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

MANAGEMENT SCIENCE
Vol. 68, No. 8, August 2022, pp. 5745–5757
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Delay Information in Virtual Queues: A Large-Scale Field Experiment on a Major Ride-Sharing Platform

Qiuping Yu,[a,b,*] Yiming Zhang,[c,d] Yong-Pin Zhou[e]

[a] Scheller College of Business, Georgia Institute of Technology, Atlanta, Georgia 30308; [b] McDonough School of Business, Georgetown University, Washington, DC 20057; [c] School of Management and Economics, The Chinese University of Hong Kong, Shenzhen, Guangdong 518172, China; [d] Shenzhen Finance Institute, Shenzhen, Guangdong 518000, China; [e] Foster School of Business, University of Washington, Seattle, Washington 98195
*Corresponding author

Contact: Qiuping.Yu@georgetown.edu, http://orcid.org/0000-0001-7476-866X (QY); zhangyiming@cuhk.edu.cn, http://orcid.org/0000-0003-2922-7877 (YZ); yongpin@uw.edu, http://orcid.org/0000-0001-5841-8387 (Y-PZ)

Received: April 21, 2021
Revised: July 7, 2021
Accepted: August 1, 2021
Published Online in Articles in Advance: June 8, 2022

https://doi.org/10.1287/mnsc.2022.4448

Copyright: © 2022 INFORMS

Abstract. The growing adoption of virtual queues in the service and retail industries has been greatly accelerated in recent times. In collaboration with a major ride-sharing platform, we study how the wait time information (WTI), both its initial magnitude and its subsequent progress over time, impacts customers' abandonment behavior in virtual queues. The study was conducted through a randomized field experiment that included 1,425,745 rides: one-third of the rides received a neutral WTI, one-third received an optimistic WTI shorter than the neutral WTI (hence less frequent updates), and one-third received a pessimistic WTI (hence more frequent updates). The underlying wait time did not vary across the three groups. We find that both the magnitude of the initial WTI and the update frequency of the WTI have a significant impact on customer abandonment. Specifically, when adjusting the initial WTI by one minute, it did not impact customer abandonment. This is because the magnitude effect of the initial WTI is cancelled out by the opposite update-frequency effect. However, when adjusting the WTI by more than one minute, the magnitude effect dominates: when comparing the pessimistic WTI of four minutes with the neutral initial WTI of two minutes, five minutes with three minutes, and eight minutes with five minutes, customers' likelihood to abandon increases by 6.2%, 14.1%, and 19.6%, respectively. Similar but opposite effects are found when comparing the optimistic WTI with the neutral WTI. We discuss how firms can use our findings and insights to design and operate better virtual queues.

History: Accepted by Vishal Gaur, operations management.
Funding: The authors received research funding from the partner platform to help conduct this study.
Supplemental Material: The data files and online appendix are available at https://doi.org/10.1287/mnsc.2022.4448.

Keywords: virtual queue • wait time information • large-scale field experiment • queue abandonment behavior • ridesharing platform

1. Introduction

Americans spend roughly 37 billion hours each year waiting in line (Stone 2012). The emotional cost of waiting is high because of stress, boredom, and a sense of waste, especially when the wait time is highly uncertain (Maister 1984). Providing wait time information has been shown to improve customer experience because it reduces the uncertainty of waiting (Larson 1987, Hui and Tse 1996). It becomes even more important for virtual queues, where customers cannot infer their remaining wait time or progress directly through visual cues. Virtual queues have existed for a long time in call centers, theme parks, and restaurants. The COVID-19 pandemic has accelerated its adoption and growth in other industries such as retailing, daycare, hospitals, and the fast-expanding food delivery (e.g., Doordash and Uber Eats) (Deighton 2020).

Beyond reducing the sense of uncertainty, providing wait time information (WTI) is also an important mechanism virtual queue managers can use to manage customers' behavior, by better facilitating the matching between the demand and system capacity. For example, during low demand period, a virtual queue manager may want to reduce abandonment, but during peak time, it may want to have at least some customers to abandon because of a lack of capacity to reduce the overall wait time for other customers who will be served (Allon et al. 2011, Yu et al. 2018). In this paper, we study how WTI impacts customers' abandonment behavior in virtual queues, which is an important first step toward developing WTI policies that optimize system performance and customer experience.

A longer initial WTI may lead to customers' belief of a longer waiting time, which may in turn cause a

higher initial likelihood of abandonment (Carmon and Kahneman 1996, Webb et al. 2017), especially when there are no cancellation fees as in the context of our study. We refer to this as the *magnitude effect*. After the initial WTI, frequent downward (one-minute) updates of the WTI signal quick progression of customers' waiting, which may induce positive customer affection and reduce customers' likelihood to abandon (Carmon and Kahneman 1996, Batt and Terwiesch 2015, Bolandifar et al. 2019). We refer to this effect as the *update-frequency effect*. Because customers who receive a longer initial WTI also subsequently experience more frequent (one-minute) downward updates, we study the combined magnitude effect and the update-frequency effect.

We partner with a major global ride-sharing platform to study these effects. When there is a shortage of drivers, our partner platform puts newly arrived customers in virtual queues based on the locations of their trip origins. During a span of four weeks (August 23, 2018, to September 20, 2018), we designed and implemented a large-scale field experiment in a major international city that is divided into more than 500 geographically exclusive virtual queues. Once a customer joins a virtual queue, the platform's forecasting algorithm calculates an *estimated wait time* (EWT) and constantly updates it in real time. Because of the stochastic nature of customers' waiting time, the EWT has an intrinsic variation. As such, there is a range of possible values that can be displayed to the customer. In our experiment, we study the impact of different WTI derived from the EWT and its associated variations: one-third of rides were randomly assigned to the control group and received a *neutral* initial WTI based on the platform's EWT and its variation; one-third received an *optimistic* initial WTI that is lower than the control group's (hence less frequent updates); and the last third received a *pessimistic* initial WTI that is higher than the control group's (hence more frequent updates).[1] The underlying wait time did not vary across the three groups.

We collected detailed ride-level information on a total of 1,425,745 rides, including the location of the customer, the distance of the ride, quoted prices, the waiting outcome (abandoned or matched with a driver), and other variables. Importantly, the data include all the WTIs customers received and the exact time stamps of customers' arrival, abandonment (if abandoned), and the matching (if matched). The granularity of the data allows to conduct detailed survival analysis on customers' likelihood to abandon at any given time. This together with the randomized design of our experiment enables clean identification of the real-time impact of WTI on customer abandonment behavior and the underlying mechanisms that drive the impact. Our analysis shows that there is an intricate tradeoff between the magnitude and update-frequency effects. When the pessimistic initial WTI is one minute higher than the neutral initial WTI, the resulting increase in the customers' likelihood to abandon is offset by the opposing impact of more frequent downward updates. Overall, the customers' likelihood to abandon is not affected.[2] When the pessimistic initial WTI exceeds the neutral one by more than one minute, however, the magnitude effect dominates. In particular, when comparing the pessimistic initial WTI of four minutes with the neutral initial WTI of two minutes, five minutes with three minutes, and eight minutes with five minutes, the customers' likelihood to abandon increases by 6.2%, 14.1%, and 19.6%, respectively. When the neutral initial WTI is between 6 and 10 minutes, the corresponding pessimistic initial WTI, which is on average 3.8 minutes longer than the neutral one, increases the customers' likelihood to abandon by 16.4%. Similar but opposite effects are found when comparing the optimistic initial WTI to the neutral ones.

Our findings have implications at the system level. Although a significantly optimistic WTI (at least two minutes less than the neutral WTI) may reduce abandonment rate, it may also lead to bad customer experience when customers get stuck on "the last minute" for too long, resulting in poor customer satisfaction (Carmon and Kahneman 1996). This will in turn hurt the retention and acquisition of customers (Lemon and Verhoef 2016) and thus the growth of the platform in our context. Virtual queue managers who care about customers' waiting experiences, such as our partner platform, should carefully weigh the positive short-term effect on abandonment against the negative long-term effect on customer experience and thus retention and acquisition when considering providing optimistic WTIs. We note that virtual queue managers should at least avoid providing moderately optimistic WTI (one minute less than the neutral WTI) as it does not reduce abandonment but hurts customer experience. In contrast, providing a moderately pessimistic WTI (one minute longer than the neutral WTI) is always beneficial. It does not increase customer abandonment but may pleasantly surprise customers later when they wait less than expected (Kahneman and Tversky 1979). Furthermore, it may also be beneficial to provide a significantly pessimistic WTI (at least two minutes higher) when the congestion level is really high. During such time, virtual queue managers will not be able to serve all the customers because of a lack of capacity; thus, it is best if less-patient customers abandon the system and do so quickly to avoid unnecessary wait times (Allon and Bassamboo 2011, Yu et al. 2018). This benefits both the customers and the service provider.

Our study is related to, but significantly different from, two streams of research. The majority of the

literature on delay information in virtual queues is analytical in nature (Hassin 1986, Guo and Zipkin 2007, Armony et al. 2009, Allon et al. 2011, Hu et al. 2018, Yu et al. 2018, Ibrahim 2018). Recent empirical studies focus on the magnitude effect of the initial WTI (Akşin et al. 2016; Yu et al. 2017, 2020; Dong et al. 2019) and have been conducted in settings of call centers and hospitals. In addition, all these empirical studies use observational data, where the WTI is not truly exogenous but intrinsically linked to the underlying congestion level of the system.[3] As such, it is challenging to cleanly identify the true impact of WTI without certain structural assumptions. Our study contributes to this literature by designing a truly randomized field experiment that also incorporates the update-frequency effect of WTI. In addition, our study broadly contributes to the literature on operational transparency (Buell and Norton 2011) by studying the impact of WTI on customer behavior in virtual queues.

There are studies that explore the impact of wait progress on customers' satisfaction (Carmon and Kahneman 1996, Kumar et al. 1997, Bray 2020), purchasing behavior (Lu et al. 2013), retrospective choices of different services routes (e.g., transportation routes) (Soman and Shi 2003), or abandonment (Batt and Terwiesch 2015, Akşin et al. 2019, Bolandifar et al. 2019, Buell 2020). Most of these studies were conducted using controlled laboratory experiments (Carmon and Kahneman 1996, Kumar et al. 1997, Soman and Shi 2003, Akşin et al. 2019). It is not clear how the results can be generalized to real-world settings. Our paper is most related to the few field studies on customer abandonment decisions using observational data (Batt and Terwiesch 2015, Bolandifar et al. 2019). These studies on abandonment focus on physical queues where customers infer the progress based on visual cues and no WTI was provided to the customers. They collectively show that queue progress inferred from visual cues improves customer experience and reduces customer abandonment. We complement this literature by studying the impact of WTI (its magnitude and update frequency) in *virtual queues* and through a *large-scale randomized field experiment* on a *ride-sharing platform*. Moreover, whereas these studies suffer from major data limitations because of the lack of exact abandonment time of their subjects, our study contains all the relevant timing information that allows us to perform real-time nonparametric and parametric survival analyses on customers' likelihood to abandon at any given time. This in turn allows us to better understand the real-time tradeoffs between the magnitude and the update-frequency effects of the WTI.

To the best of our knowledge, our study is the first large-scale randomized field experiment on both the impact of WTI and customers' abandonment behavior in queues. It is also the first field study that explores the update-frequency effects of the WTI on customers' abandonment behavior. From a methodological perspective, we develop a novel stratified experiment design, where we use customers who share the same EWT on arrival as a stratum and rotate the treatment on the level of day and virtual queue combination. This is effective in addressing the treatment interference issues (customers from one group is impacted by multiple treatments) caused by queuing network externality (where one customer's abandonment decision may impact other customers' decisions through their wait times) and communications among customers on digital platforms; we elaborate on this in Section 2.3.

2. Data and Preliminary Results

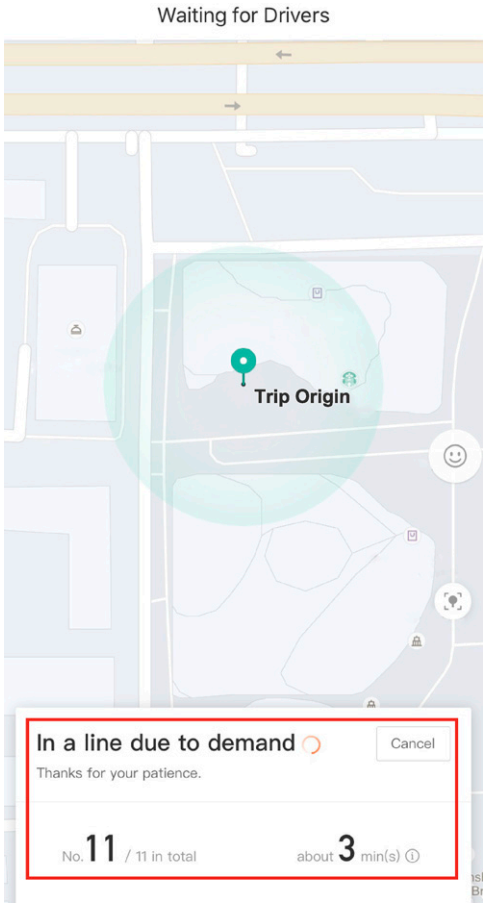
We collected our data from a large-scale field experiment on a major global ride-sharing platform.

2.1. Empirical Context

Ride-sharing companies provide service in thousands of cities across the world and have changed the way people travel. Customers are matched to nearby drivers on the platform without the need for customers to wave their hands on the street. Before the COVID-19 pandemic, there were more than 20 million operating drivers, hundreds of millions of rides happened every month, and the ride-sharing market was valued at $73 billion in 2019 and was growing rapidly (MordorIntelligence 2019). In its early days, our partner platform adopted surge pricing, which its peers such as Uber and Lyft effectively used (Lyft 2019, Uber 2019), only to find poor customer reception. This along with government regulation caused it to limit surge pricing to 1.5 times the regular price and eventually remove the use of surge pricing for most services in 2017. Instead, a virtual queue system was put into place where a large city may be divided into hundreds or thousands of geographically exclusive virtual queues.

After the customer places an order, if there is shortage of drivers, this customer will be put in a virtual queue based on the trip origin until being matched with a driver. Through the app interface, the customer is provided with WTI, the number of customers in the virtual queue, and the current position of this customer in the queue (Figure 1), which are constantly updated by the platform. The queue length information is very straightforward: because customers are mostly served in a first-come, first-served manner within the same virtual queue (with certain exceptions because of drivers' preferences), the platform simply counts the total number of customers. In the current paper, we do not treat the queue length information and focus on studying how the change in WTI can affect customers' abandonment behavior after they join the virtual queue but before they are matched with a driver. Customers can abandon any time during this period without any cancellation fee. As

Figure 1. (Color online) Ride-Sharing App Interface



such, customers' abandonment rate is rather high during this period as we will report in Table 2, which highlights the importance of optimizing WTI to better manage customers' expectation in our context.

The map, on the upper part of the app interface in Figure 1, only shows customers' trip origins and no other information or visual cues are displayed. On the upper right corner of the red rectangle, there is a "Cancel" button to cancel the trip if the customer decides to abandon. Once the customer clicks that "Cancel" button, the app will record the time of abandonment and stop collecting further information about the ride request. Our partner platform provides multiple types of services to customers, which include rides with standard cars, rides with luxury cars, rides with adapted cars for children and those with disabilities, taxi, and carpooling services. In our experiment, we focus on private rides with standard cars, which is the most basic and commonly used service on the platform.

2.2. WTI

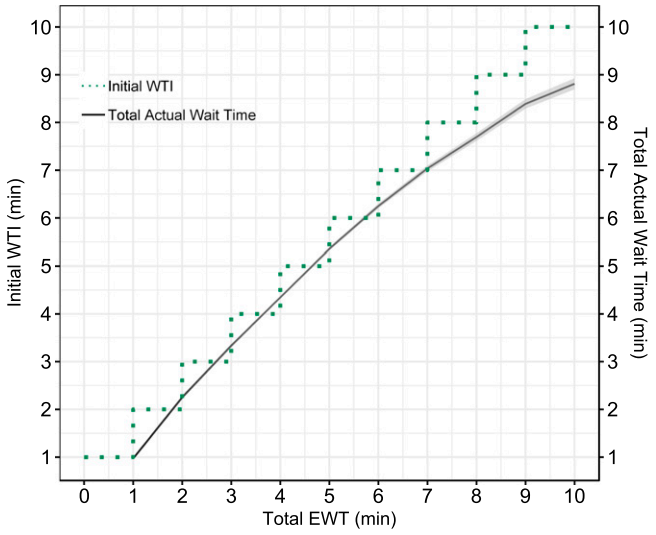
To predict customers' wait time, the platform uses a machine-learning algorithm on features such as the current numbers of available drivers and customers waiting on the platform, time and location of the order, weather, and historical delay information. Upon each customer's arrival in the virtual queue, the platform makes a prediction about the total EWT. Subsequently, although the customer waits in the virtual queue, the platform calculates the remaining EWT every three to five seconds. Over time, the EWT does not always move smoothly and may even increase because of unexpected events; thus, the platform does not show them to customers directly. Rather, the platform applies a smoothing algorithm to the subsequent EWT (not the initial total EWT) to ensure it is nonincreasing and relatively smooth over time. We elaborate on how the smoothing algorithm works in Section 2.3. Finally, the platform *rounds up* the smoothed estimates to the nearest integer minute and displays this to the customer. We refer to the information the customer sees upon arrival as the *initial WTI* and the updated ones as the *subsequent WTIs*. Figure 2 shows that the initial WTI closely aligns with the average total offered wait time of those customers who received the information and were matched with a driver.[4] Because of the integral nature of the WTI, customers will only get updates when their WTI drops to a lower integer minute. That is, for a customer who receives an initial WTI of m minutes, the customer can get up to $m-1$ updates if this customer waits on the platform sufficiently long and keeps the app active. The platform only records the WTI displayed to customers while the app is active.

Although customers' wait time is inherently uncertain, prior to our study, our partner platform displayed WTI based on a point estimator, which we refer to as the *neutral EWT*. Our discussion in this section thus far has followed the platform's approach prior to our study and focused on just the neutral EWT. In the next section, we will discuss the uncertainty in the EWT and how we used that to design our experiment.

2.3. Experiment Design

Our large-scale randomized field experiment was run on the partner platform in a major international city for four weeks from August 23, 2018, to September 20, 2018. During this time, our experiment was the only experiment being run on the platform; this was done to avoid any interference from other experiments. In our experiment, we consider three treatment conditions: (1) the control group where we provide the neutral initial WTI based on the aforementioned neutral total EWT, denoted as $total_EWT_{neu}$, (2) the high treatment group where customers receive a pessimistic initial WTI higher than those in the control group and, thus, more subsequent updates, and (3) the low treatment group where customers receive an optimistic initial WTI lower than those in the control group and thus fewer subsequent updates. Whereas the neutral WTI comes directly from the neutral EWT currently

Figure 2. (Color online) Total EWT vs. Total Actual Wait Time



calculated by the platform's algorithm, the pessimistic and optimistic WTI values are new and are based on both the neutral EWT and its inherent variation.

Consider a customer whose $total_EWT_{neu}$ is x_m seconds, which, when rounded up, is m minutes. If this customer is randomly assigned to the control group, this customer will receive an initial neutral WTI of m minutes. If the customer is assigned to a high (low) treatment group, this customer will receive a pessimistic (optimistic) initial WTI picked from an interval of likely total EWT values constructed as follows: First, we select customers who received an initial WTI of m minutes and were matched with a driver in the month prior to our experiment.[5] Second, we compute the standard deviation of these customers' actual wait times, the 20% quantile, and the 80% quantile in seconds, which we denote as δ_m, Quantile20_m, and Quantile80_m, respectively. We also compute the average $total_EWT_{neu}$ of these customers, here denoted as $\bar{x}_m$. Finally, we let the interval of likely total EWT values be $[total_EWT_{low}, total_EWT_{high}]$, where $total_EWT_{low} = \max\{\text{Quantile20}_m, \bar{x}_m - \delta_m\}$ and $total_EWT_{high} = \min\{\text{Quantile80}_m, \bar{x}_m + \delta_m\}$, which are both measured in seconds. Essentially, we construct the interval to span from the 20% quantile to 80% quantile of the actual wait time and bound it by one standard deviation from the mean. The platform then rounds the lower and upper ends of the interval, $total_EWT_{low}$ and $total_EWT_{high}$, up to the nearest integer minute and uses them as the optimistic initial WTI and pessimistic initial WTI, respectively, to show to the low and high treatment groups.

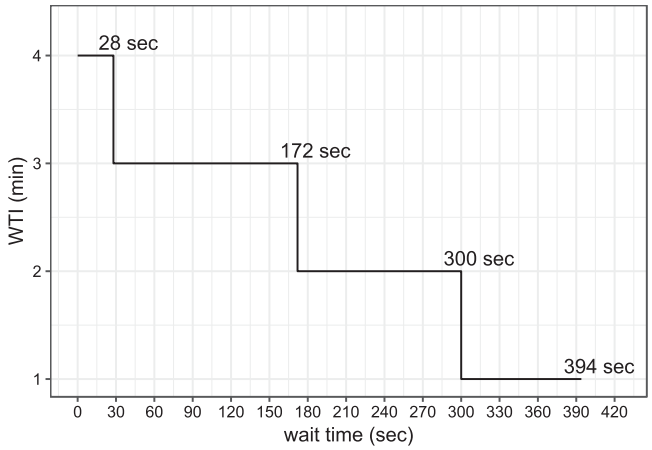
Table 1 shows the initial WTI provided to customers in the control, low, and high treatment groups for any given $total_EWT_{neu}$. Using a combination of standard deviations and quantiles to construct the treatments ensures that WTI in the high (low) treatment is substantially higher (lower) than that in the control group while still within a reasonable range acceptable to the platform.

In addition to the initial WTI, the subsequent WTI(s) customers receive are also treated. Figure 3 illustrates the WTI a customer in the control group received over time while this customer was waiting in the virtual

Table 1. Construction of Treatments

	Interval of likely total EWT (s)		Initial WTI (min)		
$total_EWT_{neu}$ (sec)	[*total_EWT_low*,	*total_EWT_high*]	Control	Low	High
(1, 60]	4	100	1	1	2
(60, 120]	47	185	2	1	4
(120, 179]	90	269	3	2	5
(180, 240]	132	353	4	3	6
(240, 300]	175	437	5	3	8
(300, 360]	218	521	6	4	9
(360, 420]	260	606	7	5	11
(420, 480]	303	690	8	6	12
(480, 540]	346	774	9	6	13
(540, 600]	388	858	10	7	15

Figure 3. WTI a Customer Received While Waiting in the Virtual Queue



queue. This customer received an initial WTI of four minutes on arrival at time 0. The WTI dropped to three, two, and one minutes at 28, 172, and 300 seconds, respectively. This customer was matched with a driver after waiting for 394 seconds in total. We define each customer's collection of downward-updating time epochs as the *update timings*. Customers' subsequent WTI(s) are determined by a combination of the EWT (from the platform's forecasting algorithm), the platform's smoothing algorithm, and our treatments. Consider a customer who is assigned to the control group. Let the time of the arrival to the virtual queue be time 0. The WTI this customer sees in the app at any time $t \geq 0$ is generated by the following steps:

1. The platform generates an estimate of the customer's remaining wait time in seconds, using its forecasting algorithm but *prior to* using the smoothing algorithm or the treatment. Denote this by EWT_t^{prior}. By definition, $EWT_0^{prior} = total_EWT_{neu}$.

2. The platform then applies its smoothing algorithm, whose purpose is to make sure that the WTIs customers receive never increase while they wait in the queue. In the control group, the smoothing algorithm uses the initial total estimated wait time ($total_EWT_{neu}$) and the current estimated remaining wait time (EWT_t^{prior}) as inputs, and generates $EWT_t^{post,control}$ (in seconds), as follows:

$$EWT_t^{post,control} = total_EWT_{neu} - f(total_EWT_{neu}, EWT_t^{prior}), \tag{1}$$

where $f(\cdot)$ is the nonnegative smoothing function and $f(total_EWT_{neu}, EWT_0^{prior}) = 0$ by definition.

3. Finally, the platform rounds ($EWT_t^{post,control}/60$) up to the nearest minutes, which we denote as $WTI_t^{control}$ and display it to the customers in the control group.

For customers who are assigned to the treatment groups, their EWT_t^{prior} in Step 1 at any time t is the same as that for the control group. However, the treatment they receive at time 0 means that their total EWT is transformed to $total_EWT_{high}$ or $total_EWT_{low}$, instead of $total_EWT_{neu}$. Once the same smoothing function $f(\cdot)$ (in Step 2) is applied, we have the following equations.

$$EWT_t^{post,high} = total_EWT_{high} - f(total_EWT_{neu}, EWT_t^{prior}). \tag{2}$$

$$EWT_t^{post,low} = total_EWT_{low} - f(total_EWT_{neu}, EWT_t^{prior}). \tag{3}$$

The platform will then apply the same Step 3 to round ($EWT_t^{post,high}/60$) and ($EWT_t^{post,low}/60$) up to the nearest minutes and display them to the customers in the respective treatment group. We denote them as WTI_t^{high} and WTI_t^{low}.

In our experiment, we let the smoothing function $f(\cdot)$ in (2) and (3) use the same inputs as those in the control group in (1). As such, the difference in the estimated remaining wait time *after* the smoothing algorithm between the treatment ($EWT_t^{post,high/low}$) and control ($EWT_t^{post,control}$) is only from their difference in the initial total EWT ($total_EWT_{high/low}$ versus $total_EWT_{neu}$). Given the integrality of the WTI, the update timings of the WTIs depend on the residual of the initial total EWT *after* the treatment (in seconds) divided by 60. As a result, when $total_EWT_{neu}$ and its corresponding $total_EWT_{high/low}$ have the same residual divided by 60, customers in the treatment and control groups will share (statistically) similar update timings.[6] This is an important feature of our experiment design, which enables a clean identification of the magnitude and update-frequency effects of the WTI as we will elaborate further in Section 3.2.

There are alternative ways to construct the EWT values for the low and high treatment groups. For example, instead of solely based on historical information,

the treatment groups' EWT can be based on the lower and upper bounds of the confidence interval of the control group's EWT *at that time*. Our approach, however, has the advantages of being computationally efficient (just look up a static Table 1) and less noisy (we will always use the same $total_EWT_{low}$ and $total_EWT_{high}$ in Table 1 for customers with the same $total_EWT_{neu}$ on arrival).

We next explain how we randomly assign rides into different treatment groups. We want to avoid assigning customers who are waiting next to each other to different treatment groups because customers may talk to each other and question the credibility of the different WTI they receive.[7] As such, we assign the treatments at the level of day and virtual queue combination (recall that there are more than 500 virtual queues in the city we run our field experiment in, and the experiment lasted 28 days). For any given day during our experiment, we randomly assign about a third of the virtual queue (meaning all its rides) to each of the three treatment conditions (control, low, or high) based on a hash function of both the virtual queue IDs and the dates. To later tease out the virtual-queue and the day-of-week fixed effect, we repeat this randomized assignment on each day for 28 days.[8] The realized assignment of treatments to the virtual queues vary across different days in a randomized fashion.

Our design is similar to the stratified randomization design (Imbens and Rubin 2015), where customers are stratified based on their $total_EWT_{neu}$ (after rounding up to the nearest integer minute). Within each stratum, unlike the traditional stratified randomized experiments where treatments are assigned randomly at the individual customer level, we do so at the level of virtual queues and day combination to avoid a bad customer experience and potential treatment interference issues. Our analysis shows that the randomization works well: (1) within each stratum, samples across the control and treatment groups are well balanced in terms of their pretreatment variables such as trip distances, quoted prices, and percentage of rides stemming from experienced customers; (2) there is sufficient variability in treatments within the same virtual queue and the same day of the week, allowing us to tease out the potential virtual-queue and day-of-week fixed effects; and (3) given that customers are stratified by their $total_EWT_{neu}$, customer-predicted wait time until being matched (if they never abandon) is statistically the same across the three treatment conditions. When predicting the $total_EWT_{neu}$ for a customer, the platform's algorithm accounts for the other customers' abandonment behavior through historical data along with concurrent features. For more details on the effectiveness of our randomization process, please refer to Online Appendix A.

2.4. Data Descriptions

Our data include detailed information for all the individual rides that joined the virtual queues in the focal city during our experiment. For each ride, we observe the time when the customer placed the order, the pickup location characterized by the virtual queue, the trip distance, the quoted price, all the EWTs and WTIs, the customer's actual total wait time in the virtual queue, and whether the customers abandoned or waited until they were matched with a driver. Our data set has three unique features: (1) for each ride, we observe not only the initial WTI customers received on their arrival but all the subsequent WTI updates customers received while they were waiting in the virtual queue, as illustrated in Figure 3; (2) we observe the exact timings when customers abandoned or were matched with a driver and when they received the WTIs; and (3) customers were randomly assigned to each treatment, and hence, we could observe the treatment group each ride was assigned to.

As mentioned earlier, besides the WTI, the platform also provides information on the queue length and queue position of the customers, all of which were updated in real time. Because no intervention was imposed on the queue length and queue position information during our experiment, the queue length the customer would observe upon arrival, conditional on the same total EWT on arrival, is similar across the control and treatment groups (see Online Appendix F).

In our analysis, we remove the rides whose total actual wait time is in the bottom 2.5% (one second or less) and in the top 2.5% (24 minutes or more) as these customers might be outliers in their patience time.[9] As there are not sufficient data for $total_EWT_{neu} > 10$ minutes, we focus our analysis on customers whose $total_EWT_{neu}$ is within 10 minutes, which accounts for 92.2% of the rides. Finally, we focus on the rides placed between 6:00 a.m. to 11:59 p.m. because only 0.7% of the orders were placed outside this time period.[10] After the filtering, 1,425,745 rides remain in our data.

In Table 2, we present the numbers of rides, the abandonment rate, and the average wait time of the abandoned customers for each stratum ($total_EWT_{neu}$). We observe that customers' abandonment rate in the high treatment group is higher than the control group when the $total_EWT_{neu}$ is high. When abandoning, customers seem to abandon earlier in the high treatment group, regardless of the $total_EWT_{neu}$. The opposite holds for customers in the low treatment group. We caution that although the table provides some preliminary information on behavior, we cannot infer the treatment effects by simply comparing customers' overall abandonment rate and wait time across the control and treatment groups. There are two main reasons for this: (1) although the assignment of the treatment is randomized and the samples across the three treatment conditions are well balanced in their

Table 2. Summary Statistics Across the Three Treatment Conditions

	Number of rides			Abandonment rate			Total actual wait time: Abandoned (s)		
$Total_EWT_{neu}$ (min)	Low	Control	High	Low	Control	High	Low	Control	High
1	205,845	220,271	225,534	10.3%	10.2%	10.1%	58	58	54
2	75,834	79,627	83,518	26.2%	26.0%	25.6%	82	80	75
3	47,255	50,807	53,367	34.0%	34.1%	34.3%	98	95	88
4	34,521	37,582	38,921	39.0%	40.0%	40.4%	116	107	102
5	25,762	27,931	29,418	43.6%	44.3%	45.1%	125	119	112
6	18,987	19,761	21,743	47.1%	47.9%	49.7%	136	130	121
7	14,221	14,692	16,381	49.9%	50.4%	52.7%	142	138	131
8	11,357	11,307	12,786	52.0%	54.3%	54.7%	152	145	137
9	8,400	8,904	9,667	55.0%	56.2%	57.9%	155	153	139
10	6,596	7,181	7,569	56.3%	58.3%	59.5%	164	156	149

pretreatment variables (e.g., trip distance, quoted price, and initial queue length), the potential virtual-queue and day-of-week fixed effects remain to be teased out; and (2) it is difficult to infer the mechanisms of the treatment effects only based on the overall statistics because multiple features associated with the treatment, such as magnitude, update frequency, and update timings, can all play a role.

3. Model and Identification Strategy

We next present the model and identification strategy for the combined magnitude and update-frequency effects. In our context, when customers in the high (low) group receive pessimistic (optimistic) initial WTIs, they also experience more (less) frequent updates subsequently. On the one hand, a pessimistic initial WTI indicates a longer wait time, making customers more likely to abandon (Carmon and Kahneman 1996, Webb et al. 2017). On the other hand, more frequent updates may improve customers' perceived progress of their waiting and thus may reduce customers' likelihood to abandon (Carmon and Kahneman 1996, Munichor and Rafaeli 2007). As such, it is important that we study the tradeoff between the magnitude effect and update-frequency effect.

3.1. Model

We define the maximum time that a customer is willing to wait for service before abandoning as the *patience time*. It is right censored for customers who waited until being matched with a driver (because they were matched before their patience time ran out). As such, we propose using a proportional hazard rate model that accounts for such censoring issues and characterizes customers' patience time as an outcome of a base hazard rate modified by a collection of variables. Its specific form is as follows:

$$h_i(t) = h_i^0(t)\exp(\boldsymbol{\beta}'\boldsymbol{X_i}). \tag{4}$$

Note that $h_i(t)$ is the hazard rate of customer i's patience time at time t, which is the probability that the customer will abandon at time t given that this customer is still waiting in the queue at the time. $h_i^0(t)$ is the base hazard rate that captures the customer's intrinsic likelihood to abandon and may vary over time as the customer waits in the queue. We use the Weibull distribution to characterize customers' base hazard rates and include a gamma distributed frailty term to characterize customers' heterogeneity (in their intrinsic patience time).[11] $\boldsymbol{X_i}$ is the vector of variables that may impact customer i's likelihood to abandon, whereas $\boldsymbol{\beta}$ is a vector of coefficients representing the magnitude of the corresponding effect.

To characterize the treatment effects, we let the $high_i$ and low_i be two dummy variables, which equal one when the ride was assigned to the high and low treatment, respectively, and zero otherwise. Because we will estimate the treatment effect for each stratum (categorized based on $total_EWT_{neu}$), we discretize the $total_EWT_{neu}$ with an interval of one minute and let $total_EWT_{neu}^n$ be the dummy variable, which equals one when the $total_EWT_{neu}$ is in the nth interval, with $n = 1, 2, \cdots, 10$. We then include the following interaction terms in $\boldsymbol{X_i}$: $high_i \times total_EWT_{neu}^n$ and $low_i \times total_EWT_{neu}^n$ for each n. In addition, we include the following variables as controls: (1) the collection of dummy variables $total_EWT_{neu}^n$ for $n = 1, 2, \cdots, 10$, which represent the fixed effects of each stratum, (2) trip distance denoted as $Distance_i$, (3) quoted price denoted as $QuotedPrice_i$, (4) virtual-queue fixed effects, which account for any time-unvarying characteristics associated with each virtual queue, such as its location, and (5) the time-of-day and day-of-week effects. Recall that in our stratified experiment design, the random assignment of treatments is done at the level of virtual queue and day combination. As such, to estimate the causal treatment effects besides the stratum fixed effect, it is important to control for the virtual-queue and day-of-week fixed effects.[12] We include the other controls to improve estimation efficiency.[13] Recall from Section 2.4 that we did not impose an intervention on the queue length and queue position information. As such, when estimating the treatment effect, we did not include queue-position related controls in our model.[14]

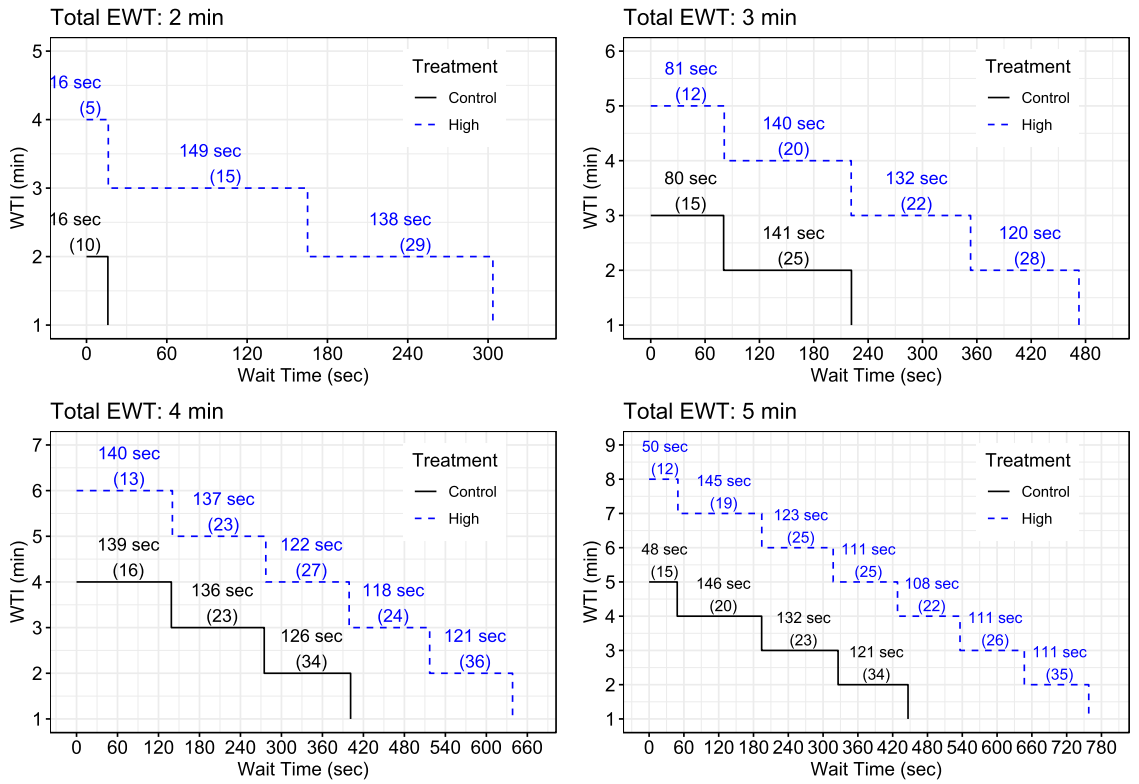
3.2. Identification

The treatment and the control groups are different not only in their initial WTI and the subsequent update frequencies, but also in when the updates happen. To cleanly identify the combined magnitude and update frequency effects, we shall tease out the effect of update timings. As we described in Section 2.3, the update timings depend on the residuals of the total EWT after the treatment (i.e., $total_EWT_{neu}$, $total_EWT_{high}$ and $total_EWT_{low}$). As such, to tease out the update-timing effect, we match the update timings across the treatment and control groups based on these residuals, which are exogenous to the focal customers' abandonment decisions.

For example, all low treatment customers whose $total_EWT_{low}$ is 132 seconds (initial WTI = 3 minutes, residual = 12 seconds) share similar update timings with control group customers whose $total_EWT_{neu}$ is 192 seconds (initial WTI = 4 minutes, residual = 12 seconds). Therefore, we can match all the low-treatment customers who have an initial WTI of 3 minutes (all of whom have the same $total_EWT_{low}$ of 132 seconds) with control customers who have an initial WTI of 4 minutes (but only those who have a $total_EWT_{neu}$ of 192 seconds). Because the latter has a much more limited sample size, we make two adjustments to balance the number of observations and the matching precision in the update timings. (1) We choose customers in the control group whose $total_EWT_{neu}$ is within the interval [187, 197] instead of exactly 192 seconds. These customers should have similar update timings to the ones experienced by the customers in the corresponding low treatment group. (2) We also match the total EWT *before* the treatment ($total_EWT_{neu}$) across the control and treatment groups. With the previous example, instead of using all the customers in the low treatment group, we only use those customers whose corresponding $total_EWT_{neu}$ is within the interval of [187, 197]. The same matching procedure is applied to the high treatment group as well. For example, all the high treatment customers with $total_EWT_{high} = 353$ seconds (initial WTI = 6 minutes) share similar update timings with the control group customers with $total_EWT_{neu} = 233$ seconds (initial WTI = 4 minutes). Therefore, in the matched data, we only use customers with $total_EWT_{neu} \in [228, 238]$ seconds from both the control and high treatment groups. This matched interval in the control group is different from that matched with the low-treatment customers ([187, 197]). As such, we will estimate the high treatment effect and low treatment effect relative to the control group separately.

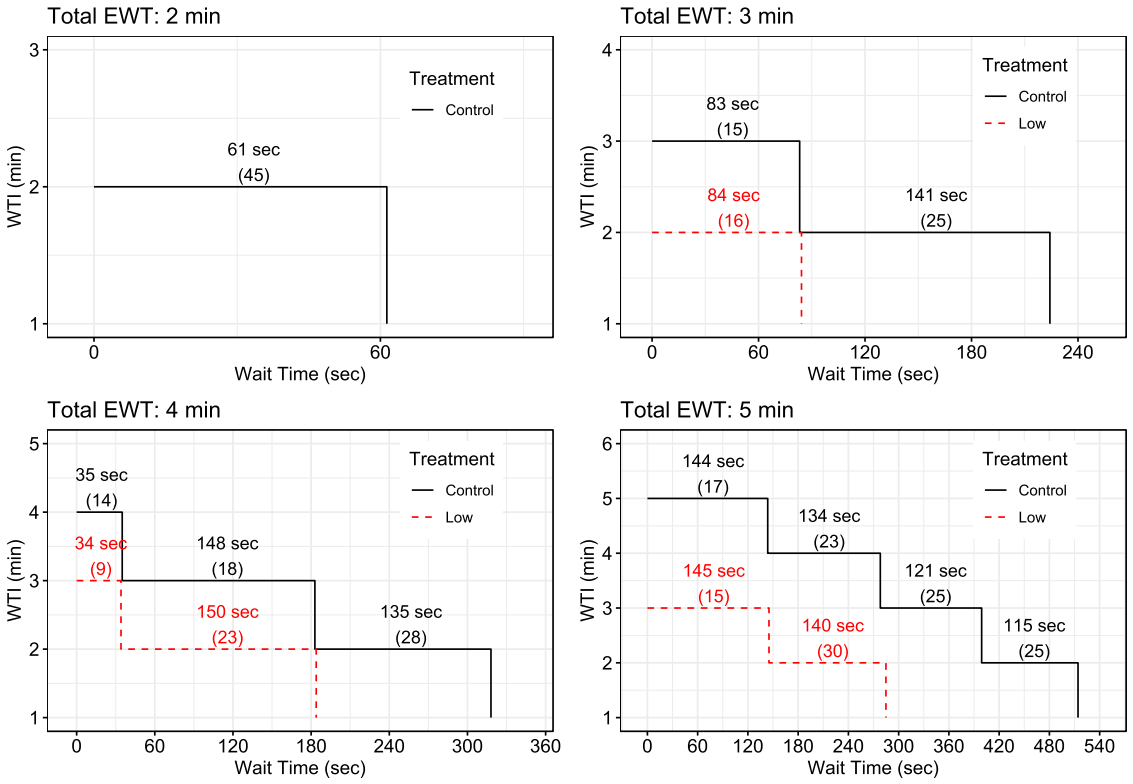
In the matched data, the update timings of the WTIs become *statistically* the same across the control and treatment groups (Figures 4 and 5). Thus, any difference in abandonment behavior found by the model

Figure 4. (Color online) WTI Progress for the Control and High Treatment Groups After Matching Procedure



Note. We provide in seconds the average time each WTI was displayed and the corresponding standard deviations (in parentheses) of those customers who *did not abandon before* the WTI dropped to the next minute.

Figure 5. (Color online) WTI Progress for the Control and the Low Treatment Groups After Matching Procedure



Note. We provide in seconds the average time each WTI was displayed and the corresponding standard deviations (in parentheses) of those customers who *did not abandon before* the WTI dropped to the next minute.

in Section 3.1 can be attributed to just the combined magnitude and update-frequency effects.[15]

4. Results

We present our estimation results on the combined magnitude and update-frequency effects induced by the low and high treatments in columns (2) and (5) of Table 3, respectively. We observe that when the optimistic or pessimistic WTI deviate from the neutral WTI by one minute, it does not impact the customers' overall likelihood to abandon. As we will show later in this section and Online Appendix B, both the

Table 3. Estimation Results for the Combined Magnitude and Update-Frequency Effects

High and control			Low and control		
(1) Covariates	(2) Overall	(3) [0, 30]	(4) Covariates	(5) Overall	(6) [0, 30]
High × $total_EWT^n_{neu}$ (initial WTI: n + Adjustment)			Low × $total_EWT^n_{neu}$ (initial WTI: n − Adjustment)		
n = 1: 1 + Δ 1	0.001	0.067***	n = 1: 1−Δ 0	−0.001	−0.004
n = 2: 2 + Δ 2	0.060**	0.141***	n = 2: 2−Δ 1	−0.018	−0.068***
n = 3: 3 + Δ 2	0.132***	0.139***	n = 3: 3−Δ 1	−0.044	−0.038
n = 4: 4 + Δ 2	0.044	0.108**	n = 4: 4−Δ 1	−0.031	−0.023
n = 5: 5 + Δ 3	0.179***	0.149***	n = 5: 5−Δ 2	−0.115**	−0.260***
n = 6: 6 + Δ 3			n = 6: 6−Δ 2		
n = 7: 7 + Δ 4			n = 7: 7−Δ 2		
n = 8: 8 + Δ 4	0.152***	0.155***	n = 8: 8−Δ 2	−0.106***	−0.117***
n = 9: 9 + Δ 4			n = 9: 9−Δ 3		
n = 10: 10 + Δ 5			n = 10: 10−Δ 3		
$total_EWT^n_{neu}$ fixed effects	Yes	Yes	$total_EWT^n_{neu}$ fixed effects	Yes	Yes
Time fixed effects	Yes	Yes	Time fixed effects	Yes	Yes
Virtual-queue fixed effects	Yes	Yes	Virtual-queue fixed effects	Yes	Yes
Other controls	Yes	Yes	Other controls	Yes	Yes
Number of rides	548,285	548,285	Number of rides	644,718	644,718

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

magnitude effect and update-frequency effect are sizable and significant by themselves. Hence, the insignificant overall impact indicates that the opposing effects of a higher initial WTI and more frequent subsequent updates have offset each other.

When the magnitude of the initial WTI differs by two or more minutes, however, the magnitude effect dominates the update-frequency effect. Specifically, when comparing the pessimistic initial WTI of four minutes with the neutral initial WTI of two minutes, five minutes to three minutes, and eight minutes to five minutes, the customers' likelihood to abandon increases by 6.2%,[16] 14.1%, and 19.6%, respectively. When the neutral initial WTI is between 6 and 10 minutes, the corresponding pessimistic WTI, which is higher by 3.8 minutes on average, increases customers' likelihood to abandon by 16.4%. We observe similar results in the opposite direction when comparing optimistic WTIs with neutral WTIs. In particular, comparing the optimistic initial WTI of three minutes to the neutral one of five minutes, customers' likelihood to abandon decreases by 10.9%. In addition, when the neutral initial WTI is between 6 and 10 minutes, providing an optimistic initial WTI that is on average 2.3 minutes shorter, reduces customers' likelihood to abandon by 10.1%. We observe that both the congestion level ($total_EWT_{neu}$) and the treatment size can moderate the treatment effect: the treatment effect may increase with the treatment size and, conditional on the same treatment size, the treatment effect may slightly decrease with the congestion level. However, because of the lack of variability in treatment size and congestion level within the range we consider, we cannot formally quantify their moderating effect.

Columns (2) and (5) in Table 3 estimate the combined magnitude and update-frequency effects during the customers' entire waiting episodes. To demonstrate the existence and size of the magnitude effect alone, we zero in on the earlier stage of customers' waiting episode (e.g., the first 30 seconds on customers' arrival). This is because, although the update frequencies differ between the control and the treatment groups, the update timings are statistically the same within each stratum until the WTI drops to one minute in one of the groups (thanks to the matching procedure described in Section 3.2.). Thus, the update-frequency effect is unlikely to kick in within the first 30 seconds, and we can attribute any difference in the abandonment behavior to only the magnitude effect. As such, we conduct the same analysis as shown in columns (2) and (5) of Table 3 but only on the customers' abandonment behavior within the first 30 seconds after arrival. We present the estimation results for the high and low treatments in columns (3) and (6) of Table 3, respectively. The estimation results in column (3) show that when the initial WTI increases from one to two minutes, customers becomes 6.9% more likely to abandon within the first 30 seconds. This magnitude effect increases as the size of treatment in the initial WTI increases. Conditional on the size of the treatment, the magnitude effect seems to decrease with the congestion level ($total_EWT_{neu}$). We observe similar but opposite results when the initial WTI decreases, as shown in column (6) of Table 3.

To demonstrate the existence and size of the update-frequency effect alone, we perform additional time-varying nonparametric and parametric survival analyses in Online Appendix B, where we find a significant and sizable reduction in customers' likelihood to abandon around the times of downward WTI updates. Such an effect persists for some time after the updates but diminishes eventually. To further demonstrate the economic significance of the combined magnitude and update-frequency effects of the WTI, we present additional summary statistics of the matched data in Online Appendix H.1.

In summary, we show that the magnitude effect, happening at the beginning of a customer's waiting episode, can be sizable and significant even when the initial WTI differs by only one minute. The update-frequency effect, triggered later in the episode, can be equally sizable but in the opposite direction. When considering customers' entire waiting episode, these two effects can offset each other if the WTI difference is moderate, rending the total effect insignificant. If the WTI difference is larger, the magnitude effect dominates the update frequency effect. We present clear visual demonstrations of the real-time tradeoff between the magnitude and update frequency effect in Online Appendix B.

5. Conclusion

In the present paper, we report on a large-scale field experiment that we ran on a ride-sharing platform to understand how customer abandonment behavior is influenced by the wait time information provided on their phone app. We characterize how the magnitude of the initial wait time information, and its subsequent update frequency, affect customers' likelihood to abandon. These individual-level discoveries have also led to practice-changing insights on a system level to help improve service for the general public during a supply shortage. For example, a moderately more pessimistic estimate of the wait time followed by more frequent downward updates could help the platform to improve customer wait experiences by reducing unnecessary wait times without increasing customer abandonment. To further explore these findings, we are also working with our partner platform on an experiment to provide range estimates rather than the current point estimates. We expect a properly designed range, higher/lower or wider/narrower all still covering the mean estimate,

can be more flexible and effective in improving customer experience and system performance.

We believe our study is both important and timely. As the economy goes increasingly digital in the post–COVID-19 world, many businesses have to rethink how they interact with customers. Virtual queue systems and their information provision policies are bound to be an important part of such a digital transformation. The present paper is but the first step in that direction; many interesting topics remain for future research. Our field experiment was conducted on a ride-sharing platform. Although it represents an important and novel industry, future studies should test the generalizability of our results to other industries, such as the retail industry. Moreover, in our study, we focus on causally identifying and quantifying the impact of the WTI and its updates on customers' abandonment behavior. We have identified two major mechanisms (the magnitude effect and the update-frequency effect) both nonparametrically and parametrically. It may be interesting to explore other effects of WTI in future studies. For example, there could be a magnitude-based expectation effect induced by the initial WTI. Specifically, longer initial WTI may set an expectation in customers' mind about their total wait time, which makes them less likely to be disappointed later and may in turn cause lower likelihood of abandonment (Kahneman and Tversky 1979). Yu et al. (2020) show the existence of such an effect in the call center context where customers are provided with WTI on their arrival but are not informed about their remaining wait as they progress in the queue. It would be interesting to examine the magnitude-based expectation effect when customers receive multiple WTIs over time.

In this paper, we control for the update timing effect by using a nonparametric matching procedure, but it would be interesting to isolate and quantify the update timing effect by itself: What if one group receives a faster update than the other group? Finally, we note customer learning in our experiment is limited because of the relative short study horizon and the fact that customers may receive different treatments on different days. Additionally, the limited learning (if exists) did not impact how customers respond to the treatment in our experiment (see Online Appendix D). If the platform adopts one policy (optimistic or pessimistic) and uses it consistently over a long period of time, however, customers may learn, update their beliefs about the WTI, and adjust their abandonment behavior accordingly. This represents a very important and promising future research direction.

Endnotes

[1] We emphasize that we did not provide false information to the customers. Given the inherent uncertainty in customers' waiting time, we simply chose WTI that were near the upper end, middle, and lower end of the range of likely estimates based on the platform's algorithm.

[2] The average wait time before being matched with a driver among customer who joined the virtual queues is about 130 seconds. As such, increasing the WTI by one minute is not a small adjustment in our context.

[3] Yu et al. (2020) is an exception in their conducting a *quasi-randomized* field experiment to study customer reference-dependent abandonment behavior when they are provided with delay announcements compared with the case when they are not. However, the announcements provided by the call center turned out to be inaccurate and the authors show that customers completely ignore the announcements when making abandonment decisions. They complement their study with additional observational data from a different call center where announcements are generally accurate.

[4] Offered wait time is the wait time before a customer is matched with a driver if this customer never abandons the platform. The offered wait time of the matched customers is the same as their actual wait time. For abandoned customers, however, their offered wait time is right censored, as they abandoned before they could have been matched with a driver. As such, when computing the average offered waiting, we only use the actual wait time of matched customers. This is also the approach the platform takes when training their forecasting algorithm.

[5] We use the data in the month prior to our experiment to construct our treatments. We, however, do not have access or permission to use ride-level data prior to our experiment for analysis.

[6] For each individual customer, in addition to the residual of their total EWT, their update timings are also determined by the real-time dynamic supply/demand condition. As such, for customers who share the same residual of the total EWT (e.g., 12 seconds), their update timings are random, but are independently and identically distributed. For customers who have different residuals, their update-timings will be statistically different.

[7] This is possible even in a virtual queue, for example, when a group of people from the same office request rides together at the end of their work shift.

[8] We reassign the treatments at 3 am each day to minimize the interference from the treatment transition because there are few customers (if any) around that time.

[9] Our results are consistent with or without removing these outliers; see Table G.1 in Online Appendix G.

[10] Our results are consistent with or without removing customers from 12:00 am to 5:59 am; see Table G.1 in Online Appendix G.

[11] Our estimation results are robust to the parametric assumptions on the base hazard rate; see Online Appendix E.

[12] We also repeat the same analysis but with the fixed effect for the combination of virtual queue and day of week. Our estimation results continue to hold both qualitatively and quantitatively (see Table G.2 in Online Appendix G).

[13] We have conducted the analysis where we only include the fixed effects of the stratum ($total_EWT_{neu}$), virtual queue, and day-of-week as controls but not the other variables; our estimation results are almost identical to our main model as expected. When we only control for the stratum fixed effect, results are generally consistent qualitatively with our main model, with aq slight difference in magnitude and statistic significance. This indicates that, we shall include the virtual queue and the day-of-week fixed effects as we argued to tease out these (although mostly minor) effects. We provide more details about the previous analysis and results in Table G.3 of Online Appendix G.

[14] We have conducted the same analysis but controlling for the initial queue length on customers' arrival and their queue position

updates. Our main insights continue to hold both qualitatively and quantitatively. In addition, we have considered models where we include the potential moderating effect of the initial queue length on our treatment effects. Such effects are generally insignificant statistically at the 10% level. We refer to Online Appendix F for more details about the previous analysis.

[15] With our matching procedure, the estimated treatment effect is under a fixed residual, and the treatment effect may be different under a different residual or update timing pattern (i.e., the potential moderating effect of update timings). Although such moderating effects may exist, we believe it will not impact our main insights, which we further discuss in Online Appendix C.

[16] In Table 3, the coefficient for the high treatment is 0.060 when the initial WTI is four minutes in the high treatment group. This means that customers' likelihood to abandon increases by 6.2% (= $|\exp(0.060) - 1|$) when being assigned to the high treatment group compared with the control group. A similar logic applies to the interpretation of other focal coefficients in our model.

References

Akşin Z, Ata B, Emadi SM, Su C-L (2016) Impact of delay announcements in call centers: An empirical approach. *Oper. Res.* 65(1):242–265.

Akşin Z, Gencer B, Gunes ED (2019) How observed queue length and service times drive queue behavior in the laboratory. Preprint, submitted June 1, https://dx.doi.org/10.2139/ssrn.3387077.

Allon G, Bassamboo A (2011) The impact of delaying the delay announcements. *Oper. Res.* 59(5):1198–1210.

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

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

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Sci.* 61(1):39–59.

Bolandifar E, DeHoratius N, Olsen T, Wiler J (2019) An empirical study of the behavior of patients who leave the emergency department without being seen. *J. Oper. Management* 65(5):430–446.

Bray RL (2020) Operational transparency: Showing when work gets done. *Manufacturing Service Oper. Management*.

Buell RW (2020) Last-place aversion in queues. *Management Sci.* 67(3):1430–1452.

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

Carmon Z, Kahneman D (1996) *The Experienced Utility of Queuing: Real Time Affect and Retrospective Evaluations of Simulated Queues* (Duke University, Durham, NC).

Deighton K (2020) Your child will be ready soon: Day cares and other businesses use restaurant-style tech to fight long lines. *Wall Street Journal*, http://www.wsj.com/articles/your-child-will-be-ready-soon-day-cares-and-other-businesses-use-restaurant-style-tech-to-fight-long-lines-11593597601.

Dong J, Yom-Tov E, Yom-Tov GB (2019) The impact of delay announcements on hospital network coordination and waiting times. *Management Sci.* 65(5):1969–1994.

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

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54(5):1185–1195.

Hu M, Li Y, Wang J (2018) Efficient ignorance: Information heterogeneity in a queue. *Management Sci.* 64(6):2650–2671.

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

Ibrahim R (2018) Sharing delay information in service systems: A literature survey. *Queueing Systems* 89(1-2):49–79.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, UK).

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–292.

Kumar P, Kalwani MU, Dada M (1997) The impact of waiting time guarantees on customers' waiting experiences. *Marketing Sci.* 16(4):295–314.

Larson RC (1987) OR Forum—perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.* 35(6):895–905.

Lemon KN, Verhoef PC (2016) Understanding customer experience throughout the customer journey. *J. Marketing* 80(6):69–96.

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

Lyft (2019) Prime time for drivers. Accessed August 17, 2019, https://help.lyft.com/hc/en-us/articles/115012926467-Prime-Time-for-drivers.

Maister DH (1984) *The Psychology of Waiting Lines* (Harvard Business School, Cambridge, MA).

MordorIntelligence (2019) Ridesharing market: Growth, trends, and forecast (2020-2025), Orbis Research, Accessed June 17, 2021, https://www.orbisresearch.com/reports/index/ridesharing-market-growth-trends-and-forecast-2020-2025.

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

Soman D, Shi M (2003) Virtual progress: The effect of path characteristics on perceptions of progress and choice. *Management Sci.* 49(9):1229–1250.

Stone A (2012) Why waiting is torture. *New York Times*, http://www.nytimes.com/2012/08/19/opinion/sunday/why-waiting-in-line-is-torture.

Uber (2019) How surge pricing works. Accessed August 17, 2019, https://www.uber.com/us/en/drive/partner-app/how-surge-works.

Webb E, Yu Q, Bretthauer K (2017) Linking delay announcements, abandonment, and staffing: A behavioral perspective. Preprint, submitted March 25, http://dx.doi.org/10.2139/ssrn.3706565.

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

Yu Q, Allon G, Bassamboo A (2020) The reference effect of delay announcements: A field experiment. *Management Sci.* 67(12):17–26.

Yu Q, Allon G, Bassamboo A, Iravani S (2018) Managing customer expectations and priorities in service systems. *Management Sci.* 64(8):3942–3970.

Copyright 2022, by INFORMS, all rights reserved. Copyright of Management Science is the property of INFORMS: Institute for Operations Research and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.