Don't Call Us, We'll Call You: An Empirical Study of Callers' Behavior Under a Callback Option

October 4, 2017

Brett Alan Hathaway, Seyed Morteza Emadi, Vinayak Deshpande
Kenan-Flagler Business School, University of North Carolina at Chapel Hill

Abstract

While call centers have recently invested in callback technology, the impact of this innovation on callers' behavior and call center performance has been less clearly understood. Using call center data from a US commercial bank, we perform an empirical study of callers' decision-making process in the presence of a callback option. We formulate a structural model of callers' decision-making process, and impute their underlying preferences from the data. Our estimates of callers' preferences show that callers experience almost no discomfort while waiting for a callback. We also find that callers incur a high cost of switching from their offline tasks to answer a callback. We conduct a counterfactual analysis of how various callback policies affect the service quality and system throughput of this call center. Our results indicate that offering callbacks increases service quality, as measured by the average disutility callers accumulate while waiting for service, without substantially impacting throughput.

1 Introduction

Customer queueing is an inseparable part of service encounters, which may be detrimental to firms in two ways. First, customers' perceptions of service quality may drop as they generally find waiting in queue unpleasant (Bitran et al. (2008); Suck and Holling (1997); Leclerc et al. (1995)). Second, customers may lose patience and abandon the queue, which could reduce system throughput. A reduction in system throughput leads to underutilization of resources and may result in lost revenue and/or churn. Thus, it is not surprising that firms use a variety of remedies to mitigate the negative effects of waiting.

A key to quantifying the effectiveness of these remedies is to have a clear definition of service quality. Historically, queuing literature has defined service quality as a function of how long customers wait, e.g., their average waiting time (Aksin et al., 2007). However, this definition concentrates only on how long customers wait, while ignoring how unpleasant it is to wait per unit

of time. An alternative definition of service quality can be formed from the economic framework used in recent literature that treats waiting customers as utility-maximizing agents (Mandelbaum and Shimkin (2000); Hassin and Haviv (2003); Aksin et al. (2013)). In this framework, customers incur a per period cost of waiting and accumulate 'waiting disutility' until they receive service or abandon the system. We therefore define 'service quality' as the average waiting disutility that customers accumulate while waiting for service, which is determined by the average waiting time of customers in the system and their cost of waiting per unit of time. Thus, firms may increase service quality by reducing customer waiting times or by making the waiting experience more pleasant per unit of time (reducing waiting cost).

Call centers have been on the forefront of providing technological innovations to increase service quality and/or system throughput. One of these innovations is providing delay announcements where callers receive information regarding their future waiting time such as their position in line or a delay estimate. The idea behind providing delay announcements is that service quality will increase for the subset of callers who choose to abandon due to their announcement. Rather than accumulating disutility while waiting in queue, these callers exit the system, and accumulate little to no disutility until they reenter the system later. However, a major shortcoming of providing delay announcements is that the call center has no control over when the callers will reenter the system. Consequently, some callers may reenter the system when the call center has insufficient service capacity, which may lead to longer waiting times and, hence, lower service quality.

The recent innovation of callback technology seeks to address the shortcoming of providing delay announcements, while retaining its benefits. This innovation is gaining popularity as a recent survey of call center managers found that 22% of them have implemented callback technology (ContactBabel (2016)). Under callback technology, callers are offered a callback, where they may wait 'offline' until they are contacted by a call center agent, or remain in the 'online' queue to wait for service. Callback technology enables callers who accept a callback offer to perform offline tasks while waiting for their callback, rather than waiting for service while tethered to a phone. This should decrease the per period cost that callers incur while waiting for service. In addition, the call center determines the policy of when to initiate callbacks, which gives them control over when callers reenter the system. This allows the call center to initiate callbacks in periods where they have capacity to serve additional callers with little or no impact on the waiting times of callers in the online queue. Finally, callback technology could reduce lost calls by providing callbacks to callers who otherwise would have abandoned without returning to the system. In fact, callback technology could even increase system throughput in call centers that offer callbacks when they have insufficient service capacity and initiate callbacks when they have excess service capacity. This form of demand postponement has been shown to be advantageous in a variety of practical situations (Iyer et al. (2003); Biyalogorsky et al. (1999); Odlyzko (1999)).

The extant literature has been positive regarding the effects of offering callbacks. Using diffusion approximations, Armony and Maglaras (2004a,b) characterize the performance of a call center that offers callbacks, and find that the average waiting time of callers who wait in the online queue

decreases while system throughput increases. Legros et al. (2016) use an MDP framework to perform callback policy analyses, and show that offering callbacks may substantially decrease the combined cost of caller waiting and abandonment, while Ata and Peng (2017) show that offering callbacks reduces online waiting times in a system where the arrival rate is a time-varying stochastic process. While the literature has been favorable towards callbacks, it has been exclusively analytical in nature. This leaves two gaps which, to the best of our knowledge, have yet to be addressed empirically. The first gap is in understanding how callers choose whether to accept a callback offer. For example, what preferences do callers consider when making their decisions, and how do these preferences affect the likelihood that they will accept the callback offer when it is presented to them or the likelihood that they will actually answer the callback when it arrives? The previous literature recognizes this gap, and highlights the importance of understanding caller preferences regarding callbacks using real data. For example, Armony and Maglaras (2004a, pg. 276) write, "*An interesting problem that will not be broached here is the estimation of customer preferences using observed data.*" The second gap is in finding the impact of offering callbacks in a real world setting supported by data. This requires a model that could be used to find the impact of different policies on callers' behaviors, which in turn would impact system performance. For example, managers may be interested in understanding how service quality and system throughput are affected by policy parameters such as when to offer and when to initiate callbacks.

In this paper we fill the aforementioned gaps in the literature and take a data-driven approach to studying callbacks by providing a framework to understand callers' decision making process and preferences under a callback policy and by conducting an analysis of the impact of different callback policies on the service quality and system throughput of a call center. To perform our study we leverage a dataset of over 1 million calls from a US-based banking call center. When callers enter this call center, the system first estimates their expected waiting time. If the callers' expected waiting time is less than 5 minutes or greater than 59 minutes, callers are sent to the online queue without receiving a callback offer or delay announcement. If their expected waiting time is between 5 and 59 minutes, the system announces the callers' expected delay and offers callers a callback with a 'service guarantee', which is an assurance that they will receive a callback within their announced delay. Callers who accept a callback offer enter the 'offline queue' and are contacted by a call center agent within the guarantee duration, while callers who decline the offer may enter the 'online queue' and wait for service. In our data we are able to observe whether callers received a callback offer, whether they accepted the callback offer, and whether they answered the callback when it arrived. In addition, we are also able to observe the abandonment decisions of callers who entered the online queue.

To characterize the callers' decision-making process and to capture their preferences, we formulate and estimate the parameters of a structural model of callers' decisions under a callback policy. We build on the recent stream of structural models that treat callers who wait in queue as utility maximizing agents (Aksin et al. (2013); Akşin et al. (2016); Yu et al. (2016); Emadi and Swaminathan (2017)). These models characterize callers' decision-making process for how long to

wait before abandoning a queue, and impute callers' reward for receiving service and their per unit waiting cost from the data. We extend their framework to include the callers' decisions of whether to accept a callback offer and whether to answer a callback when it arrives. To model these decisions, we formulate three additional primitives to capture the callers' callback preferences. The first is the utility that callers obtain by receiving a guarantee that their callback will arrive within a given time frame. The second is the per unit cost that callers incur to wait offline for their callback to arrive, and the third is the cost that callers incur to switch away from their offline tasks to answer a callback when it arrives.

Our estimation results reveal a number of interesting findings. While the callers' per unit online waiting cost is statistically significant and a positive value, the callers' per unit offline waiting cost is statistically equal to zero. One of the implications of this result is that the average online waiting time of all callers can be used as a metric for service quality, because callers accumulate disutility while waiting online but accumulate no disutility while waiting offline. Hence, managers who are concerned with maximizing service quality should be focused on reducing this metric by substituting offline waiting time for online waiting time through offering callbacks. Moreover, we find that the callers' valuation of their service guarantee for accepting a callback offer does not depend on the duration of the guarantee. This indicates that callers value the assurance that they will receive a callback within a specified timeframe, but aren't too concerned about when the callback is guaranteed to arrive. In other words, callers value a guarantee of 5 minutes as much as a guarantee of 30 minutes. We also observe that the process of switching from offline tasks in order to answer a callback is costly for callers. We base this observation on the fact that the callers' average estimated switching cost is 2.99, while their estimated service reward is 5.57. However, since their reward for receiving service is sufficiently high, most callers still answer a callback when it arrives. Finally, our estimates indicate that if this call center offered callbacks with guarantee durations longer than the current policy, callers would have nearly the same likelihood of accepting callback offers as they do under the current policy. This result indicates that, without sacrificing callback offer acceptance rates, this call center can extend their guarantee durations to implement a demand postponement policy where they offer callbacks when they have insufficient service capacity and initiate callbacks when they have excess service capacity.

Given that we impute the callers' preferences through our estimation process, we are able to predict how callers would behave under callback policies that differ from the call center's current policy, and thus ascertain the effects of those policies on service quality and system throughput. We do this by performing a counterfactual analysis of four callback policies under different loads. We find that offering callbacks increases service quality regardless of system load. This occurs because the average time that callers in the system wait in the online queue reduces, which reduces the callers' average waiting disutility. However, we find that the magnitude of the service quality improvement depends on the callback policy. For example, we show that offering callbacks with a fixed guarantee duration of 30 minutes increases service quality more than the current policy. Moreover, we find that offering callers a window, rather than a specific time, in which the callback

is guaranteed to arrive may further increase service quality.

While offering callbacks increases service quality, it does not have a substantial effect on system throughput. Under high loads system throughput is unaffected since the system is operating near maximum capacity. Under low loads offering callbacks may slightly increase or decrease system throughput depending on the policy, but the magnitude of the changes are small.

Our paper has three main contributions. First, we provide a model for callers' decisions in a call center that offers callbacks. To the best of our knowledge this is the first empirical work in this area. Through our model we provide a framework for how callers choose whether to accept a callback offer and whether to answer a callback when it arrives. Our model extends the current structural models of callers' behaviors and contains additional primitives that capture callers' callback preferences. Second, we find interesting insights about callers' preferences that expand our understanding of callers' decision making process. For example, our negligible estimate of the callers' offline waiting cost indicates that waiting for a callback does not reduce service quality, and that the average online waiting time of all callers in the system can be used a metric for service quality. Also, the high estimate of the cost that callers incur to switch tasks in order to answer a callback demonstrates why certain callers choose not to answer an arriving callback. Third, our framework can be used to conduct counterfactual analyses of how different callback policies will perform under a variety of operating conditions. We show in our setting that offering callbacks improves service quality but has no substantial impact on system throughput. We also show that this call center could substantially improve its service quality by implementing policies that offer callbacks during periods of insufficient service capacity and initiate callbacks during periods of excess capacity.

2 Literature Review

Our work relates to two streams of literature. The first stream consists of models that treat callers as utility maximizing agents. The second stream is research regarding callbacks. In this section we review both of these streams and articulate our contribution to each.

Models of Callers as Utility Maximizing Agents: There are several extant models in the literature that treat callers as utility maximizing agents. The seminal paper in this stream is by Naor (1969). He models an M/M/1 system in which a manager selects an admission price that customers must pay to join the queue. Upon observing the queue length, customers decide whether to balk or join the queue. Naor derives the equilibrium behaviors of the manager and the customers. Naor's model began a long stream of research to which Hassin and Haviv (2003) provide an overview. Because callers in our setting may abandon after joining the online queue, the models in this stream that are most pertinent to ours are those that deal with callers' abandonment decisions (Hassin and Haviv (1995); Mandelbaum and Shimkin (2000); Shimkin and Mandelbaum (2004)).

The more recent papers in this stream have modeled callers' abandonment decisions using a structural estimation approach, which leverages data to impute the underlying preferences that

drive callers' abandonment behaviors. The first work in this vein is that of Aksin et al. (2013), where the authors model callers who wait in queue as solving an optimal stopping problem. Given their service reward and their per unit waiting cost, callers decide when to stop waiting by abandoning the system. Aksin, et al. estimate the callers' reward and cost parameters for four different classes and find that their callers' parameters differ by class. Emadi and Swaminathan (2017) relax the rational expectation assumption and instead model callers in a Bayesian learning framework, where callers update their beliefs about the distribution of waiting times in the call center each time they call. They find that callers who have no experience with the call center are optimistic about their probability of receiving service in a given period. Akşin et al. (2016) and Yu et al. (2016) extend the approach of Aksin et al. (2013) to call centers that provide delay announcements. In both works callers use the information in their delay announcement as a signal of the current waiting times in the call center and change their abandonment behavior accordingly. Our contribution to this stream of research is the formulation of a structural model of callers' behavior in the presence of a callback option.

Callback Research: Only a handful of studies regarding call centers that offer callbacks have been conducted. The two studies that are most pertinent to our work are those of Armony and Maglaras (2004a,b). In Armony and Maglaras (2004a) the authors model a call center as an $M/M/N$ multiclass system with an online queue for real-time service and an offline queue for callers who postpone service by receiving a callback. Upon entering the system, callers are told the steady-state mean waiting time in the online queue and receive a callback offer with a fixed guarantee of the maximum delay they will experience before receiving a callback. Callers then choose either to immediately balk, join the online queue, or accept the callback offer. Relying on diffusion approximations, Armony and Maglaras find the unique equilibrium of the system and characterize the system performance. They show under a wide range of caller preferences that offering callbacks decreases the average waiting time of callers who join the online queue, while increasing the throughput of the system. Armony and Maglaras (2004b) extend their analysis to a system where callers receive an estimate of their waiting time in the online queue upon arrival, which is based on the current state of the system. The authors show that providing real-time delay estimates magnifies the performance improvements from their first analysis.

A pair of more recent papers have concentrated on the decision of when to offer callbacks. Legros et al. (2016) explore this question using an MDP approach, where the manager's objective is to minimize the sum of the expected online waiting costs, offline waiting costs, and abandonment costs. They show in the case of two servers that callbacks should only be offered when the number of callers waiting in the offline queue is below some threshold. They also numerically characterize a policy for how many servers should be reserved to answer callers in the online queue. Ata and Peng (2017) study the question of when to offer callbacks in a system where the arrival rate is a time-varying stochastic process. They propose a threshold policy which depends on the length of the online queue, the current arrival rate, and the maximum service rate of the system, and demonstrate that their policy performs well in simulations.

Our contribution to the stream of callback research is two-fold. First, to the best of our knowledge we perform the only empirical study of callers' behavior in a call center that offers callbacks. We use the callers' decisions of whether to accept a callback offer and whether to answer a callback when it arrives in imputing the callers' callback preferences. Second, we provide a framework for conducting counterfactual analyses of callback policies and explore how various callback policies would affect the service quality and system throughput of this call center.

3 Data Description

We obtained our data from the call center of a US-based commercial bank. The call center offers customer service 24 hours a day, seven days a week. These services include deposit account inquiries, credit and debit card services, and online banking technical support. Our data includes a record of each of the 1,215,776 calls that were routed to queues staffed by agents between April 13, 2016 and July 31, 2016. The call center handles approximately 13,000 calls on the weekdays, 7,500 calls on Saturdays, and 4,500 calls on Sundays.

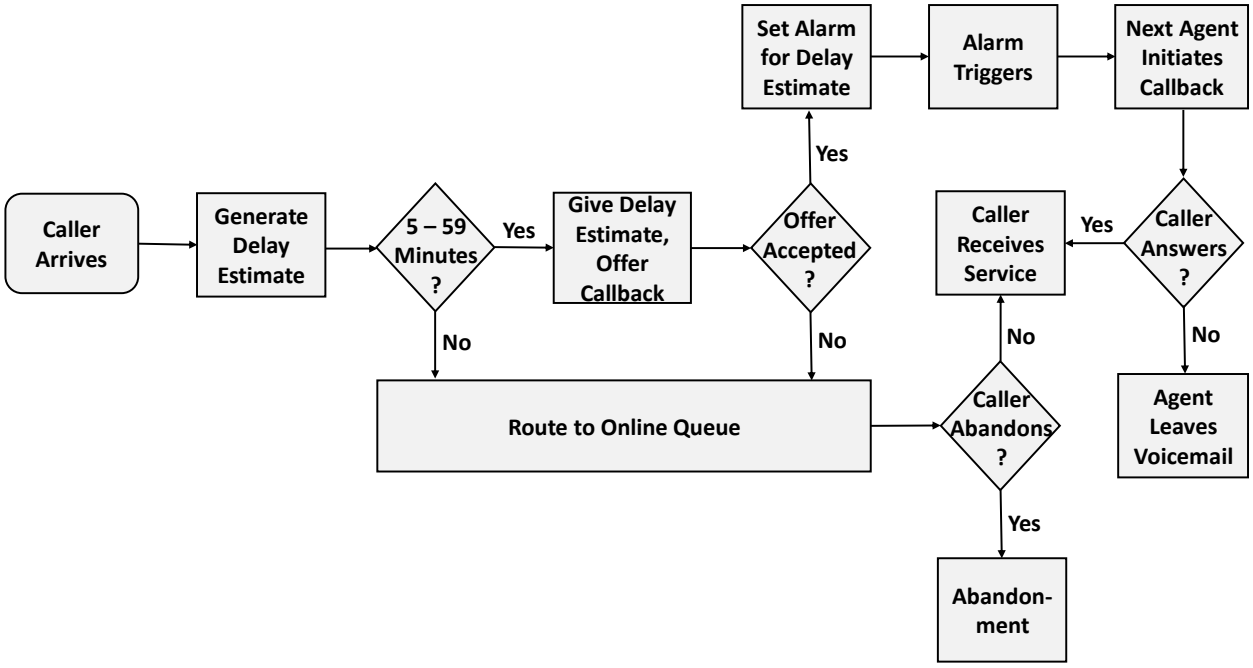
This call center offers callbacks to callers depending on the congestion in the system. In Figure 1 we provide a diagram of the routing process under the callback policy. Upon each caller's arrival, the call center generates a queue-length-based delay estimate (rounded down to the nearest minute) based on current queue lengths and staffing levels.[1] If the delay estimate is between 5 and 59 minutes, callers are informed of their delay estimate and receive a callback offer with a service guarantee of when the caller will receive the callback.[2] If a caller accepts a callback offer, the routing software schedules an alarm for one minute before the caller's delay estimate has elapsed. When the alarm triggers, the next available agent initiates the callback. If a caller's delay estimate is less than 5 minutes or greater than 59 minutes, the caller receives no delay announcement and no callback offer. For callers who receive a callback offer the delay announcement and callback offer are provided only once at the beginning of the call. Therefore, callers must immediately decide whether to accept the callback offer or wait for service in the online queue. Callers who choose to receive a callback may choose not to answer their phone when the callback arrives. In this case, the agent who initiated the callback leaves the caller a voicemail.

For every call in our dataset we observe the caller's estimated delay, whether the caller received a callback offer, whether the caller accepted a callback offer, whether the caller answered an arriving callback, and how long the caller waited until the callback arrived. For callers who joined the online queue we observe each caller's waiting time and whether the caller abandoned the queue before receiving service. To decrease the computational burden of estimating our model, we only analyze calls where the waiting time was 75 minutes or less; this restriction eliminates only .03% of the

[1] The delay is estimated using the following formula: mean service time of a call · calls waiting for service (both in the online and offline queues) / number of available servers. In literature this is called the *simple queue-length-based delay estimator*; see Ibrahim and Whitt (2009) for details.

[2] Callers who receive a callback offer receive the following message: "Your estimated wait time is x minutes. You may continue to hold for the next available agent, or we will save your place in line and call you back. To continue to hold, press 1. To save your place and receive a call back, press 2."

Figure 1: Diagram of the Callback Offer Policy



calls from the analysis.

In Figure 2 we provide summary statistics of the callback process. We observe that 50.1% of all callers received a callback offer, 37.5% of callers who received an offer accepted the offer, and 93% of callers who accepted a callback offer answered it when it arrived. Finally, the average waiting time for callers who received and answered a callback was 15 minutes, 59 seconds.

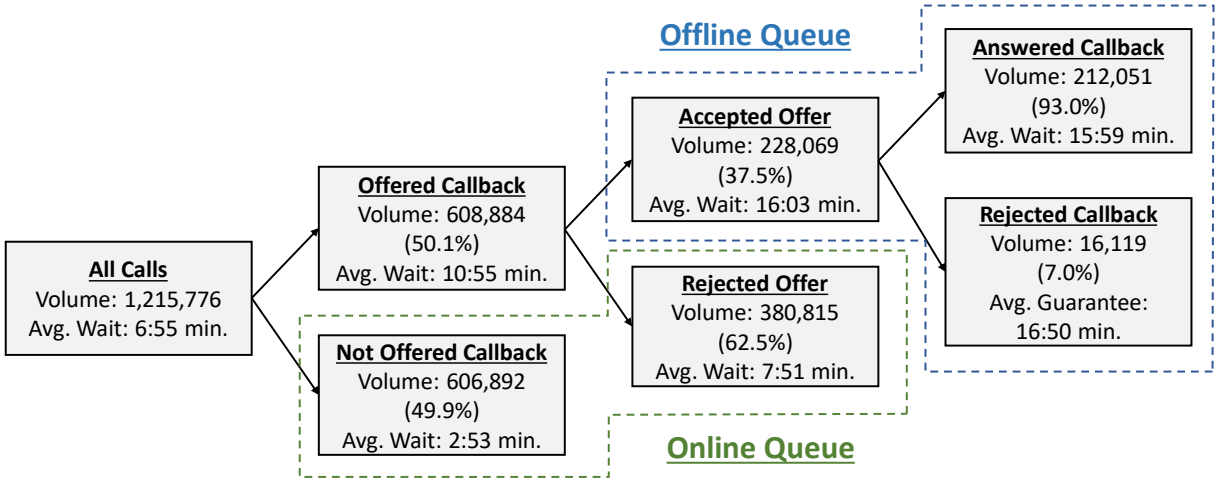
In Table 1 we display the summary statistics for callers who were routed to the online queue; this includes callers who did not receive a callback offer and callers who received an offer but declined it. The average waiting time for callers who waited in the online queue was 4 minutes and 48 seconds, and 22.1% of callers who entered the online queue abandoned.

Table 1: Summary Statistics of Calls Routed to Online Queue

Call Outcome	**Volume**	**% Total**	**Avg. Wait**
Routed to Online Queue	987,717	—	4:48
Answered from online queue	769,289	77.9%	4:41
Abandoned from online queue	218,428	22.1%	5:14

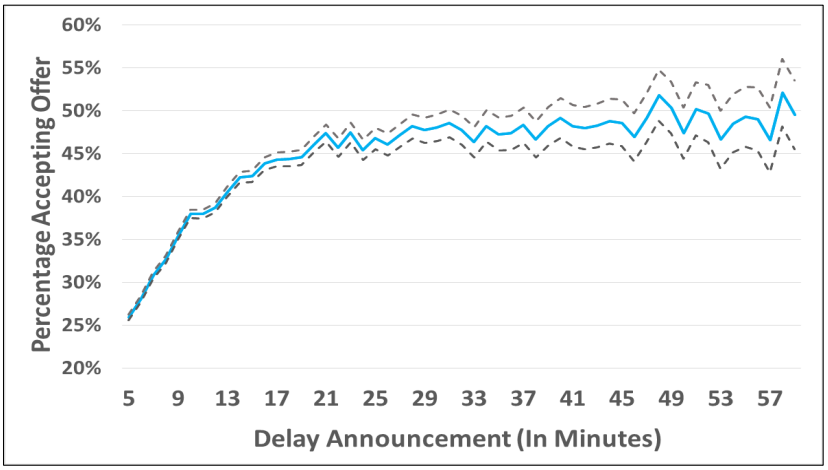
We next examine how delay announcements affect callers' callback decisions. In Figure 3 we graph the percentage of callers who accepted a callback offer by their delay announcement, and include the bounds of the 95% confidence interval. While only 26% of callers accept a callback offer with a 5 minute delay estimate, nearly half of callers accept an offer if the estimate is greater than 30 minutes. In Figure 4 we display the percentage of callers who answered an arriving callback

Figure 2: Callback Volume and Mean Waiting Times



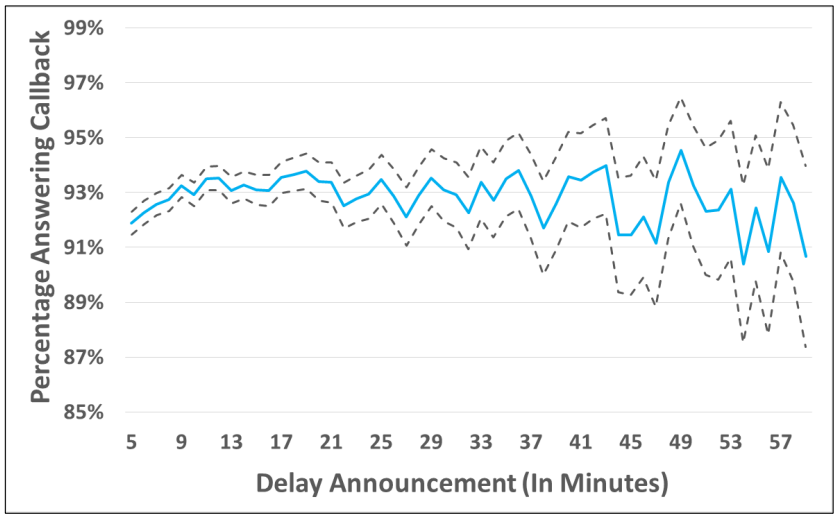
by delay announcement. Other than slightly lower answer rates for announcements of less than 7 minutes, we see no discernible relationship between delay announcements and the callers' propensity to answer arriving callbacks. This shows that even though delay announcements appear to have a large impact on callers' acceptance of a callback offer, their impact on callers' propensity to answer an arriving callback is not significant.

Figure 3: Percentage of Callers Accepting Callback Offer by Delay Announcement



Discussion: The high callback acceptance and answer rate suggests that receiving a callback is an attractive option for callers. While these descriptive statistics may be informative for managers who are considering implementing a callback policy, the data alone leaves two sets of questions unanswered. The first set of questions pertain to the callers' underlying callback decision-making process. For example, what caller preferences affect their willingness to accept a callback offer? One reason callers may accept a callback offer is if their discomfort of waiting online exceeds their

Figure 4: Percentage of Callers Answering Arriving Callback by Delay Announcement



discomfort of waiting offline. Another reason is that callers may value the certainty of receiving a callback within a guaranteed timeframe over the stochasticity of waiting in the online queue. We are interested in understanding how callers consider these two benefits when they are deciding whether to accept a callback offer.

The second set of questions pertain to the performance of this call center under callback policies that differ from its current policy. For example, how would service quality and system throughput change if the call center offered callbacks when a caller's delay estimate is less than the current threshold of five minutes? What if the callers' service guarantee were fixed, rather than varying with the callers' delay estimate? What if the call center offered callers a window in which their callback will arrive rather than a fixed guarantee? Finally, how would these callback policies perform relative to a policy where no callbacks are offered? Understanding how different callback policies affect call center performance should be critical for managers who are seeking to increase service quality and/or system throughput.

Tackling the above questions requires a model that characterizes the callers' decision-making process. One approach would be to use reduced-form regressions to capture how callers' callback and abandonment decisions vary under different delay announcements and waiting times. However, callers make these decisions based on two things - their underlying preferences such as their service valuation and discomfort of waiting, and their beliefs about the distribution of waiting times in the call center. While the callers' preferences are policy-independent, the callers' beliefs regarding the waiting time distribution are a reflection of the actual waiting times induced by the call center's policies. Since the reduced-form approach does not separate the impact of callers' preferences and beliefs on their decisions, any insights and predictions from the regressions would be limited to the current policy of the call center. We, therefore, require a model that recovers callers' underlying preferences regardless of the call center's policies. To do this, we formulate and estimate the

parameters of a structural model of callers' callback behaviors, which we explain next.

4 Model and Estimation Strategy

4.1 Model

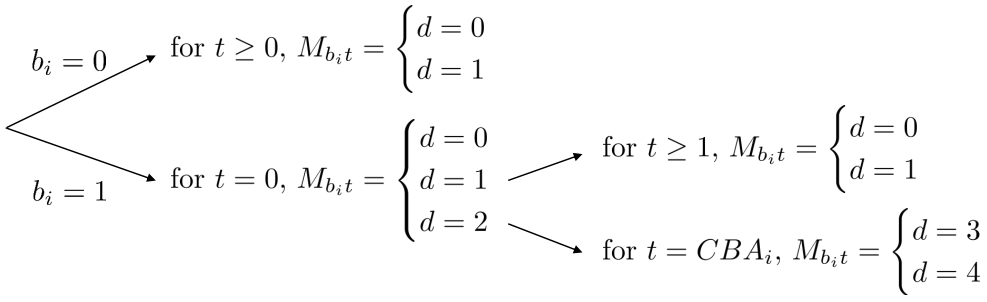
We model callers' behavior under a callback policy as a series of actions, denoted by d, made at discrete periods indexed by t, where t is the number of periods since arrival. The following are the five actions callers may take:

- **Abandon** ($d = 0$): Callers may abandon the system.
- **Wait Online** ($d = 1$): Callers may decide to wait in the online queue.
- **Accept Callback Offer** ($d = 2$): Callers may accept a callback offer and wait offline for the callback to arrive.
- **Reject Arriving Callback** ($d = 3$): When a callback arrives, callers may reject it by not answering their phone.
- **Answer Arriving Callback** ($d = 4$): When a callback arrives, callers may answer the callback and immediately receive service.

We index callers by $i \in I$. Let b_i be an indicator that caller i received a callback offer upon entering the system. In this call center, if a caller's estimated delay is between 5 and 59 minutes, $b_i = 1$ and the caller receives a delay announcement and a callback offer. Otherwise, $b_i = 0$ and the caller receives no delay announcement and no callback offer. Let $M_{b_i t}$ be the set of available actions (the caller's choice set) that caller i may take in period t, depending on whether caller i received a callback offer. In Figure 5 we provide an illustration of how $M_{b_i t}$ evolves as caller i moves through the system. If upon entering the system the caller does not immediately receive service, the caller may receive a callback offer, depending on the call center's policy. If caller i does not receive a callback offer ($b_i = 0$), she is routed to the online queue, where she chooses between abandoning and continuing to wait for $t \geq 0$. Caller i will continue through this waiting process until she either receives service or abandons. If caller i receives a callback offer ($b_i = 1$), she immediately chooses among three actions in period 0. She may abandon the system ($d = 0$), join the online queue ($d = 1$), or accept the callback offer ($d = 2$). If she joins the online queue and does not receive service in period 0, she chooses between abandoning and continuing to wait for $t \geq 1$, where she continues waiting until she receives service or abandons. If she accepts the callback offer, she waits offline until $t = CBA_i$, where CBA_i is the period in which the callback arrives (CBA stands for Call Back Arrival). Finally, when a callback arrives, callers may reject ($d = 3$) or answer the callback ($d = 4$).

Upon entering the system, caller i may receive a delay message and a guarantee message. She first receives a delay message, denoted by A_i. In general, the delay message contains information

Figure 5: Callers' Choice Set at Different Periods, Depending on Whether a Callback is Offered



about the current state of the online queue, such as a delay estimate or caller i's position in the online queue. In our setting the caller receives her expected waiting time in the online queue. Caller i then receives a guarantee message, denoted by G_i. This message includes information regarding her service guarantee, including a commitment to initiate her callback within D_{G_i} periods, should she choose to accept a callback offer. Note that under this call center's policy no message is announced if caller i does not receive a callback offer, i.e., if $b_i = 0$, then $A_i = \emptyset$, $G_i = \emptyset$. In the remainder of this paper, we use the superscript n for the online queue and f for the offline queue to index the parameters of the callers.

Callers act as forward-looking, utility-maximizing agents whose decisions depend on a vector of structural parameters, denoted by Θ. The utility caller i receives by choosing action d in period t is given by

$$u(t, A_i, G_i, d, \epsilon_{it}(d); \Theta) = v(t, A_i, G_i, d; \Theta) + \epsilon_{it}(d),$$

where $\epsilon_{it}(d)$ is the idiosyncratic shock of choosing action d in period t. The idiosyncratic shocks may be attributed to the effects of random external events on the utilities of callers' actions. We assume that the idiosyncratic shocks are type-1 extreme value distributed and are identically and independently distributed (iid) across different callers, periods, and actions. The term $v(t, A_i, G_i, d; \Theta)$ is the nominal utility of choosing action d in period t, which is independent of the idiosyncratic shocks.

We first explain the utilities for $d = 0$ and $d = 1$, which correspond to the actions that callers may take while waiting in the online queue. Then we explain the utility for $d = 2$, which is the utility of deciding to accept a callback offer upon arriving to the system. Finally, we explain the utilities for $d = 3$ and $d = 4$, which correspond to the actions that callers who have accepted a callback offer may take when their callback arrives.

We model the callers' decision-making process while waiting in the online queue as an optimal stopping problem as described in Aksin et al. (2013). At the beginning of each period, callers compare the utilities of abandoning and continuing to wait. If the utility of abandoning is greater than the utility of waiting, callers "stop" by immediately abandoning the queue. Callers who choose to wait at the beginning of a period and do not receive service by the end of the period

repeat the decision-making process at the beginning of the next period. This process continues until callers receive service or abandon. We assume that callers who join the online queue know the probability of receiving service in period t. This is consistent with the rational expectation equilibrium assumption in the extant literature (Aksin et al. (2013); Akşin et al. (2016); Yu et al. (2016)). We also assume that callers know they will receive service by no later than a terminal period, denoted by T.

We treat abandonment as an outside option and normalize the nominal utility to zero. Hence, if caller i chooses to abandon in the online queue the nominal utility is given by

$$v(t, A_i, G_i, 0; \Theta) = 0. \tag{1}$$

If caller i chooses to wait in the online queue, the nominal utility is given by

$$v(t, A_i, G_i, 1; \Theta) = -c^n + \pi_{A_i}(t)r + (1 - \pi_{A_i}(t)) \cdot V^n(t, r, c^n, \pi_{A_i}(\cdot)) \tag{2}$$

where c^n is the callers' per period cost of waiting in the online queue, r is the callers' reward for service, $\pi_{A_i}(t)$ is the caller's probability of receiving service in period t given the delay message A_i, and $V^n(t, r, c^n, \pi_{A_i}(\cdot))$ is the caller's online integrated value function. Callers consider the waiting cost for the current period, the utility they expect to receive from service in the current period, and the utility they expect to receive if they do not receive service in the current period but make the optimal decision in the next period.

The service probability $\pi_{A_i}(t)$ of caller i in period t is defined as the probability of caller i receiving service in period t, conditional on the caller not yet receiving service by period t. Like Akşin et al. (2016) we consider $\pi_A(\cdot)$ separately for different delay messages. We assume that $\pi_A(\cdot)$ is the equilibrium outcome of the system, where callers' beliefs regarding the distribution of online waiting times match the actual waiting time distribution. We also assume that $\pi_A(\cdot)$ is common knowledge among all callers, and that $\pi_A(T) = 1$, i.e., all callers will receive service by period T. Finally, given the nominal utilities of the actions callers may take while waiting in the online queue in Equations (1) and (2), from Ben-Akiva and Lerman (1985), §5.2, the online integrated value function is given by[3]

$$V^n(t, r, c^n, \pi_{A_i}(t)) = \tag{3}$$
$$\begin{cases} \log(1 + \exp(-c^n + \pi_{A_i}(t+1)r + (1 - \pi_{A_i}(t+1)) \cdot V^n(t+1, r, c^n, \pi_{A_i}(\cdot)), & \text{for } t < T, \\ 0, & \text{for } T. \end{cases}$$

Next we focus on the utilities at $t = 0$, given that the caller receives a callback offer ($b_i = 1$). We have already explained the utilites of $d = 0$ and $d = 1$. So, we next explain the utility of $d = 2$, which corresponds to accepting a callback offer. If caller i chooses to accept the callback offer, the

[3]The closed-form representation of the online value function relies on the fact that under the type-1 extreme value distribution the expected utility of choosing among a set of actions is given by the log of the sum of the exponentiated nominal utilities. In this case, the set of available actions in the next period are abandoning and waiting in the online queue. For a proof of this result, see Aksin et al. (2013), Appendix A.

nominal utility is given by

$$v(t=0, A_i, G_i, 2; \Theta) = -c^f D_{G_i} + \mu^f_{G_i} + V^f(r, s^f_{G_i}), \tag{4}$$

where c^f is the per period cost of waiting offline, $\mu^f_{G_i}$ is the utility of receiving a guarantee message of G_i, and $V^f(r, s^f_{G_i})$ is the caller's offline integrated value function, which depends on r (reward) and the switching cost for different guarantee messages $s^f_{G_i}$, which we later explain. The first term on the right in (4) is the callers' expected cost of waiting offline. Like Armony and Maglaras (2004a,b), we assume that callers who are offered a callback believe they will wait the entire duration of the service guarantee (D_{G_i}) before the callback arrives. In other words, they assume the worst-case scenario, which is that the callback will arrive in D_{G_i} periods. We refer to $\mu^f_{G_i}$ as the guarantee utility, which is the value callers derive from receiving a guarantee message of G_i. To account for any differences in callers' valuation of various guarantee messages, we consider $\mu^f_{G_i}$ separately for all values of G_i; we denote by U^f the vector of $\mu^f_G, \forall G$. The offline integrated value function is the caller's expected utility of making the optimal decision when the callback arrives. We provide a closed-form representation of $V^f(r, s^f_{G_i})$ after reviewing the nominal utilities of the actions callers may take when a callback arrives.

To characterize $V^f(r, s^f_{G_i})$ we need to lay out the utilities callers may receive when the callback arrives ($t = CBA_i$). If caller i chooses to reject a callback when it arrives ($d = 3$), the nominal utility is given by

$$v(t = CBA_i, A_i, G_i, 3; \Theta) = 0. \tag{5}$$

Because caller i receives no service by rejecting an arriving callback, this action is treated as an outside option and the nominal utility is normalized to zero.

If caller i chooses to answer a callback when it arrives, the nominal utility is given by

$$v(t = CBA_i, A_i, G_i, 4; \Theta) = r - s^f_{G_i}, \tag{6}$$

where $s^f_{G_i}$ is the cost caller i incurs to switch from her offline task to answer the callback. Callers who accept a callback offer may perform offline tasks while waiting for their callback to arrive, and would incur a cost to switch from their task to answer a callback. Therefore, callers' nominal utility for answering a callback is their service reward less their switching cost. To account for any differences in callers' switching costs based on their guarantee message, we consider s^f_G separately for all values of G, and denote by S^f the vector of $s^f_G, \forall G$.

We now provide the closed-form representation of the offline value function $V^f(r, s^f_{G_i})$ from Equation (4). Recall that when callers determine the nominal utility of accepting a callback, they consider the expected utility of making the optimal decision when the callback arrives. Given the nominal utilities of the actions callers may take when the callback arrives in Equations (5) and (6),

from Ben-Akiva and Lerman (1985), §5.2, we obtain [4]

$$V^f(r, s_G^f) = \log(1 + \exp(r - s_G^f)). \tag{7}$$

Given A_i, G_i, b_i, and $\Theta = \{r, c^n, c^f, U^f, S^f\}$, the optimal decision of caller i in period t is given by

$$d_{it} = \arg\max_{d \in M_{b_i t}} u(t, A_i, G_i, d, \epsilon_{it}(d); \Theta).$$

Finally, under the assumption that the callers' idiosyncratic shocks are iid type 1 extreme value distributed, callers' choice probabilities are given by a closed-form representation in a Logit form. Denoting by $P_{it}(d_{it}; A_i, G_i, b_i, \Theta)$ the probability that caller i chooses action d_{it} in period t, we have

$$P_{it}(d_{it}; A_i, G_i, b_i, \Theta) = \frac{\exp v(t, A_i, G_i, d_{it}; \Theta)}{\sum_{d \in M_{b_i t}} \exp v(t, A_i, G_i, d; \Theta)}. \tag{8}$$

4.2 Estimation Strategy

Prior to explaining our estimation strategy, we first discuss how we partition callers' delay messages (A) and guarantee messages (G) into distinct subsets to reduce the complexity of the estimation procedure. Recall that upon entering the system, callers may receive a delay message containing their estimated delay rounded down to the nearest minute, and a callback offer with a corresponding guarantee message, which is the call center's commitment to call the caller back within a guaranteed duration. Also recall that under this call center's policy the callers' guarantee duration is set equal to their delay estimate, i.e., $A_i = G_i$. Finally, recall that callers who receive no callback offer also receive no message, i.e., $A_i, G_i = \emptyset$ when $b_i = 0$; we refer to this case as the 'no-announcement message'. Let $\mathscr{L} = \{\emptyset, 5, 6, \cdots, 59\}$ denote the set of delay and guarantee messages, which includes the no-announcement message and one message for each minute of delay or guarantee duration between 5 and 59 minutes. We group the messages into 13 subsets of $\mathscr{L}$ with a partition $\mathscr{A}_0, \mathscr{A}_1, \cdots, \mathscr{A}_{12}$, where $\mathscr{A}_0 = \{\emptyset\}, \mathscr{A}_1 = \{5\}, \mathscr{A}_2 = \{6\}, \mathscr{A}_3 = \{7\}, \mathscr{A}_4 = \{8\}, \mathscr{A}_5 = \{9\}, \mathscr{A}_6 = \{10, 11\}, \mathscr{A}_7 = \{12, 13\}, \mathscr{A}_8 = \{14, 15\}, \mathscr{A}_9 = \{16, 17, 18\}, \mathscr{A}_{10} = \{19, 20, 21, 22\}, \mathscr{A}_{11} = \{23, 24, \cdots, 29\}$, and $\mathscr{A}_{12} = \{30, 31, \cdots, 59\}$. We refer to this partition as the message partition. We assume that callers who receive delay messages (or guarantee messages) that are in the same subset have the same service probabilities and same parameters. In other words, we assume that the impact of messages in the same subset on callers' behavior and expectations are the same. We choose this partition to make the volume of callers who receive messages from each subset to be as close to equal as possible.

We estimate our model in two steps. First, we estimate the service probabilities $\pi_A(\cdot)$ separately

[4] The closed-form representation of the offline value function again relies on the fact that under the type-1 extreme value distribution the expected utility of choosing among a set of actions is given by the log of the sum of the exponentiated nominal utilities. In this case, the set of actions are rejecting a callback and answering a callback and the exponentiated nominal utilities are 1 and $\exp(r - s_G^f)$.

for each subset from the message partition directly from the data. Then, using our estimate of $\pi_A(\cdot)$ and the callers' observed decisions, we estimate the callers' parameters via the maximum likelihood approach.

Recall that the callers' service probability for period t is the probability of a caller receiving service in period t, conditional on the caller not yet receiving service by period t. It is essentially the hazard rate of the callers' waiting time distribution. More formally, the service probabilities are given by

$$\pi_A(t) = \frac{F_A(t+1) - F_A(t)}{1 - F_A(t)},$$

where $F_A(\cdot)$ is the cumulative distribution of a caller's waiting time given the caller's delay message. We estimate $F_A(\cdot)$ separately for each subset of the message partition using the Kaplan-Meier estimator (Kaplan and Meier (1958)), which accounts for censoring due to abandonment. We assume that our estimate of $\pi_A(\cdot)$ is the equilibrium outcome of the system.

Using our estimate of $\pi_A(\cdot)$ and the observed decisions of the callers, we estimate the callers' parameters $\Theta = \{r, c^n, c^f, U^f, S^f\}$ via maximum likelihood. We index callers by $i \in I$, where I is the total number of callers in the data, and we let τ_i be the final decision period of caller i. Recall that the probability of caller i choosing action d_{it} in period t is denoted by $P_{it}(d_{it}; A_i, G_i, b_i, \Theta)$. Then the likelihood of observing the entire sample, which we denote by $L(\Theta)$, is given by

$$L(\Theta) = \prod_{i=1}^{I} \prod_{t=0}^{\tau_i} P_{it}(d_{it}; A_i, G_i, b_i, \Theta).$$

The maximum likelihood estimation is solved by maximizing the log-likelihood function, $\log L(\Theta)$, over the parameters Θ subject to the following constraints: the nominal utilities in Equations (1), (2), (4), (5), and (6) and the integrated value functions in Equations (3) and (7).

We assume that callers who wait in the online queue make decisions every ten seconds. Since our data is collected every second, we round caller waiting times upward and abandonment times downward. We assume that callers know they will be answered within 450 periods, i.e., $T = 450$. Thus, we run our estimation only on calls with waiting times that are no greater than 4500 seconds; this eliminates only .03% of the data. We maximize the likelihood function at 100 random starting points, and select the solution with the highest likelihood. To obtain standard errors, we perform nonparametric bootstrapping (Horowitz (2001)).

Identification: In Appendix A we formally discuss what sources of variation in the data allow us to identify the callers' parameters. We provide a brief summary here. The callers' reward and online waiting cost parameters are identified by variation in the service probabilities $\pi_A(\cdot)$ and the callers' abandonment decisions over different periods. Once the reward is identified, the vector of switching costs (S^f) is identified by variation in the callers' probability of rejecting a callback over the different guarantee messages. Finally, once the switching costs are identified, the offline waiting cost (c^f) and the vector of guarantee utilities (U^f) are identified by variation in the callers' probability of accepting a callback offer over different guarantee messages and guarantee durations.

5 Estimation Results and Model Validation

5.1 Estimation Results

In this section we present the estimation results and highlight some important insights drawn from our results.

Insight 1: The callers' offline waiting cost (c^f) is nearly zero, implying that waiting offline for a callback does not reduce service quality. In Table 2 we present the estimates of the callers' service reward (r) along with their per-minute online and offline waiting costs (c^n, c^f). While the callers' online waiting cost is statistically significant and a positive value, the offline waiting cost is statistically equal to zero. One interpretation of the low offline waiting estimate is that callers may be more concerned about the manner in which they wait for service than the duration of their wait. While callers who accept a callback offer are free to perform offline tasks as they wait for their callback to arrive, callers in the online queue must stay on their phones, which limits the set of tasks they can perform while they wait for service. However, we note that callers in our sample only receive guarantee durations of no greater than 59 minutes. It is possible that under longer guarantee durations, callers in this call center would have greater offline waiting costs. Finally, our waiting cost estimates have important implications regarding how caller waiting affects service quality. Recall that each period that callers wait for service, they accumulate waiting disutility equal to their estimated online or offline waiting cost, and that we define service quality as the average waiting disutility of callers in the system. The important implication from our estimates is that service quality reduces as callers wait in the online queue, but does not significantly reduce as they wait offline for a callback.

Table 2: Estimates of Callers' Reward, and per Minute Online and Offline Waiting Cost

Parameter Name	Symbol	Estimate	Standard Error
Service reward ($)	r	5.570**	(0.006)
Online waiting cost ($/minute)	c^n	0.112**	(0.001)
Offline waiting cost ($/minute)	c^f	1.73E-11	(7.31E-07)

(**) denotes significant at $p < .0001$.

Insight 2: Managers may use the average online waiting time of *all* callers in the system as a metric for service quality. Recall that historically service quality has been measured as a function of how long callers wait for service. For example, in many call centers service quality is measured by the callers' average waiting time. However, in call centers that offer callbacks, callers may experience two types of waiting - online waiting and offline waiting. Thus, a pertinent question managers may have is what metric they should use to measure service quality under a callback policy. While the most direct metric for service quality would be the average disutility callers accumulate while waiting for service, in practice managers may not be able to capture this directly from the data. However, our estimates show that callers only accumulate disutility while

waiting in the online queue, indicating that the average time all callers in the system wait in the offline queue has no bearing on service quality. Hence, managers can measure service quality simply as the average online waiting time of all callers in the system, which includes callers who accepted a callback offer (and, hence, had zero online wait). Thus, managers who are interested in maximizing service quality should concentrate on reducing this metric by substituting offline waiting time for online waiting time through offering callbacks.

Insight 3: Callers' valuation of service guarantees doesn't depend on the duration of the guarantee. In Table 3 we display the estimates of the callers' guarantee utilities (μ_G^f), which depend on the callers' guarantee message (G). Recall that we grouped callers' guarantee messages into distinct subsets, and assumed that callers who receive guarantee messages in the same subset have the same parameters. Thus, we estimated a separate guarantee utility and switching cost for callers whose messages are in each subset. We first discuss the callers' guarantee utility, which is their utility of receiving a guarantee message of G, and is a component of the nominal utility of accepting a callback offer in (4). The estimated guarantee utilities are positive but do not appear to vary greatly with the callers' guarantee message, as the estimates only range between 1.52 and 1.61. This small range of estimates suggests that receiving a service guarantee of any duration is more important to callers than the duration of the guarantee itself. One possible reason for this is that regardless of the duration of their guarantee, callers value the assurance that they will receive a callback at a specified time. This assurance allows callers to confidently plan their offline tasks, and lies in contrast with the online waiting process, where callers may receive a delay estimate but no promise of when they will receive service.

Table 3: Estimates of Callers' Guarantee Utility, Switching Cost, and Utility of Accepting Callback Offer by Subset of Guarantee Messages

Message Subset ($\mathscr{A}$)	Duration(s) (Minutes)	μ_G^f: Guarantee Utility Estimate ($)	St. Error	s_G^f: Switching Cost Estimate ($)	St. Error	Utility of Accept CB
1	5	1.57**	(0.02)	3.14**	(0.03)	4.09
2	6	1.55**	(0.02)	3.09**	(0.03)	4.11
3	7	1.58**	(0.03)	3.04**	(0.03)	4.18
4	8	1.58**	(0.03)	3.02**	(0.03)	4.21
5	9	1.56**	(0.03)	2.94**	(0.03)	4.27
6	10-11	1.60**	(0.02)	2.95**	(0.02)	4.30
7	12-13	1.56**	(0.02)	2.93**	(0.02)	4.27
8	14-15	1.61**	(0.02)	2.95**	(0.02)	4.30
9	16-18	1.56**	(0.03)	2.92**	(0.03)	4.28
10	19-22	1.55**	(0.03)	2.93**	(0.02)	4.26
11	23-29	1.56**	(0.03)	2.99**	(0.03)	4.22
12	30-59	1.52**	(0.03)	3.00**	(0.02)	4.16

(**) denotes significant at $p < .0001$.

Insight 4: The process of switching from offline tasks in order to answer a callback is

costly for callers. In Table 3 we display the callers' switching costs (s_G^f), which depend on the callers' guarantee message (G). Recall that the callers' task switching cost (s_G^f) is the cost that callers incur to switch from their offline task to answer a callback, and that from (6) the nominal utility of answering a callback is the callers' reward minus their switching cost ($r - s_g^f$). Depending on their guarantee message, the callers' estimated switching costs range between 2.92 and 3.14, all of which are more than half of their estimated reward of 5.570 in Table 2. The size of the callers' estimated switching costs suggests that callers perceive the loss of continuity from switching their offline tasks to answer a callback to be costly. However, since the callers' reward parameter is sufficiently high, most callers answer a callback when it arrives. Based on the callers' estimated reward and switching costs, we calculate that the callers' probability of answering a callback ranges between 92.6% and 94.0%, which is consistent with what we observe in the data (see Figure 4 in §3). We also note that, similar to their guarantee utilities, the callers' switching costs do not vary greatly with their guarantee messages, as they range between 2.92 and 3.14.

Insight 5: This call center could change their current policy by offering callbacks with longer guarantee durations without sacrificing callback offer acceptance rates. We conclude this section by discussing how the estimated parameters relate to the callers' probability of accepting callbacks under various policies. In the last column of Table 3, we display the callers' utilities of accepting a callback offer. As a reminder, from (4) the callers' utility of accepting a callback offer depends on their reward (r), guarantee utility (μ_G^f), offline waiting cost (c^f), and switching cost (s_G^f). Note that the utilities only range between 4.06 and 4.30. This is because callers incur negligible offline waiting costs and their guarantee utilities and switching costs have little dependence on their guarantee message. Consequently, the callers' utility of accepting a callback offer is insensitive to the duration of their service guarantee. This means that for a given delay announcement a caller would be nearly as likely to accept a callback offer with a guarantee duration of 5 minutes as an offer with a duration of 59 minutes. Consequently, this call center could change their current policy by offering callbacks with longer guarantee durations without sacrificing callback offer acceptance rates. Indeed, in our counterfactual section we test policies with guarantee durations longer than the current policy and find that they may increase service quality and/or system throughput.

5.2 Model Validation

To test our model's accuracy in predicting callers' abandonment and callback behavior, we conduct out of sample testing. We break our data into a training set, which contains calls that arrived between April 13, 2016 and June 30, 2016, and a holdout set, which contains calls that arrived between July 1, 2016 and July 31, 2016. Recall that our data spans from April 13, 2016 to July 31, 2016. We first use the data from the training set to estimate the callers' parameters. We then use the estimates from the training set to predict the callers' decisions during the holdout period using a Monte Carlo procedure. We collect the callers' decisions from our Monte Carlo procedure

and calculate the predicted aggregate performance measures of the call center during the holdout period. Finally, we compare the predicted performance measures with the actual performance measures of the call center during the holdout period.

In Table 4 we compare the out of sample predictions to the actual performance during the holdout period using relative error, which is given by $|\text{predicted} - \text{actual}|/\text{actual}$. We observe that our model provides reasonable predictions of the performance of the online queue, with relative errors of 1.9% and 9.1% for the average online waiting time and the abandonment rate, respectively. We find that our model also provides reasonable predictions of callers' callback decisions, with a relative error of 11.1% for the percentage of callers who accepted a callback when offered and 0.4% for the percentage of callers who answered their callback after accepting a callback offer.

Table 4: Out of Sample Performance

Performance Measure	**Actual**	**Predicted**	**Error**
Average online waiting time (seconds)	225.6	229.9	1.9%
Abandonment rate from online queue	22.7%	20.6%	9.1%
% Accepting callback when offered	35.2%	39.1%	11.1%
% Answering callback after accepting offer	92.7%	93.1%	0.4%

6 Counterfactual Analysis

In this section we would like to understand how various callback policies affect service quality and system throughput. For example, does the current policy increase service quality and system throughput? If so, could these increase even further under another policy? Because we have characterized the callers' decision making process, we can explore these questions by performing a counterfactual analysis of how various callback policies affect the service quality and system throughput of this call center.

We begin this section by providing an overview of our simulation process. We then describe four callback policies that we use in our analysis. We next explain how we measure service quality and system throughput and define two additional measures that assist us in explaining our results. Finally, we provide the simulation conditions and present the results.

Simulation Process: We conduct our counterfactual analysis through a series of discrete event simulations, where we simulate the operations of this call center in which callers make their decisions as described by the process in §4. Unless otherwise stated, we keep all of the previous assumptions from our model and estimation strategy. In our estimation strategy we assumed that callers who wait in the online queue make decisions every 10 seconds. We maintain this assumption and therefore set the unit of time in our simulations to 10 seconds. In all of our simulations we use the arrival rate and service time distribution from the peak operating hours, which we define as Monday through Friday, 9:00 AM to 5:00 PM. In each analysis we perform separate simulations of

the call center under various staffing levels, where the number of servers varies between 132 and 152. We pick these staffing levels to test the the policies under a variety of system loads.

A unique challenge of our counterfactual analysis is that we must simulate the callers' decisions within a rational expectation equilibrium framework. Recall that in our model the callers' beliefs about their probability of receiving service in a given period match their actual probabilities of receiving service, in a rational expectation equilibrium. Since the service probabilities $(\pi_A(t), t \geq 0)$ are an outcome of the callers' decisions under a specific policy, they must be redetermined under each unique set of policies and operating conditions. To do this, we use an iterative process similar to Aksin et al. (2013). We first simulate the system under the assumption that $\pi_A(\cdot)$ is a vector of zeros. We then estimate the service probabilities $(\pi_A^l(\cdot), l = 1)$ from the simulated data. We next simulate the system where callers believe that the service probabilities $(\pi_A^l(\cdot), l = 1)$ estimated from the first iteration are the service probabilities in the second iteration, and then obtain new estimates of the service probabilities $(\pi_A^l(\cdot), l = 2)$ to be used in the following iteration. We iterate through this process $(l \geq 3, 4, \cdots)$ until the average waiting times in the online queue converge.

Policies: We select four policies for our analysis, which include a policy where no callbacks are offered, the current policy of this call center, a policy from the extant literature that offers callers a fixed guarantee duration (Armony and Maglaras (2004b)), and a policy that provides callers with a window in which their callback will arrive, rather than providing a specific predetermined time. We describe each of these policies:

- **No-Callback with Delay Announcements Policy (N):** Callers receive a delay message, which contains their estimated delay based on the state of the system[5], but receive no callback offer. Upon entering the system, all callers who are not immediately served or abandon join the online queue and wait until they receive service or abandon from the online queue. We choose this policy as a baseline to compare against other policies that offer callbacks.[6]

- **Status Quo Policy (SQ):** This is the call center's current policy as illustrated in Figure 1 in §3. Under this policy if the callers' delay estimate is between 5 and 59 minutes, callers are offered a callback with a guarantee duration equal to the delay estimate. A theoretical disadvantage of this policy is that it does not postpone callers' service requests from periods where the call

[5]In simulations of policy N we provide arriving callers with a delay message, which contains their expected waiting time if they join the online queue, rounded down to the nearest minute. Similar to our message partitions in §4.2 we group the delay messages into 13 distinct subsets for estimating $\pi_A(\cdot)$ after each simulation iteration. We choose the following message partition: $\mathscr{A}_0 = \{0, 1, \cdots, 4\}, \mathscr{A}_1 = \{5\}, \mathscr{A}_2 = \{6\}, \mathscr{A}_3 = \{7\}, \mathscr{A}_4 = \{8\}, \mathscr{A}_5 = \{9\}, \mathscr{A}_6 = \{10, 11\}, \mathscr{A}_7 = \{12, 13\}, \mathscr{A}_8 = \{14, 15\}, \mathscr{A}_9 = \{16, 17, 18\}, \mathscr{A}_{10} = \{19, 20, 21, 22\}, \mathscr{A}_{11} = \{23, 24, \cdots, 29\}$, and $\mathscr{A}_{12} = \{30, 31, \cdots\}$. This partition is similar to the one we used in our estimation procedure, but accounts for delay messages with expected waiting times less than 5 minutes and greater than 59 minutes. We use the waiting time of the most recent caller who received service in the online queue as the estimate of an arriving caller's expected waiting time in the online queue. This estimator is referred to as the Last-to-Enter-Service (LES) Delay Estimator; see Ibrahim and Whitt (2009) for an analysis of the performance of the LES Estimator. We find that this estimator provides reasonable predictions of callers' online waiting times.

[6]We also performed our analysis under a no-callback policy that does not provide callers with delay messages and found that the performance of the call center under this policy varied little with the performance of the call center under policy N.

center has insufficient service capacity to periods where the call center has excess service capacity. Rather, this policy has the effect of holding the callers' place in line. Consequently, we test two additional policies which may be used by this call center as demand postponement strategies.

- **Fixed Guarantee Policy (FG(D)):** This is the callback policy that was analyzed in Armony and Maglaras (2004b). All callers who enter the system receive a delay message containing their expected waiting time[7] if they join the online queue and a callback offer with a fixed service guarantee of D periods.[8] Unlike this call center's current policy, the callers' guarantee duration does not vary with their delay estimate but remains fixed regardless of the state of the system. The service guarantees are met using a queue-length threshold, nonpreemptive, head-of-line policy, which follows:

  If $Q^f(\delta) \geq R^f(\delta) - R^f(\delta - D)$, *give priority to the offline queue, otherwise give priority to the online queue,*

  where δ is the number of periods since the beginning of the simulation, $Q^f(\delta)$ is the length of the offline queue in period δ, and $R^f(\delta)$ is the number of callers who have arrived to the offline queue in periods $\{0, 1, \cdots, \delta\}$.[9]

- **Window Policy (W(LB, UB)):** Even though the advantages of policy FG have been shown in the literature, it has two shortcomings. First, all callers who enter the system receive a callback offer even when there are servers available to immediately serve them. Second, the system determines when to initiate callbacks only based on the length of the offline queue without considering the state of the online queue. Consequently, there may be cases where there are no callers waiting in the online queue but the system does not initiate callbacks because the offline queue length does not exceed some threshold. We therefore test a policy which seeks to address these two shortcomings and refer to it as the Window Policy (W(LB,UB)). To address the first shortcoming, new arrivals only receive a callback offer if there are no servers available to immediately serve them. To address the second shortcoming, callers who receive a callback offer receive a window in which their callback may arrive, where LB (UB) is the earliest (latest) period in which the callback may arrive. Under policy W callers who accept a callback offer will not be contacted before LB. After LB, they will be contacted only if there are no callers waiting in the

[7] Similar to policy N, we estimate the callers' expected waiting time in the online queue under policy FG using the LES delay estimator. We also choose the same message partition as policy N.

[8] In Armony and Maglaras (2004a) the authors analyze the performance of a callback policy where callers are offered a callback with a fixed service guarantee, but rather than receiving a delay estimate based on the current state of the system, callers only receive the steady-state mean waiting time of callers who join the online queue. In Armony and Maglaras (2004b) the authors show that providing callers with delay estimates based on the current state of the system leads to lower waiting times and higher system throughput than the limited information policy in Armony and Maglaras (2004a). Consequently, we choose to test the policy from Armony and Maglaras (2004b).

[9] Armony and Maglaras demonstrate that this threshold policy is asymptotically compliant, which means that under heavy traffic and a traffic intensity near 1, as the number of servers in the system increases the percentage of guarantees that are missed becomes negligible. We find that this threshold policy consistently meets callers' service guarantees in our simulations.

online queue. To ensure that guarantees are met, offline callers receive priority once they have reached the end of their window (UB). Thus, under this policy the system finds opportunities to offer and initiate callbacks when they will have little or no impact on online waiting times while maintaining the call center's commitment to initiate callbacks within a guaranteed timeframe. We assume that callers believe the actual time until the callback arrives is uniformly distributed between LB and UB minutes. We then find the callers' expected utility of accepting a callback offer based on the estimates from Table 3 in §5.1.

Finally, to be clear, in Armony and Maglaras (2004b) the authors did not claim that policy FG achieved some optimal performance measure. Rather, they were interested in analyzing the performance of the policy under heavy traffic and suggesting a threshold policy for meeting callers' service guarantees. We also make no claim that any of the other policies that we test achieve some optimal performance measure in this call center. Our aim in testing these policies is to gain insights into how this call center would perform under various callback policies.

Measures: We next explain how we measure the effect of these policies on service quality and system throughput.

- **Average Waiting Disutility of All Callers who Entered the System (AWD_All):** We calculate this measure as AWD_All = (total online waiting disutility + total offline waiting disutility) / number of callers who entered the system. Recall that we defined waiting disutility as the disutility that callers accumulate while waiting online or offline to receive service. We use the average waiting disutility of all callers in the system as the measure of service quality, where higher values of this measure indicate lower service quality. Also recall that the callers' per period cost of waiting in the offline queue is statically zero. Hence, this measure is essentially the total online waiting disutility divided by the number of callers who entered the system.

- **System Throughput:** We calculate this measure as System Throughput = (number of callers who received service in the online queue + number of callers who received service in the offline queue (answered a callback)) / simulation duration. We are able to directly measure system throughput by accounting for the two ways that callers may receive service, which include receiving service from the online queue and answering a callback.

In addition to our performance measures, we collect two waiting time measures from our simulation data to assist us in explaining the results:

- **Average Online Waiting Time of All Callers who Entered the System (AWT_All):** We calculate this measure as AWT_All = total online waiting time in the system / number of callers who entered the system. This is a measure of how the online waiting time spreads across all callers who entered the system, including callers who accepted a callback offer and thus did not experience online waiting. Recall from §5.1 that managers may use AWT_All as a suitable metric for service quality since it is directly proportional to the callers' average waiting disutility (AWD_All).

- **Average Online Waiting Time of Callers who Waited in the Online Queue (AWT_On):** We calculate this measure as AWT_On = total online waiting time in the system / number of callers who waited in the online queue[10]. This measures the mean waiting time of callers who waited in the online queue. We include this measure since managers may want to know how various callback policies affect the waiting times of callers who wait in the online queue.

Simulation Conditions: To determine how the four callback policies perform in this call center under different loads, we vary the number of servers in the system while maintaining the peak arrival rate and service time distribution. For each policy we simulate the system with 132, 136, 140, 144, 148, and 152 servers to explore a broad range of loads. For the FG policy we select a guarantee duration of 30 minutes but find that our insights hold for the entire range of available guarantee durations (between 5 and 59 minutes). We simulate the system under two cases of policy W, where the callers' window is centered around 30 minutes. In the first case we set the callers' window for receiving a callback to be between 28 to 32 minutes. In the second case the window is set to be between 20 and 40 minutes. We also find that our insights hold for various other windows.

Result 1: Offering callbacks increases service quality by decreasing the callers' average waiting disutility (AWD_All). In Table 5 we present the results of our analysis. We observe that AWD_All is highest under policy N for all loads, indicating that offering callbacks reduces AWD_All. This occurs because offering callbacks reduces the average time that callers in the system wait in the online queue (AWT_All) by channeling callers to the offline queue. Given that the offline waiting cost is zero, this in turn reduces the callers' average waiting disutility, which increases service quality. However, we note that the magnitude of the service quality improvement depends on the callback policy. For example, policy SQ leads to higher average waiting disutility than policies FG and W, since callbacks are only offered under policy SQ when online waiting times are between 5 and 59 minutes. We also find that policy W results in the lowest average waiting disutility due to its flexibility in determining when to offer and when to initiate callbacks. Finally, we note that under low loads (148 and 152 servers), policy W(20,40) leads to lower average waiting disutility than policy W(28,32). This is because under policy W(20,40) callers have a longer guarantee window, which gives the system more flexibility in determining when to initiate callbacks. We also note that under low loads, average waiting disutility is only slightly lower under policy SQ than policy N,[11] and that under high loads, policies FG and W result in nearly the same

[10] We consider callers who are immediately served upon entering the system as receiving service from the online queue with a waiting time of zero periods.

[11] Under low loads (140, 144, 148, and 152 servers) AWD_All is only slightly lower under policy SQ than policy N. This is due to the fact that under low loads the system rarely offers callbacks under policy SQ as online waiting times rarely exceed five minutes.

average waiting disutility.[12]

Table 5: System Performance under Various Callback Policies and Loads

Servers	Policy	AWD_All	Throughput	AWT_All	AWT_On
132	N	0.488	3.62	261	261
132	SQ	0.457	3.62	245	269
132	FG(30)	0.402	3.62	215	298
132	W(28,32)	0.399	3.62	213	298
132	W(20,40)	0.399	3.63	214	299
136	N	0.389	3.74	208	208
136	SQ	0.384	3.74	205	209
136	FG(30)	0.314	3.74	168	226
136	W(28,32)	0.307	3.74	164	223
136	W(20,40)	0.306	3.74	164	223
140	N	0.285	3.86	152	152
140	SQ	0.284	3.86	152	152
140	FG(30)	0.228	3.85	122	161
140	W(28,32)	0.217	3.86	116	153
140	W(20,40)	0.218	3.86	117	155
144	N	0.205	3.94	110	110
144	SQ	0.205	3.94	109	109
144	FG(30)	0.160	3.93	86	111
144	W(28,32)	0.124	3.96	67	85
144	W(20,40)	0.123	3.97	66	85
148	N	0.125	4.03	67	67
148	SQ	0.124	4.03	67	67
148	FG(30)	0.089	4.00	47	61
148	W(28,32)	0.058	4.04	31	37
148	W(20,40)	0.047	4.05	25	31
152	N	0.068	4.08	36	36
152	SQ	0.068	4.08	36	36
152	FG(30)	0.048	4.04	26	33
152	W(28,32)	0.016	4.10	8	9
152	W(20,40)	0.013	4.10	7	8

AWT_On and AWT_All are measured in seconds.

[12]Under high loads (132 and 136 servers), policies FG and W result in nearly the same AWD_All for two reasons. First, because servers are almost always busy under high loads, almost all new arrivals are offered a callback under policy W. This is nearly the same as policy FG, which offers all new arrivals a callback. Second, recall that under policy W callbacks are only initiated before the end of the callers' window if there are no calls waiting in the online queue. However, under high loads there are almost always calls waiting in the online queue, which means that nearly all offline callers receive their callbacks at the end of their window. Hence, under heavy loads policy W loses nearly all flexibility in determining when to offer and when to initiate callbacks.

Result 2: Under low loads, offering callbacks decreases the average waiting time of callers who wait in the online queue (AWT_On), but increases it under high loads. As can be seen in Table 5, under low loads (148 and 152 servers), offering callbacks decreases the average waiting time of callers who join the online queue (AWT_On). However, under high loads (136 and 132 servers) offering callbacks increases AWT_On. Because we are interested in understanding why this occurs, we explore this phenomenon in Appendix B. In short, we find that offering callbacks under low (high) loads decreases (increases) the traffic intensity within the online queue, which we posit decreases (increases) the average waiting time of callers who wait in the online queue.

Result 3: Offering callbacks under high loads has no substantial impact on system throughput. Offering callbacks under low loads may slightly increase or decrease system throughput depending on the callback policy. Under high loads (132, 136, and 140 servers) system throughput appears to be unaffected by the policy choice. This is because the system is operating near its maximum capacity. Under low system loads (144, 148, and 152 servers) system throughput increases under policy W, but decreases under policy FG.[13] However, the magnitude of the changes are no more than 1%. Hence, offering callbacks in this setting does not appear to have a substantial impact on system throughput.

7 Discussion and Summary

Call centers are increasingly turning to callback technology to mitigate the negative effects of caller waiting. While the extant literature has analytically characterized the performance of call centers that offer callbacks, to the best of our knowledge, we perform the first empirical study of callers' behavior in the presence of a callback option. To characterize the callers' decision-making process, we formulated and estimated the parameters of a structural model of callers' behavior under a callback policy. Our estimates of the callers' preferences reveal important insights regarding how callers perceive the value of callback offers. For example, we found that the callers' per unit offline waiting cost is statistically equal to zero, which implies that waiting offline to receive a callback does not reduce service quality. Furthermore, managers whose objective is to maximize service quality should focus on reducing the average online waiting time of all callers in the system by substituting offline waiting for online waiting through offering callbacks. We also note that this metric may be used in future analytical work to determine the optimal callback policy with respect to service quality. Our estimates also reveal that the callers' utility of accepting a callback offer is insensitive to the duration of their service guarantee, indicating that this call center may implement

[13] In Armony and Maglaras (2004a,b) the authors show that system throughput increases under policy FG. Our result differs from theirs due to differences between the settings. First, in our setting, callers may balk upon entering the system but may also abandon after joining the online queue, while callers in Armony and Maglaras (2004a,b) may only balk. Second, in our setting callers who accept a callback offer may later reject the callback when it arrives, while all callers who accept a callback offer answer their callbacks in Armony and Maglaras (2004a,b). Consequently, offering callbacks in our setting may increase or decrease system throughput, depending on how the policy affects the volume of callers who abandon and the volume of callers who reject callbacks when they arrive.

demand postponement policies without sacrificing callback offer acceptance rates.

Using the estimates from our model, we performed a counterfactual analysis of how various callback policies affect the service quality and system throughput of this call center. We found that offering callbacks improves service quality but has no substantial impact on system throughput. Based on our results, we recommended that this call center adopt a policy where callers receive a window in which the callback may arrive, rather than a specific guaranteed duration. This policy allows the call center to postpone service requests from periods of insufficient service capacity to periods of excess capacity. In testing this policy we assumed that the callers' utility of accepting a callback offer with a window is based on their estimated utility of accepting a callback offer with a specific guarantee duration. In reality, callers may value callback windows differently than our assumption. Considering the potential benefits to service quality, we recommend that this call center explore how readily callers would accept callback offers with various callback windows.

There are several ways this work can be extended. First, our framework could be applied to other service settings. For example, a recent trend in amusement parks is to offer visitors who arrive at the entrance of a ride a ticket containing a window in which they may later return to the ride and bypass other visitors who are waiting in line. It could be interesting to determine how park visitors use the observed length of the line and the conditions of their ticket to make their decisions. Another interesting extension would be to use our framework to explore how other callback policies affect service quality and system throughput. For example, one could test different thresholds for when to offer callbacks and when to initiate callbacks based on the state of the system. Finally, we demonstrated in our study that service quality substantially improves under policy W, where callers receive a window in which their callback will arrive. Thus, a potentially valuable extension of this study would be to explore how callers will respond to callback offers with various windows.

References

Aksin, Z., M. Armony, and V. Mehrotra (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management 16*(6), 665–688.

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

Akşin, Z., B. Ata, S. M. Emadi, and C.-L. Su (2016). Impact of delay announcements in call centers: An empirical approach. *Operations Research*.

Armony, M. and C. Maglaras (2004a). Contact centers with a call-back option and real-time delay information. *Operations Research 52*(4), 527–545.

Armony, M. and C. Maglaras (2004b). On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research 52*(2), 271–292.

Ata, B. and X. Peng (2017). On the control of a call center with the callback option. *Working Paper*.

Ben-Akiva, M. E. and S. R. Lerman (1985). *Discrete choice analysis: theory and application to travel demand*, Volume 9. MIT press.

Bitran, G. R., J.-C. Ferrer, and P. Rocha e Oliveira (2008). Om forum-managing customer experiences: Perspectives on the temporal aspects of service encounters. *Manufacturing & Service Operations Management 10*(1), 61–83.

Biyalogorsky, E., Z. Carmon, G. E. Fruchter, and E. Gerstner (1999). Research note: Overselling with opportunistic cancellations. *Marketing Science 18*(4), 605–610.

ContactBabel (2016). *US contact center decision makers' guide, 9th edition.* http://www.contactbabel.com/reports.cfm.

Emadi, S. and J. Swaminathan (2017). Impact of callers' history of abandonment: Model and implications. *Working Paper*.

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

Hassin, R. and M. Haviv (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*, Volume 59. Springer Science & Business Media.

Horowitz, J. L. (2001). The bootstrap. *Handbook of econometrics 5*, 3159–3228.

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

Iyer, A. V., V. Deshpande, and Z. Wu (2003). A postponement model for demand management. *Management Science 49*(8), 983–1002.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association 53*(282), 457–481.

Leclerc, F., B. H. Schmitt, and L. Dube (1995). Waiting time and decision making: Is time like money? *Journal of Consumer Research 22*(1), 110–119.

Legros, B., O. Jouini, and G. Koole (2016). Optimal scheduling in call centers with a callback option. *Performance Evaluation 95*, 1–40.

Mandelbaum, A. and N. Shimkin (2000). A model for rational abandonments from invisible queues. *Queueing Systems 36*(1-3), 141–173.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society*, 15–24.

Odlyzko, A. (1999). Paris metro pricing for the internet. In *Proceedings of the 1st ACM conference on Electronic commerce*, pp. 140–147. ACM.

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

Suck, R. and H. Holling (1997). Stress caused by waiting: A theoretical evaluation of a mathematical model. *Journal of mathematical psychology 41*(3), 280–286.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Yu, Q., G. Allon, and A. Bassamboo (2016). How do delay announcements shape customer behavior? an empirical study. *Management Science*.

A Model Identification

In this appendix we discuss how we identify the callers' parameters. We first discuss identification of the callers' service reward and online waiting cost. We then show identification of the switching cost. We finish by discussing how we identify the callers' offline waiting cost and guarantee utilities.

Reward and Online Waiting Cost: To show identification of the callers' reward and online waiting cost (r, c^n), we rely on the reasoning provided in Emadi and Swaminathan (2017). Given Equation (8), for $d = 0$ we obtain

$$P_{i(t+1)}(0; A_i, G_i, b_i, \Theta) = \frac{1}{1 + \exp(v(t+1, A_i, G_i, 1; \Theta))}. \tag{9}$$

Using Equation (9) and (3) gives us

$$\begin{aligned} V^n(t, r, c^n, \pi_{A_i}(\cdot)) &= \log(1/P_{i(t+1)}(0; A_i, G_i, b_i, \Theta)) \\ &= -\log(P_{i(t+1)}(0; A_i, G_i, b_i, \Theta)). \end{aligned}$$

Given (2) and (8), we have

$$\begin{aligned} &\log(1/P_{it}(0; A_i, G_i, b_i, \Theta) - 1) \\ &= v(t, A_i, G_i, 1; \Theta) \\ &= -c^n + \pi_{A_i}(t)r - (1 - \pi_{A_i}(t)) \log(P_{i(t+1)}(0; A_i, G_i, b_i, \Theta)). \end{aligned} \tag{10}$$

We denote the probability of caller i abandoning in period t by P^{ab}_{it}, and rearrange Equation (10) to obtain

$$\log(1/P^{ab}_{it} - 1) + (1 - \pi_{A_i}(t)) \log(P^{ab}_{i(t+1)}) = -c^n + \pi_{A_i}(t)r.$$

Note that the terms on the left hand side of the equation are all observed in the data, which indicates that r and c^n on the right hand side are identified by observed variation in the left hand

terms. We provide a brief explanation using a regression analogy. Assume that $y_t = \log(1/P_{it}^{ab} - 1) + (1 - \pi_{A_i}(t))\log(P_{i(t+1)}^{ab})$ and $x_t = \pi_{A_i}(t)$. Then regressing y_t on x_t would yield an intercept $(-c^n)$ and a slope (r) that depends on $\pi_{A_i}(t)$. This demonstrates that variation in the callers' abandonment behavior and service probabilities $\pi_A(\cdot)$ across different periods and delay messages provides identification of the reward and online waiting cost parameters.

Switching Cost: Given that r is identified above, the vector of callers' switching costs (S^f) is identified by variation in callers' decisions of whether to reject an arriving callback over different guarantee messages. Given Equations (8), (6), and (5), the probability of rejecting an arriving callback is given by

$$P_{it}(3; A_i, G_i, b_i, \Theta) = 1/(1 + \exp(r - s_{G_i}^f)).$$

Denoting by P_{it}^{rj} the probability of caller i rejecting an arriving callback in period t, and solving for $s_{G_i}^f$, we obtain

$$s_{G_i}^f = r + \ln(P_{it}^{rj}/(1 - P_{it}^{rj})).$$

Since r on the right hand side has been previously identified, each switching cost is uniquely identified by the callers' probability of rejecting an arriving callback with a given guarantee message.

Offline Waiting Cost and Guarantee Utility: Given that the remaining caller parameters are identified above, the callers' offline waiting cost (c^f) and vector of guarantee utilities (U^f) is identified by variation in the callers' probability of accepting a callback offer over different guarantee messages and guarantee durations. To simplify exposition, we denote by v_{i0}^n the nominal utility that caller i receives by waiting online in period 0, as given in (2). Then given Equations (8), (7), (4), and (1), the probability of accepting a callback offer is given by

$$P_{i0}(2; A_i, G_i, b_i, \Theta) = \frac{\exp(-c^f D_{G_i} + \mu_{G_i}^f + \log(1 + \exp(r - s_{G_i}^f)))}{1 + \exp(v_{i0}^n) + \exp(-c^f D_{G_i} + \mu_{G_i}^f + \log(1 + \exp(r - s_{G_i}^f)))}. \tag{11}$$

We denote by P_{i0}^{ac} the probability of caller i accepting a callback offer in period 0, and perform algebraic manipulations of (11) to obtain

$$\log\left(\frac{P_{i0}^{ac}(1 + \exp(v_{i0}^n))}{(1 - P_{i0}^{ac})(1 + \exp(r - s_{G_i}^f))}\right) = -c^f D_{G_i} + \mu_{G_i}^f.$$

Note that term on the left side is identified through the observed probability of callers accepting a callback offer and the previously identified structural parameters. Relying on the regression analogy from earlier, we see that the left hand term acts as the dependent variable, $\mu_{G_i}^f$ acts as an intercept for a given guarantee message, and c^f acts as a slope which depends on the guarantee duration (D_{G_i}). This demonstrates that variation in the callers' probability of accepting a callback offer over different guarantee messages and guarantee durations provides identification of the callers' offline waiting cost and their vector of guarantee utilities.

B Effect of Offering Callbacks on AWT_On

In this appendix we discuss how offering callbacks affects the average online waiting time of callers who wait in the online queue (AWT_On). Interestingly, we observe in Table 5 in §6 that offering callbacks increases AWT_On under high loads but decreases it under low loads. To help us understand why this occurs, we calculate an additional measure which we call the online traffic intensity, denoted by ρ_{on}. We calculate this measure as ρ_{on} = arrival rate to the online queue / maximum service rate in the online queue, where the maximum service rate in the online queue = maximum service rate of the system - callbacks answered per period. We use this measure as a proxy for the load in the online queue, and explain our intuition behind this measure. On the one hand, offering callbacks decreases the arrival rate to the online queue by channeling callers who accept a callback offer to the offline queue. On the other hand, the maximum service rate in the online queue decreases due to the servers' requirement to serve callers who answer their callbacks. The rate by which the maximum service rate in the online queue decreases is equal to the rate at which callers answer callbacks. This means that offering callbacks may cause the online traffic intensity to be greater than or less than the system traffic intensity, depending on how much the arrival rate to the online queue and the maximum service rate to the online queue decrease.

Table 6: AWT_On and Online Traffic Intensity under Various Callback Policies and Loads

(a)

Servers	Policy	AWT_On	ρ_{on}
132	N	261	1.138
132	SQ	269	1.145
132	FG(30)	298	1.165
132	W(28,32)	298	1.167
132	W(20,40)	299	1.168
136	N	208	1.105
136	SQ	209	1.105
136	FG(30)	226	1.116
136	W(28,32)	223	1.117
136	W(20,40)	223	1.115
140	N	152	1.073
140	SQ	152	1.073
140	FG(30)	161	1.074
140	W(28,32)	153	1.073
140	W(20,40)	155	1.073

(b)

Servers	Policy	AWT_On	ρ_{on}
144	N	110	1.043
144	SQ	109	1.043
144	FG(30)	111	1.034
144	W(28,32)	85	1.035
144	W(20,40)	85	1.036
148	N	67	1.015
148	SQ	67	1.015
148	FG(30)	61	1.000
148	W(28,32)	37	1.004
148	W(20,40)	31	1.003
152	N	36	0.988
152	SQ	36	0.988
152	FG(30)	33	0.968
152	W(28,32)	9	0.979
152	W(20,40)	8	0.978

In Tables 6(a) and 6(b) we present AWT_On and the online traffic intensity (ρ_{on}) from our analysis in §6. Under low loads (148 and 152 severs) the average waiting time of callers who joined the online queue (AWT_On) is lowest under the FG and W policies. However, AWT_On is lowest under policy N when loads are higher (132 and 136 servers), which means that under high loads

offering callbacks can actually increase the mean waiting times of callers who wait in the online queue. While it is surprising that offering callbacks can increase average online waiting times, the online traffic intensity offers a potential explanation for why this occurs. We observe that when loads are low (148 and 152 servers), ρ_{on} is lowest under policies FG and W. Conversely, when loads are high (132 and 136 servers), ρ_{on} is highest under policies FG and W. We give a brief explanation for this. Recall that ρ_{on} is calculated as the arrival rate to the online queue divided by the maximum service rate in the online queue, and that offering callbacks reduces both the numerator (the arrival rate to the online queue) and the denominator (the maximum service rate in the online queue) of the online traffic intensity formula. Thus, the effect of offering callbacks on ρ_{on} depends on how the reduction in the online arrival rate compares to the reduction in the maximum service rate in the online queue. We find that under higher (lower) loads the percentage reduction in the online arrival rate is less than (greater than) the percentage reduction in the maximum service rate of the online queue, resulting in higher (lower) values of ρ_{on}. Because ρ_{on} is an indication of the load in the online queue, we would expect that ρ_{on} would be positively correlated with the callers' average waiting times in the online queue. Indeed, we observe this correlation in our results. Under higher loads (132 and 136 servers) both ρ_{on} and AWT_On are highest under the FG and W policies, while under lower loads (148 and 152 servers) both ρ_{on} and AWT_On are lowest under the FG and W policies. This trend suggests that offering callbacks above (below) some threshold system load leads to a higher (lower) online traffic intensity, which correlates with the average waiting times of callers who wait in the online queue. An interesting extension of this work would be to establish analytically whether our intuition regarding this threshold is correct and, if so, to find a solution for the threshold above (below) which offering callbacks increases (decreases) the average waiting time of callers who wait in the online queue.