-hand side of (29) is the pdf of an inverse gamma distribution with parameters μ_{in}^{pr} and $\delta_{in}^{pr}+t^{k_0}$. Therefore, for the case in which the caller does enter the service stage or abandons by time t, we have

$$\mu_{in}^{po}(t) = \mu_{in}^{pr} = \mu_{in}^{pr} + 1 - O_{int},$$
$$\delta_{in}^{po}(t) = \delta_{in}^{pr} + t^{k_0}.$$

Finally, under the assumption that callers do not forget what they learned in the past, callers' belief at the end of contact n would be the same as their belief at the beginning of contact $n+1$. Therefore, we have: $\mu_{in+1}^{pr,m} = \mu_{in}^{po,m}(w_{in})$ and $\delta_{in+1}^{pr,m} = \delta_{in}^{po,m}(w_{in})$.

Proof of Proposition 2. Consider caller i who updates her belief about the scale parameter of the waiting time distribution according to the updating process illustrated in Proposition 1. Recall that μ_{in+1}^{po} and δ_{in+1}^{po} denote the parameters of caller i's belief at the end of contact n (and beginning of contact $n+1$).We show that if $n \to \infty$ the mean of callers' belief distribution converges to γ_0 and its variance converges to 0.

Suppose that s_{in} is the amount if time the caller needs to wait in her n^{th} contact to receive service, which is a draw from the Weibull distribution of the waiting time with pdf and cdf given in (1). Moreover, suppose that the patience time of caller i denoted by a_{in} is a draw from a distribution with cdf and pdf given by $G(\cdot)$ and $f(\cdot)$, respectively. Caller i will receive service if her patience level is larger than the amount of time she need to wait, and will abandon otherwise. Consequently, her waiting time observed in the data is given by $w_{in} = \min(s_{in}, a_{in})$. Based on Proposition 1 we have $\mu_{in+1}^{po} = \mu^{pr} + \sum_{j=1}^{n}(1 - O_{ij})$ and $\delta_{in+1}^{po} = \delta^{pr} + \sum_{j=1}^{n} w_{ij}^{k_0}$.

Let $Mean_{in+1}$ and Var_{in+1} denote the mean and variance of callers' belief about γ_0 at the beginning of their $(n+1)^{th}$ contact. Following the inverse gamma assumption about the distribution of callers' belief, we have

$$Mean_{in+1} = \frac{\delta_{in+1}^{po}}{\mu_{in+1}^{po} - 1} = \frac{\delta^{pr} + \sum_{j=1}^{n} w_{ij}^{k_0}}{\mu^{pr} + \sum_{j=1}^{n}(1 - O_{ij}) - 1}, \tag{30}$$

and

$$Var_{in+1} = Mean_{in+1}^2 \frac{1}{\mu_{in+1}^{po} - 2} = Mean_{in+1}^2 \frac{1}{\mu^{pr} + \sum_{j=1}^{n}(1 - O_{ij}) - 2}. \tag{31}$$

We first show that $\lim_{n\to\infty} Mean_{in+1} = \gamma_0$ and then $\lim_{n\to\infty} Var_{in+1} = 0$.

Part 1. $\lim_{n\to\infty} Mean_{in+1} = \gamma_0$:

Based on Equation (30), we can write

$$\lim_{n\to\infty} Mean_{in+1} = \lim_{n\to\infty} \frac{\frac{\delta^{pr}}{n} + \frac{\sum_{j=1}^{n} w_{ij}^{k_0}}{n}}{\frac{\mu^{pr}-1}{n} + \frac{\sum_{j=1}^{n}(1-O_{ij})}{n}}. \tag{32}$$

Given that $\lim_{n\to\infty} \frac{\delta^{pr}}{n} = 0$, $\lim_{n\to\infty} \frac{\mu^{pr}-1}{n} = 0$, and the independence of the s_{in} and a_{in} (recall that $w_{in} = min(s_{in}, a_{in})$), we can use the strong law of large numbers to rewrite (32) as follows :

$$\begin{aligned}
\lim_{n\to\infty} Mean_{in+1} &= \frac{\mathbb{E}_g \mathbb{E}_f(w_{ij})}{\mathbb{E}_g \mathbb{E}_f(1 - O_{in})}, \\
&= \frac{\mathbb{E}_g(\int_0^{a_{ij}} s^{k_0} f(s)ds + \int_{a_{ij}}^{\infty} a_{ij}^{k_0} f(s)ds)}{\mathbb{E}_g(\int_0^{a_{ij}} f(s)ds)}, \\
&= \frac{\mathbb{E}_g([-(s^{k_0} + \gamma_0)\exp(\frac{-s^{k_0}}{\gamma_0})]_0^{a_{ij}} + a_{ij}^{k_0}\exp(\frac{-s^{k_0}}{\gamma_0}))}{\mathbb{E}_g(1 - \exp(\frac{-s^{k_0}}{\gamma_0}))}, \\
&= \frac{\mathbb{E}_g(-a_{ij}^{k_0}\exp(\frac{-s^{k_0}}{\gamma_0}) - \gamma_0\exp(\frac{-s^{k_0}}{\gamma_0}) + \gamma_0 + a_{ij}^{k_0}\exp(\frac{-s^{k_0}}{\gamma_0}))}{\mathbb{E}_g(1 - \exp(\frac{-s^{k_0}}{\gamma_0}))}, \\
&= \frac{\gamma_0 \mathbb{E}_g(1 - \exp(\frac{-s^{k_0}}{\gamma_0}))}{\mathbb{E}_g(1 - \exp(\frac{-s^{k_0}}{\gamma_0}))}.
\end{aligned} \tag{33}$$

Given our assumption that $G(0) < 1$, the patience time distribution will have some mass at non-zero values. Consequently, we have $\mathbb{E}_g(1 - \exp(\frac{-s^{k_0}}{\gamma_0})) \neq 0$, and we can write

$$\lim_{n\to\infty} Mean_{in+1} = \frac{\gamma_0 \mathbb{E}_g(1 - \exp(\frac{-s^{k_0}}{\gamma_0}))}{\mathbb{E}_g(1 - \exp(\frac{-s^{k_0}}{\gamma_0}))} = \gamma_0. \tag{34}$$

Part 2. $\lim_{n\to\infty} Var_{in+1} = 0$:

Given Equations (31) and (34) we have

$$\begin{aligned}\lim_{n\to\infty} Var_{in+1} &= \gamma_0^2 \lim_{n\to\infty} \frac{\frac{1}{n}}{\frac{\mu^{pr}-2}{n} + \frac{\sum_{j=1}^{n}(1-O_{ij})}{n}}, \\ &= \gamma_0^2 \lim_{n\to\infty} \frac{\frac{1}{n}}{\mathbb{E}_g\mathbb{E}_f(1-O_{in})}, \\ &= \gamma_0^2 \lim_{n\to\infty} \frac{\frac{1}{n}}{\mathbb{E}_g(\int_0^{a_{ij}} f(s)ds)}, \\ &= \gamma_0^2 \lim_{n\to\infty} \frac{\frac{1}{n}}{\mathbb{E}_g(1-\exp(\frac{-s^{k_0}}{\gamma_0}))}.\end{aligned} \tag{35}$$

Given that $\mathbb{E}_g(1-\exp(\frac{-s^{k_0}}{\gamma_0})) \neq 0$, from (35) we have $\lim_{n\to\infty} Var_{in+1} = 0$.

D Estimation Procedure for the Rational Expectation Model

We use the following procedure to estimate the parameters of the rational expectation model.

a) We first use the Kaplan-Meier estimator (Kaplan and Meier (1958)) to non-parametrically estimate callers' waiting time distribution denoted by $F^{RE,m}(t)$ and the hazard rate of this distribution denoted by $\pi^{RE,m}(t)$ for m=1, 2 corresponding to Rush-hour and Non-Rush-hours.

b) Using equations (9) to (12) we can calculate caller i's abandonment probability in her n^{th} contact under the rational expectation equilibrium assumption denoted by $P_{int}^{RE}(1, r_i, c_i)$ by replacing $\pi_{in}^{po,m}(t)$ by $\pi^{RE,m}(t)$.

c) The log-likelihood function for callers' actions denoted by $logL^{RE}$ is given by

$$\begin{aligned}\log L^{RE}(\Theta_0^{RE}) = \sum_{i=1}^{N} \log \Bigg(\sum_{l=1}^{2} \prod_{n=1}^{n_i} \mathbb{I}_{\{q_{in}=m\}} \prod_{t=0}^{w_{in}} \log \Big(P_{int}^{RE}(1, r^l, c^l)\Big)^{\mathbb{I}_{d_{in,t}=1}} \\ \times \Big(1 - P_{int}^{RE}(1, r^l, c^l)\Big)^{\mathbb{I}_{d_{in,t}=0}} \Bigg).\end{aligned} \tag{36}$$

We maximize (36) to estimate $\Theta_0^{RE} = \{(\eta^l, r^l, c^l)_{\{l=1,2\}}\}$. Please see Aksin et al. (2013) for more details.