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

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–33
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Frustration-Based Promotions: Field Experiments in Ride-Sharing

Maxime C. Cohen,[a] **Michael D. Fiszer,**[b] **Baek Jung Kim**[c]

[a] Desautels Faculty of Management, McGill University, Montreal QC H3A 1G5, Canada; [b] Via Transportation Inc., New York, New York 10013; [c] Sauder School of Business, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada

Contact: maxime.cohen@mcgill.ca, https://orcid.org/0000-0002-2474-3875 (MCC); sguibor@gmail.com (MDF); baekjung.kim@sauder.ubc.ca (BJK)

Received: March 18, 2018
Revised: October 15, 2019; December 2, 2020
Accepted: January 18, 2021
Published Online in Articles in Advance: August 31, 2021

https://doi.org/10.1287/mnsc.2021.3991

Copyright: © 2021 INFORMS

Abstract. The service industry has become increasingly competitive. One of the main drivers for increasing profits and market share is service quality. When consumers encounter a bad experience, or a *frustration*, they may be tempted to stop using the service. In collaboration with the ride-sharing platform Via, our goal is to understand the benefits of proactively compensating customers who have experienced a frustration. Motivated by historical data, we consider two types of frustrations: long waiting times and long travel times. We design and run three field experiments to investigate how different types of compensation affect the engagement of riders who experienced a frustration. We find that sending proactive compensation to frustrated riders (i) is profitable and boosts their engagement behavior, (ii) works well for long waiting times but not for long travel times, (iii) seems more effective than sending the same offer to nonfrustrated riders, and (iv) has an impact moderated by past usage frequency. We also observe that the best strategy is to send credit for future usage (as opposed to waiving the charge or sending an apologetic message).

History: Accepted by Vishal Gaur, operations management.

Keywords: ride-sharing • field experiments • quality management • service operations

1. Introduction

In an economy where customers have access to a large amount of information and can easily compare alternative services, how can a company keep customers from straying to competing firms? For services with repeat users, such as e-commerce and ride-hailing platforms, customers can easily switch between competitors. Each customer decides which service to use at a given time based on price and service quality, which are easy to obtain. Furthermore, in competitive markets, prices are roughly the same so that the main distinction lies in other factors, such as quality of service or a positive feeling toward the brand. Thus, businesses are constantly seeking ways to enhance their service quality and connection with customers.

Inevitably, in some cases the service will not achieve the desired quality level, and users will encounter a poor experience. Some of those customers will take the time to file a complaint by email, by calling the customer service hotline, or by posting on social media. It is common for companies to generously compensate the customers who reach out.[1] This is especially true for online platforms. For example, when experiencing a delivery delay for a product purchased on Amazon or on jet.com, one can obtain a $5 or $10 gift card with very little effort. Similar practices can be found in the airline industry (airline companies offer compensation under various circumstances), the hospitality sector (customers often receive free bar vouchers), and the food industry (in a restaurant, one is often offered a free dessert). However, in most cases, if the customer does not voice a complaint, no compensation will be received. In this case, the unhappy customer will simply be disappointed with the service and may decide to stop using it.

Many firms are aware of this issue and are actively working on possible solutions. One relevant business practice is when a firm fails to achieve its publicized level of service, it will provide compensation to the customer. A good example is the 20-minute delivery guarantee offered by Domino's Pizza for some orders in certain locations. When the delivery is late, Domino's will offer a free pizza voucher for the next order.[2] A second recent example is Amazon and Walmart offering store credit for late Christmas deliveries in December 2017.[3] Committing to a universal guaranteed service level may be challenging in practice as it depends on several external factors. Instead, companies typically seek to compensate for substantial frustrations (i.e., bad experiences) depending on the context, the relationship with the customer, and their competitive advantage.

It may be hard to detect the various potential frustrations experienced by customers while using the service. In addition, one may want to carefully distinguish between authentic frustrations and those that are more ambiguous. If a company could automatically detect the legitimate frustrations in real time, it would then be possible to send a targeted proactive compensation to the customer who experienced the frustration. This practice could then be optimized to target users who encountered the worst experiences with timely and proportionate forms of compensation. One recent such example is when Best Buy sent proactive apologetic emails to customers who preordered the iPhone 7 in September 2016. Several disappointed customers who experienced delays in shipments of their smartphones were proactively offered a $100 discount on their next purchase.[4]

How can we design the process of sending targeted proactive compensation to customers experiencing a frustration in the context of ride-sharing? What is the impact of such a practice? For companies that build a strong data-driven strategy, this ambitious undertaking is now possible and is the motivation of this paper.

In the context of ride-sharing, customers interact with several online platforms to request on-demand transportation services. In recent years, this means of transportation seriously disrupted the industry. In the United States, several companies compete for market share including Uber, Lyft, and Via. When ordering a ride, the customer specifies the origin and destination locations. Each service provider may offer a price and a waiting time (as well as various quality attributes that are intrinsic to each firm). After selecting a service provider, the customer waits for a vehicle to arrive at a predetermined pickup location, boards the vehicle, and is dropped off at the requested destination. In shared services, such as Via, Lyft Line, and UberPOOL (all operating in New York City), the vehicle's route may be modified to pick up and drop off other passengers. Customers may thus experience several types of frustrations, such as long waiting times, a high number of stops, significant detours, driver no-shows, and poor service interaction with the driver.

In this paper, we collaborate with one of the leading ride-sharing platforms, Via (some background on the company can be found in Section 3.2). We design and run three field experiments to study the impact of proactively compensating riders who experienced a frustration.

1.1. Summary of Results

Given the popularity of ride-sharing online platforms, this paper studies a timely practical problem. In addition, topics related to service quality are at the core of most service providers' priorities. Our results can be summarized as follows:

- Discussing two frustration types and two engagement metrics in ride-sharing. Motivated by historical data and by the ride-sharing market, we consider two types of frustration: long waiting times and long travel times (see details in Section 3.1). To measure the riders' engagement behavior, we compute the total number of rides and spending—allowing us to measure customer engagement via both the frequency and the monetary value.
- Analyzing the impact of proactively compensating frustrated riders. Our first field experiment uses four different compensation conditions: Control (no action), Comms (apologetic message), Credit (offering a $5 credit for the next ride), and Waived (reimbursing the fare). We find that the Credit condition is significantly different from each other condition. Namely, proactively offering a $5 credit to frustrated riders boosts their engagement behavior relative to not offering compensation. We also find that offering a $5 credit to frustrated riders is revenue enhancing. On the other hand, sending an apologetic text message or waiving the charge did not yield a statistically significant effect in our experiment, and hence such compensation does not seem to be effective.
- Refining our results. We perform the same analysis for each type of frustration separately. We observe that the main effect (i.e., offering a $5 credit has a positive impact on rider engagement) is significant for long waiting times but not for long travel times. We also study how the main effect is moderated by the pre-experiment engagement. We find that the effect is significant for frequent and intermediate riders but not for infrequent riders.
- Testing the robustness of our findings. We run a second experiment in a different market at a different period to test the robustness of our findings. We observe similar qualitative insights on the impact of proactively compensating frustrated riders. In addition, we consider a new condition (called Discount) in which frustrated riders receive a 50% discount on their next ride. Our results suggest that a 50% discount remains as effective as sending a $5 credit.
- Compensating nonfrustrated riders. Instead of sending compensation to riders who experienced a frustration, in our third experiment, we consider sending a reward to nonfrustrated users. Specifically, we

target a subset of riders on their "Viaversary" date (the calendar date on which they joined the service) and offer them a $5 reward. Our findings suggest that rewarding riders after a frustration seems more effective than rewarding riders for an arbitrary milestone with the company.

1.2. Related Literature

This paper is closely related to the literature on service quality management in marketing and operations. Several marketing papers focus on service quality and customer retention (see, e.g., Parasuraman et al. 1985, Zeithaml et al. 1996, Mittal and Kamakura 2001, and the references therein). These studies investigate the relationship between service quality and firms' profits (see, e.g., Zahorik and Rust 1992). Other papers examine how customers react to service failures or dissatisfaction (Anderson et al. 1994, Taylor 1994, Smith and Bolton 1998, Smith et al. 1999, Berry and Parasuraman 2004). For example, Smith and Bolton (1998) examine how customers' dissatisfaction from service failures affects their cumulative assessment. Similarly, several studies in operations management (e.g., Craighead et al. 2004, Anderson et al. 2009) consider how service failures or bad experiences (such as flight delays or unpleasant hotel stays) affect customer satisfaction and firms' profits.

In addition, there is a stream of papers that study the relationship between compensation for service failures and customer engagement (Weiner 1985, Bitner 1990, Bitner et al. 1990, Kelley et al. 1993, Smith and Bolton 1998, Hoffman et al. 2003, Tsiros et al. 2004, Bolton et al. 2007, Grewal et al. 2008, Roggeveen et al. 2012). This stream of papers conveys that there are several ways to address a service failure for the firm, such as fixing the failure quickly, cocreating a recovery strategy, and issuing an appropriate compensation. In the case of core service failures, the firm must fix the problem in a timely manner (Parasuraman et al. 1991). However, simply fixing the problem (e.g., booking the customer on the next flight) is often not enough as customers may expect to be compensated in order to preserve the equity of their relationship with the firm. Compensation is the most common method to restore equity (Walster et al. 1973, Bitner 1990, Bitner et al. 1990, Kelley et al. 1993, Smith and Bolton 1998, Hoffman et al. 2003, Grewal et al. 2008, Grewal and Levy 2009). These papers argue that compensating customers can help dissipate their dissatisfaction from a service failure. Nevertheless, offering compensation without an appropriate explanation often drives negative evaluations, as it may indicate an admission of guilt (Bitner 1990). Grewal et al. (2008) found that compensation is necessary only when the firm is responsible for the failure, and the failure occurs frequently. In addition, the authors propose a potential mechanism related to stability (i.e., the likelihood to reoccur) and locus of responsibility.

This paper contributes to the stream of literature on compensation for service failure as a recovery strategy. We aim to bridge four gaps with respect to the existing literature: (1) *Considering incentives as opposed to only using compensation.* Specifically, our paper examines which type of compensation is the most effective. We empirically investigate whether proactively offering a monetary incentive following a frustration is effective, and if so, which type of action (apologetic message, waiving the charge, or offering credit for future usage) works best. It is interesting to understand how immediate compensation is different from incentives for future usage. (2) *Role of prior engagement as a moderator.* We show that frequent users react differently relative to infrequent users. (3) *Studying this problem with field experiments as opposed to laboratory studies.* Previous studies in this domain mainly exploit laboratory experiments, surveys, or critical incident techniques. In our paper, however, we design and run field experiments to identify the causal effect of different types of compensation following a service failure, and its impact on customer engagement. (4) *The proactiveness of the apology.* Customers who receive compensation did not complain or voice their dissatisfaction as it is typically the case in earlier papers. Instead, the firm proactively offered compensation to frustrated users.

Whereas most previous papers consider the traditional service industry (e.g., airlines and hotels), our paper focuses on online ride-sharing platforms. These two contexts admit several key differences: (i) online ride-sharing platforms allow tracking service quality in real time with a high granularity, (ii) riders use the service frequently at a low cost per ride, and (iii) riders incur a low switching cost between competing service providers. As a result, we can investigate the causal effect of proactive compensation on riders' engagement by running field experiments.

This paper is also related to field experiments in online platforms. Typical platforms can decide to run a series of carefully designed experiments (often called A/B tests) to validate some intuitions on users' behavior. For more details on this topic, we refer the reader to the paper by Kohavi et al. (2013). Several researchers in the operations management community have recently used field experiments to address research questions (see, e.g., Zhang et al. 2017, Fisher et al. 2018, Gallino and Moreno 2018, Cui et al. 2019, Singh et al. 2019).

1.3. Structure of the Paper

Section 2 develops our hypotheses. Section 3 describes our data and discusses two key metrics related to riders' engagement behavior. Section 4 reports some evidence on how service quality affects engagement.

Sections 5–7 present the design and results of our three field experiments. Finally, our conclusions are reported in Section 8. Several additional analyses and robustness tests are relegated to the appendices.

2. Hypotheses Development

In this section, we develop our hypotheses on how different proactive compensation methods affect the future engagement of frustrated users (the concept of frustration is formally defined in Section 3.1). Our development is based on various theories in economics, psychology, and marketing. Our first hypothesis comprises three parts and will be tested in our first field experiment (Section 5).

Hypothesis 1. *Proactively compensating frustrated users increases their future engagement behavior:*

a. *Proactively compensating frustrated users by offering them a future credit increases their engagement behavior.*

b. *Proactively compensating frustrated users by offering them a refund increases their engagement behavior.*

c. *Proactively compensating frustrated users by sending an apology message does not increase their engagement behavior.*

Previous studies in economics and psychology have documented the effect of apologies on various outcomes related to consumer behavior. First, the economics of apologies address the importance of apologies to customers who experience service failures to retain their engagement with firms (see, e.g., Ho 2012, Halperin et al. 2019). Using the principal-agent model of a customer-firm relationship (see details in Ho 2012), these studies suggest that (i) apologies are effective only when the apology is costly to the firm, and (ii) the effect of apologies increases with the apology cost. The idea behind these papers is based on the signaling mechanism of the firm's type (i.e., high or low) through the cost of apologies. In particular, high-type firms have an incentive to send high-cost apologies to distinguish themselves from low-type firms. Then, the high-cost apologies lead consumers to view the firms that sent cost-incurring apologies as a high type, and are ultimately more likely to use their service.

Another research stream in psychology has investigated the effect of apologies in different contexts such as brand performance and interpersonal relationship (e.g., Aaker et al. 2004, Skarlicki et al. 2004, Abeler et al. 2010, De Cremer et al. 2011, Ohtsubo et al. 2012, Ohtsubo et al. 2020). Ohtsubo et al. (2012, 2020) provide an alternative explanation on why costly apologies are more effective than no-cost apologies based on the concept of sincerity. These studies found that costly apologies are viewed as a more sincere intention to restore the damaged relationship and, hence, have a greater impact.

Motivated by the aforementioned theories related to costly apologies in economics and psychology, we hypothesize that proactively offering a cost-incurring compensation will increase customers' future engagement (Hypotheses 1(a) and (b)). In our case, either providing a $5 credit for future use or offering a refund for the frustrated ride can be seen as a cost-incurring apology. On the other hand, we hypothesize that a non-cost-incurring proactive compensation—in our case, proactively sending an apology message (with no monetary benefit)—will not increase customers' engagement behavior (Hypothesis 1(c)). In particular, we investigate what type of proactive cost-incurring compensation is more effective. Our study thus adds to the previous literature by comparing different proactive actions in the context of ride-sharing. The first compensation type is to provide a $5 future credit. This type of compensation is related to the common practice of offering coupons in retail (see, e.g., Nevo and Wolfram 2002, Reimers and Xie 2019). Specifically, Nevo and Wolfram (2002) found that coupons induce customers to repurchase. Related to these previous research, we believe that providing a $5 credit will create a similar effect as coupons and, thus, we hypothesize that proactively providing a $5 credit for future usage will increase the engagement behavior (Hypothesis 1(a)). The second compensation type is to provide a refund for the frustrated ride. According to Wu et al. (2019), it is important to provide a refund option to unsatisfied consumers, and this can increase firms' profits in online retail. Since proactively providing a refund for the frustrated ride may have a similar effect as giving a refund for unsatisfied consumers in online retail, we hypothesize that a refund for the frustrated ride will increase the engagement behavior (Hypothesis 1(b)).

In contrast, as shown in the previous literature related to costly apologies (e.g., Ho 2012, Ohtsubo et al. 2012, Halperin et al. 2019, Ohtsubo et al. 2020), we hypothesize that proactively sending an apologetic message (which can be seen as a "cheap" apology) will not increase customers' engagement behavior (Hypothesis 1(c)). Since the identification of the underlying mechanism to explain why a non-cost-incurring compensation would not be effective is not the main goal of our paper, we will not attempt to disentangle the potential different explanations to support this hypothesis. However, related to the theories discussed previously, non-cost-incurring compensation may not have a significant impact on customer engagement.

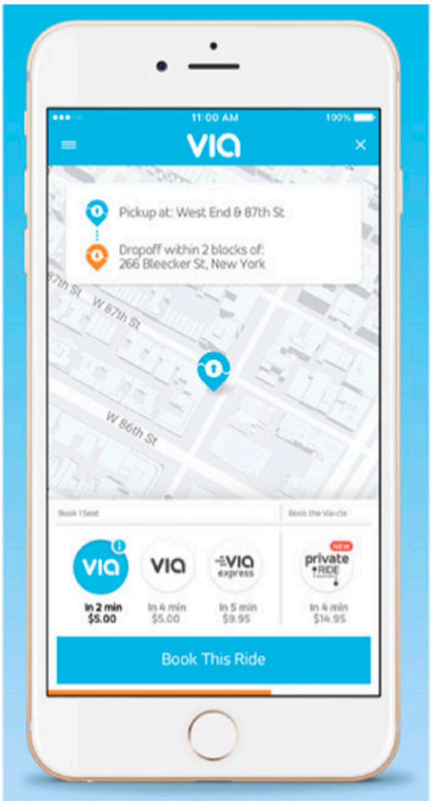
We next develop our second hypothesis, which will be tested in our second field experiment (Section 6).

Hypothesis 2. *Proactively offering a $5 credit or a lower amount (a 50% discount for the next ride) to frustrated* *users are equally effective in increasing their future engagement behavior.*

Both offering a $5 credit for future usage and a 50% discount for the next ride fall under the category of cost-incurring apologies. Thus, based on the theories mentioned earlier, both types of compensation should enhance customer engagement. Our hypothesis here is that $5 is too high of an amount, so that one can achieve the same goal by using a lower amount. On the one hand, it is well known that providing a very low amount (e.g., a $1 credit) will have no effect on customer engagement (see, e.g., Kalwani and Yim 1992). Indeed, it may be perceived as a cheap offer and will not be attractive for customers. On the other hand, providing a high amount will definitely trigger customer re-engagement. The question is, then, how can we find the right amount that is still successful in increasing customer engagement but not cost prohibitive to the firm. This question is related to the extensive literature on estimating price elasticity (see, e.g., Hoch et al. 1995, Andreyeva et al. 2010). In the context of ride-sharing, Cohen et al. (2016) has used data from Uber to estimate price elasticities and consumer surplus.

Hypothesis 2 is also related to the extensive literature that compares absolute and relative promotional discounts (see, e.g., Chen et al. 1998, McKechnie et al. 2012, Lehtimäki et al. 2019). As it shown in the literature, the framing of the offer can have a significant impact on the customer reaction.

Figure 1. (Color online) Screenshot of Via's Interface


Pickup at: West End & 87th St
Dropoff within 2 blocks of: 266 Bleecker St, New York

Source. App Store.

3. Context and Data

In this section, we present our context, industry partner, and data. We then describe our key metrics related to the riders' engagement behavior and our data-filtering procedure.

3.1. General Scope

As discussed, our goal is to proactively send compensation to riders who experienced a low quality of service (henceforth, a frustration). We focus on two types of frustration that can be experienced by riders in the context of a ride-sharing platform: (1) long waiting times and (2) long travel times. In the ride-sharing industry, these two quality dimensions are closely related to customer satisfaction. We selected these two metrics after analyzing rider feedback originating in the year 2017. When riders place a ride request by specifying pick-up and drop-off locations, they typically receive a price quote together with an estimated time of arrival (ETA) for the driver to arrive. For example, in Figure 1 (see lower left side), the rider is offered a ride for $5 with an ETA of two minutes (among other options that are beyond the scope of this paper). Subsequently, the rider can decide whether to accept the quote. Once the request is accepted, the driver will be en route to pick up the rider. It is clear that if the driver arrives later than the proposed ETA, it affects the quality of service. We call this type of frustration *a (positive) ETA error*. For example, if the proposed ETA was two minutes but the driver arrived for pick-up after 13 minutes, the ETA error is equal to 11 minutes. In such a situation, the rider experiences a frustration due to having to wait longer than anticipated.

The second frustration metric is related to travel time. Since we consider a ride-sharing platform that allows several passengers heading in the same direction to share the same vehicle, the total travel time may be affected by several factors. For example, the travel time is affected by the number of riders picked up and dropped off by the driver (i.e., the number of stops). The travel time can also be affected by traffic and weather conditions. Note that factors such as number of stops or itinerary can be controlled by Via, whereas factors such as weather cannot. Our goal is to capture an aggregated measure that is normalized for uncontrolled factors. For the purpose of our field experiments, we consider a metric called *VGR*, which stands for Via Google ratio. For each ride, we know the value of the total realized travel time (for a given origin-destination pair and a specific time), called the Via duration. Next, we use the Google Maps API to access the predicted time it would have taken to complete the same ride at the same time in a private car, according to the Google Maps estimate (referred to as the Google duration). We then compute the ratio of the Via duration divided by the Google duration. The VGR metric allows us to normalize for uncontrolled factors and to isolate the effects related to the quality of service. Note that neither frustration metric (ETA error and VGR) can be controlled by the rider.

We then define a frustration by focusing on riders who experienced either a long ETA error or a high VGR when riding with Via (precise definitions are reported later in this paper). As we discuss in Section 3.2, Via provides the vast majority of the rides within very good service levels on both dimensions (i.e., short waiting times and low VGRs).

3.2. Industry Partner and Data Description

In this section, we provide a brief overview of our industry partner, Via Transportation Inc., or simply Via. Founded in June 2012, Via is a privately held transportation company based in New York City (NYC) focusing on real-time ride-sharing. The company offers its users a smartphone application to match riders with drivers on-demand. An advanced algorithm enables multiple passengers headed in the same direction to seamlessly share a ride, managing fleets of dynamic shuttles with high efficiency, and rerouting vehicles in real time in response to demand variations. As of January 2018, Via is providing over 1.5 million rides monthly.[5]

Unlike most competitors that started as private ride providers, Via's product was designed to provide shared rides—the Via algorithm is optimized to increase the utilization of vehicles while keeping detour levels minimal for all passengers. As a result, most rides during working days/hours from anywhere in Manhattan to anywhere else in Manhattan cost $5. Via's customer service philosophy is summarized as follows: "We, at Via, LOVE each and every one of our customers! Customer Service agents are a human extension of the Via product. This means our number one priority when providing real-time support should be to prevent bad experiences from happening." Via's member service associates respond live to rider texts to help solve issues, provide context, and have considerable discretion to compensate proactively.

To guide the empirical analysis presented in this paper, we use a large historical data set. Specifically, our data set includes all the rides completed in NYC between May 1 and December 31, 2017. Each observation in our data set is a ride (i.e., a rider who is traveling from a given origin to a destination on a specific day/time). For each observation, we have access to several observables features, such as rider ID, exact times and locations (of both pick-up and drop off), distance traveled, proposed ETA, ETA error, trip duration, and price paid.

Our data set includes several million rides completed by a large number of different riders.[6] In this paper, the two relevant features are the ETA error and the VGR, which translate to two different types of frustration. It is apparent in our data that these two metrics are excellent for the vast majority of Via rides. In particular, the average ETA error amounts to 0.404 minutes (i.e., 24.24 seconds) with a standard deviation of 2.172, which means that riders wait on average less than 25 seconds more than the proposed ETA. As noted earlier, Via still strives to improve the experience for customers who encountered a high ETA error and in many cases issues compensation.

We select a 10-minute threshold for ETA error for a ride to qualify as a frustration (for our first field experiment). Note that ETA errors greater than 10 minutes occur in less than 0.34% of the rides in our data set. Even though experiencing a 10-minute ETA error is rare, riders who use the service more frequently have an increased chance of experiencing such an error. As a result, addressing this type of frustration is an important problem in practice. Similarly, we decided to set a threshold of 2.0 on the VGR to define a frustrated experience. The occurrences where the VGR is greater than 2.0 are also very rare in our data set (due to confidentiality, we cannot reveal summary statistics on VGR).

3.3. Engagement Metrics

Our goal is to identify appropriate metrics to measure riders' engagement with the platform. Capturing engagement behavior highly depends on the application under consideration. In the context of online platforms, it is clear that the engagement is related to the frequency of usage and to the amount of money spent on the platform. Nevertheless, the appropriate time scale is unclear (shall we consider a one week window or a one month period?). To measure riders' engagement behavior, we consider the following two metrics: (i) total spending (in $) during the first T weeks after being exposed to the experiment, and (ii) total number of rides completed during the first T weeks after the experiment. We vary the value of T between 1 and 4, allowing us to examine both the short-term impact and the effect on a longer time horizon (we also consider increasing the value of T up to 11 and find consistent results).

Both metrics—spending and number of rides—allow us to investigate how riders' engagement behavior varies depending on the condition—namely, how different proactive actions can effectively compensate frustrated riders. We will also compare these two metrics to their corresponding values prior to the experiment, allowing us to measure how the postexperiment engagement differs from the pre-experiment behavior. We next discuss in greater detail these engagement metrics:

$$\text{Total-spending}_T^j = \frac{1}{N_j}\sum_{i\in j}\sum_{t=1}^{T}\text{Total-spending}_{it}^j, \qquad (1)$$

$$\text{Total-rides}_T^j = \frac{1}{N_j}\sum_{i\in j}\sum_{t=1}^{T}\text{Total-rides}_{it}^j, \qquad (2)$$

where i corresponds to a rider, j to a compensation condition (defined formally in Section 5), and t to a week. As mentioned, T denotes the length of the time window (i.e., $T \in \{1,2,3,4\}$ weeks after being exposed to the experiment). In Equation (1), $\text{Total-spending}_{it}^j$ represents the dollar amount spent by rider i from condition j in week t, and N_j indicates the total number of riders in condition j. Thus, $\text{Total-spending}_T^j$ denotes the average cumulative spending of riders in condition j during T weeks. Similarly, in Equation (2), $\text{Total-rides}_{it}^j$ represents the number of rides completed by rider i from condition j in week t, so that Total-rides_T^j is the average total number of rides completed by riders in condition j during T weeks.

By comparing $\text{Total-spending}_T^j$ and Total-rides_T^j across the different conditions, we can understand how the different compensation methods affect the engagement of riders who experienced a frustration. To complement our analysis, we will also consider the total spending and number of rides prior to the experiment. These results, however, do not provide any indication on whether frustrations affect the engagement behavior. To identify the causal effect of the frustration on the engagement behavior, we will construct a sample of nonfrustrated riders (this analysis is presented in Section 5).

3.4. Data Filtering

To highlight the patterns observed in our data, we carefully refine the sample of analyses. First, we filter our experimental data by removing riders displaying exceptionally high usage.[7] Removing outliers in terms of extremely high usage levels will help make our sample more representative. To address this issue, we eliminate the top 1% of observations based on the distribution of each key metric. For example, to analyze the total spending within the first T weeks after being exposed to the experiment, we first look at the distribution of this variable and discard the top 1% of riders who spent the most (similarly, we eliminate the top 1% of observations for the total number of rides). To ensure the robustness of our results, we vary this threshold from 1% to 5% by increments of 1%. We observed consistent results under each of these thresholds.

4. Impact of Service Quality on Engagement

Before presenting the design and results of our field experiments, we confirm the common intuition that poor service encounters may adversely affect future usage in the context of our ride-sharing platform. Specifically, we explore the correlation between our frustration metrics (ETA error and VGR) and the engagement behavior (total spending and number of rides). For this analysis, we use historical (i.e., nonexperimental) data. The results of this analysis provide a concrete motivation for running our field experiments and for investigating how such decrease in engagement can be proactively compensated by the platform.

Since we cannot observe the ETA error and the VGR when riders do not complete a ride, we aggregate the variation of these variables at the week level. We estimate the following specification:

$$\begin{aligned}\log(\text{Total-rides})_{it} &= \text{ETA-error}_{it-1} + \log(\text{VGR})_{it-1}\\ &\quad + \text{Controls}_{it} + \mu_i + \nu_t + \epsilon_{it},\end{aligned}$$

where i corresponds to a rider and t to a time unit (i.e., week). The dependent variable, Total-rides_{it}, measures the total number of rides completed by rider i at time t (similar results were obtained for the total spending, as discussed later). The first two independent variables, ETA-error_{it-1} and VGR_{it-1}, capture the weekly average ETA error and VGR from the rides completed by rider i in the previous week $(t-1)$. We exploit the variation in both lagged variables since they are measured only when a ride is completed. We also include hours of day, day of week, and free/waived rides as additional controls. Finally, we include individual and time fixed effects to capture any time-invariant individual specific effects and unobserved heterogeneity across riders as well as unobserved time-specific demand shocks. For this analysis, we use a data sample with more than 100,000 riders (who completed 770,604 rides) for five months between May 1 and October 1, 2017.

The results for the total spending and number of rides are reported in Table 1. As expected, we find that both ETA error and VGR in the previous week have a negative impact on the current riders' engagement. It implies that these two metrics capture a potential frustration in the context of ride-sharing. We refrain to claim that this result is causal since ETA error and VGR may naturally be endogenous (formally addressing the endogeneity issue is beyond the scope of this paper and we keep our analysis to be correlative). Nevertheless, we partially address the endogeneity of our frustration metrics by using propensity score

Table 1. Correlation Between Frustration Metrics and Engagement Behavior

	Dependent variable					
	$\log(1 + \text{Total-spending})_{it}$			$\log(\text{Total-rides})_{it}$		
	(1)	(2)	(3)	(4)	(5)	(6)
ETA-error$_{it-1}$	−0.001** (0.0005)		−0.001** (0.0005)	−0.001** (0.0004)		−0.001* (0.0004)
$\log(\text{VGR})_{it-1}$		−0.018*** (0.005)	−0.017*** (0.005)		−0.020*** (0.004)	−0.020*** (0.004)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	770,604	770,604	770,604	770,604	770,604	770,604
R^2	0.382	0.382	0.382	0.367	0.367	0.367

Notes. All standard errors are clustered at the rider level. The variables *Total-spending* and *VGR* are log-transformed.
$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

matching in Section 5. In the next three sections, we present the design and results of our three field experiments.

5. Experiment 1: New York City

In this section, we present the design and results of our first field experiment. We then refine our results by estimating heterogenous treatment effects. Finally, we complement our findings by conducting several robustness tests. Our main econometrics method is the difference-in-differences (diff-in-diffs). Nevertheless, we also consider analysis of variance (ANOVA) and regression analyses to showcase the robustness of our results (the details are relegated to the appendix).

5.1. Design and Implementation

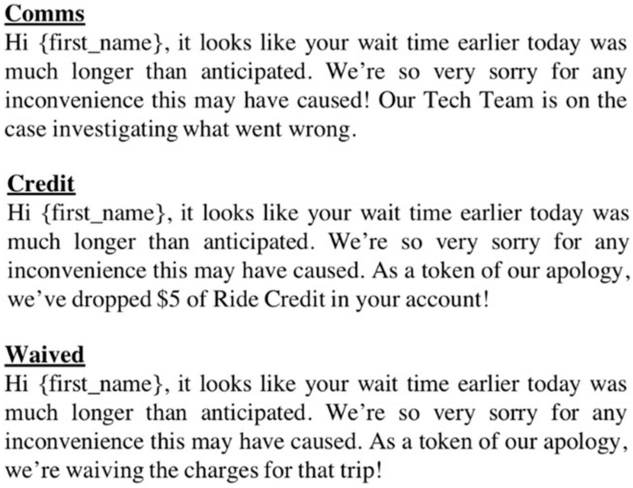
Our objective is to investigate the impact of different proactive actions on riders who experienced a frustration. As mentioned, we define a frustration by either a long ETA error or a high VGR. In particular, we consider a threshold of 10 minutes for the ETA error to qualify for a frustrated event. It is clear that waiting 10 extra minutes is perceived as a bad experience for riders. We decided to use the threshold value of 10 minutes based on riders' feedback. As mentioned in Section 3.2, such events correspond to the 0.34% worst values in our data set. Similarly, we set the threshold value for VGR to 2.0.

Our first experiment was conducted between July 5 and August 25, 2017, in NYC. During this period, we monitored a subset of riders who experienced an unexpectedly long ETA error or a high VGR.[8] If a rider experiences either an ETA error greater than 10 minutes or a VGR higher than 2.0, we classify this observation as a frustrated rider. We then randomly assign those frustrated riders to the four following conditions: (i) Control, (ii) Comms (Communications), (iii) Credit, and (vi) Waived. The Control condition represents the set of riders who experienced a frustration but did not receive any compensation (this group will be used as the baseline of our analysis). The Comms condition includes the set of riders who received a text message from Via to apologize for the inconvenience that may have been experienced (without any monetary compensation). The Credit condition represents the set of riders who received a $5 credit to be used for the next ride. Last, the Waived condition includes the riders who received a waived ride (i.e., the charges for the ride were refunded to the rider).[9] Each rider was sent the appropriate promotion via a text message. The sample of text messages sent to riders who experienced a long ETA error can be found in Figure 2 (the messages for the VGR category are similar and are reported in Figure A.1 of Appendix A).

Overall, our experiment includes a total of 3,982 riders divided as follows: Control (969), Comms (999), Credit (1,354), and Waived (660). In addition, we control for several factors. First, we ensure that the same rider is not included twice in the experiment. Second, we focus on rides that are typical and representative (i.e., we remove very long or expensive rides, very short rides, etc.). Third, we constantly monitor the number of occurring frustrations to avoid identifying frustrations that are not caused by the service quality (e.g., if the highway was closed due to a special event). Our goal is to rigorously analyze the behavior of riders from the four conditions (Control, Comms, Credit, and Waived) after being exposed to the experiment. This will allow us to reach a better understanding on how different actions affect the engagement behavior after experiencing a frustration. The results of this experiment are presented in Sections 5.2 and 5.3.

As mentioned, this experiment includes a total of 3,982 riders. After applying the filter from Section 3.4, we are left with 3,943 riders divided as follows:

Figure 2. Examples of Text Messages Sent to Riders in Our Field Experiment (for the ETA Error Category)



Comms
Hi {first_name}, it looks like your wait time earlier today was much longer than anticipated. We're so very sorry for any inconvenience this may have caused! Our Tech Team is on the case investigating what went wrong.

Credit
Hi {first_name}, it looks like your wait time earlier today was much longer than anticipated. We're so very sorry for any inconvenience this may have caused. As a token of our apology, we've dropped $5 of Ride Credit in your account!

Waived
Hi {first_name}, it looks like your wait time earlier today was much longer than anticipated. We're so very sorry for any inconvenience this may have caused. As a token of our apology, we're waiving the charges for that trip!

Control (961), Comms (992), Credit (1,341), and Waived (649). As we can see, the number of riders in the different conditions is not symmetric. This is due to the fact that our assignment suffered from two technical difficulties: (A) the Control condition (for both ETA error and VGR) did not get assigned any rider during the first week of the experiment, and (B) the Waived condition (only for VGR) did not work properly for the first 28 days (users who were supposed to be assigned to the Waived condition were assigned to the Credit condition instead). To account for these unfortunate issues, we run a series of robustness tests. First, we remove the observations from the first week of our experiment and re-estimate all our econometrics models. This fully addresses issue (A). Second, we subsample our data to ensure that we have the same number of users in each condition and re-estimate our models. Specifically, we select the minimum number of users from each condition and randomly sample the same number of users from the other conditions. Third, we consider using only the data starting from day 29 (so that the experiment length is reduced) and re-estimate our models. This will address both issues (A) and (B). We find the same qualitative results for all three robustness tests, hence strengthening the validity of our results. The results of these robustness tests are reported in Section A.7 of Appendix A. Since the third robustness test is the closest to a randomized setting, we will also report the main result of this test in Section 5.3. Finally, we present comprehensive balancedness tests to showcase the validity of our randomization in Section A.8.

5.2. Preliminary Results

Figure 3 reports the average total spending for each condition using $T = 4$. We consider different time windows in Section A.5. Our results allow us to infer the following (by taking the average across all samples):

- Riders in the Credit condition spent 11.98% more relative to riders in the Control condition.
- Riders in the Credit condition spent 13.60% more relative to riders in the Comms condition.
- Riders in the Credit condition spent 11.74% more relative to riders in the Waived condition.

For the total spending, the result of the one-way ANOVA test (see, e.g., Maxwell et al. 2017) is significant for all four conditions ($F(33,939) = 5.43$, $p < .01$). In addition, Table 2 presents the results of the pairwise comparisons among the different conditions. We find that the pairwise comparison between Credit and each other condition is positive and statistically significant at the 99% level.[10]

When riders experience a frustration (either a long ETA error or a high VGR), the service provider can make up for it by proactively offering compensation to the frustrated riders. Our results suggest that providing a $5 credit toward a future ride is significantly more effective than sending an apologetic message or waiving the charge. Thus, our results support our Hypotheses 1(a) and (c) but reject Hypothesis 1(b) (see Section 2). Such a finding bears the following important practical implications:

1. The platform can compensate for a poor experience by proactively sending compensation.

2. It seems more effective to credit a rider's account than waive the charge. We highlight that the Credit condition outperformed the Waived condition even though the average offered amount was lower ($5 versus $5.62). In addition, the expected cost of the Credit condition is smaller than the cost of the Waived condition, as some of the users who receive the future credit will not necessarily use it (and hence the cost for such users is zero). The difference between these two compensation methods may be explained by the fact that providing a credit value offers an opportunity to use the service again. Since the quality of service is high most of the time, the rider is very likely to experience a high quality of service (low ETA error and small VGR),

Table 2. Pairwise Comparisons Between Conditions for Experiment 1

	Dependent variable: *Total-spending*	
	Mean difference	*p*-value
Comms: Control	−0.140	0.732
Credit: Control	1.173	0.002
Waived: Control	0.021	0.964
Credit: Comms	1.313	0.001
Waived: Comms	0.161	0.724
Waived: Credit	−1.151	0.008

Table 3. Diff-in-Diffs Results for Experiment 1

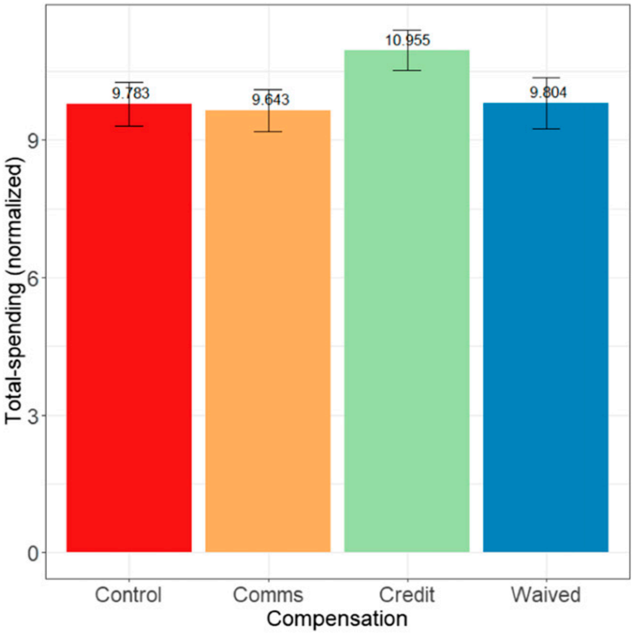
	Dependent variable				
	Total-rides$_{it}$ (1)	Total-spending$_{it}$ (2)	log(1 + Total-rides$_{it}$) (3)	log(1 + Total-spending$_{it}$) (4)	Total-spending$_{it}$ (5)
$\mathbb{1}\{\text{Comms}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.022	0.093	0.012	0.027	0.204
	(0.016)	(0.082)	(0.008)	(0.018)	(0.130)
$\mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.043***	0.213***	0.024***	0.059***	0.265**
	(0.014)	(0.075)	(0.007)	(0.016)	(0.126)
$\mathbb{1}\{\text{Waived}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.005	0.039	0.003	0.012	0.170
	(0.017)	(0.090)	(0.009)	(0.019)	(0.123)
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	220,808	220,808	220,808	220,808	87,864
R^2	0.302	0.224	0.287	0.250	0.220

Note. All standard errors are clustered at the rider level. Column (5) is based on the data starting from day 29 (robustness test for the randomization, see more details in Section A.7 of Appendix A).

$^{**}p < 0.05$; $^{***}p < 0.01$.

and as a result, this will correct for the previous frustration. On the other hand, by sending an apologetic message or by waving the charge, it can be considered as an instant treatment, without providing an incentive to try the service again. We will elaborate more on these potential mechanisms in Section 5.3. In addition, offering credit is better than waiving the charge from a pure revenue perspective. This finding is also related to theories from behavioral economics on how customers perceive future gains relative to reimbursements for past expenses.

Figure 3. (Color online) Average Total Spending for Experiment 1



Notes. This figure reports the total spending during the 28 days after the experiment for each condition. The confidence interval is reported at the 90% level. We normalize all the numbers presented in all figures to avoid revealing sensitive information. This normalization does not affect the relative differences between the conditions, which is the main focus of this paper.

3. The Credit condition is significantly different relative to each other condition. Nevertheless, Control, Comms, and Waived are not statistically different from each other.

4. Offering a $5 credit to frustrated riders is revenue enhancing (that is, the average difference in spending between riders in the Credit and Control conditions exceeds $5). Our results suggest that riders in the Credit condition will spend on average an extra 11.98% relative to riders in the Control condition. In addition, by subtracting the $5 investment, this marketing campaign is revenue enhancing.

We present additional results in the appendix. Specifically, Section A.3 shows the ANOVA results for each frustration type separately (ETA error and VGR), Section A.4 considers the pre-experiment usage level as a moderator, Section A.5 considers varying the value of T between one and four weeks, and Section A.6 reports the results of several regression analyses. In summary, we show that our main effect is statistically significant for the ETA error but not for VGR, for frequent users, and for all time windows. We find that it takes 18 days to earn the $5 back (i.e., the difference in total spending between the Credit and Control conditions becomes larger than $5 after 18 days). Such a metric is very important when designing marketing campaigns in the context of online platforms. A return on investment in 18 days is considered a high level of performance. These findings suggest that by proactively sending monetary compensation, the firm can mitigate the adverse effect of frustrated riders who decrease their engagement level.

So far, our analysis has focused on investigating how proactively compensating frustrated riders affects their aggregated total spending and number of rides during $T = 4$ weeks. An interesting question is whether the \$5 of credit offered in the Credit condition drives entirely the effect. To answer this question, we conduct two analyses: (1) subtracting \$5 for riders in the Credit condition, and (2) removing the data from the first $N \geq 1$ weeks for riders in the Credit condition. We also extend our analysis up to $T = 11$ weeks (instead of $T = 4$) after the experiment exposure.

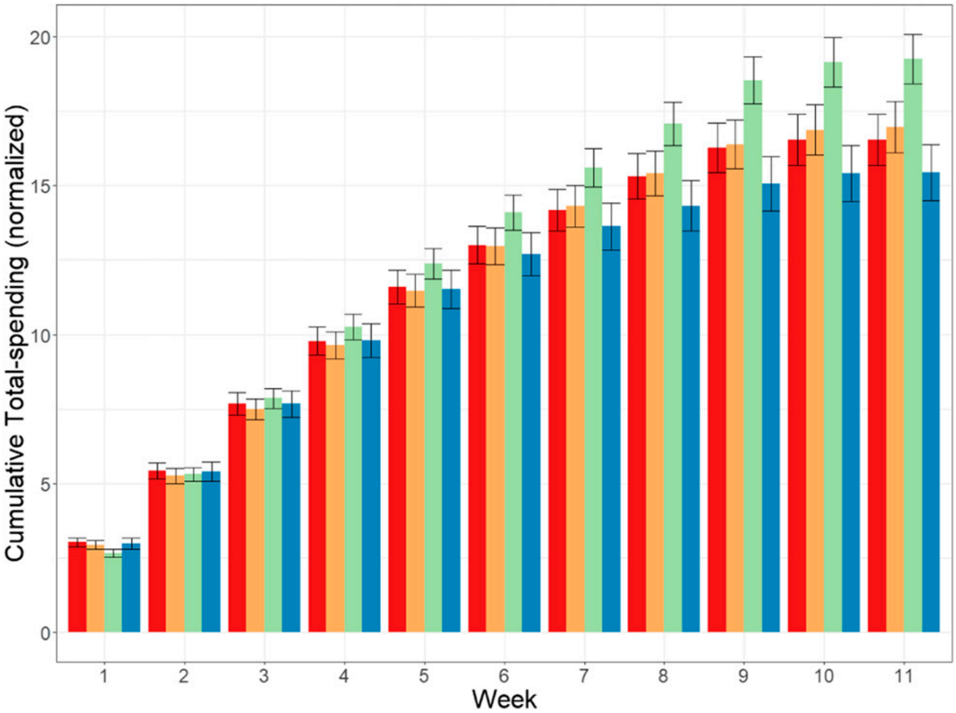
Figure 4 presents the results for the cumulative spending when removing the first ride for riders in the Credit condition (i.e., the ride for which riders use the \$5 credit). After subtracting \$5 for riders in the Credit condition, the spending in the Credit condition is not statistically higher than the other conditions for the first seven weeks. However, from the eighth week onward, we can see that riders in the Credit condition are more likely to complete rides relative to the other conditions, and this effect persists until the eleventh week (we do not have the data beyond that point). This suggests that our qualitative results still hold after subtracting \$5 (which might be treated as a free ride), so that our results are robust to the inclusion or exclusion of the \$5 credit.

We next go one step further and remove the first $N \geq 1$ weeks of data. Riders may increase their usage right after the intervention because of the (short-term) impact of the proactive compensation (especially, riders in the Credit condition). To mitigate this concern, we investigate whether the effect persists after removing the first N weeks of usage in the analysis. We vary the value of N between one and three weeks and report the results for $N = 3$ in Figure 5. As we can see, even after removing the first three weeks of data, we still find a significant treatment effect (this result is consistent when using a value of one, two, or three weeks). However, if we use $N \geq 4$ weeks, the effect is not statistically significant anymore. This analysis demonstrates that the effect of proactively compensating lasts for some time.[11] Several previous studies show that deep discounts can make consumers pay less in the future. Reasons include: (1) consumers form a price expectation and the promoted price could become a reference point (Winer 1986) or decrease future demand (Cohen et al. 2017); (2) consumers devalue the quality of the service or product (Davis et al. 1992); and (3) deep discounts (including products offered for free) lead to perceptions of lower costs and higher margins (Davis et al. 1992). Interestingly, this is not the case in our study as we find that the effect of compensating frustrated riders lasts for some time.

5.3. Empirical Strategy: Difference-in-Differences

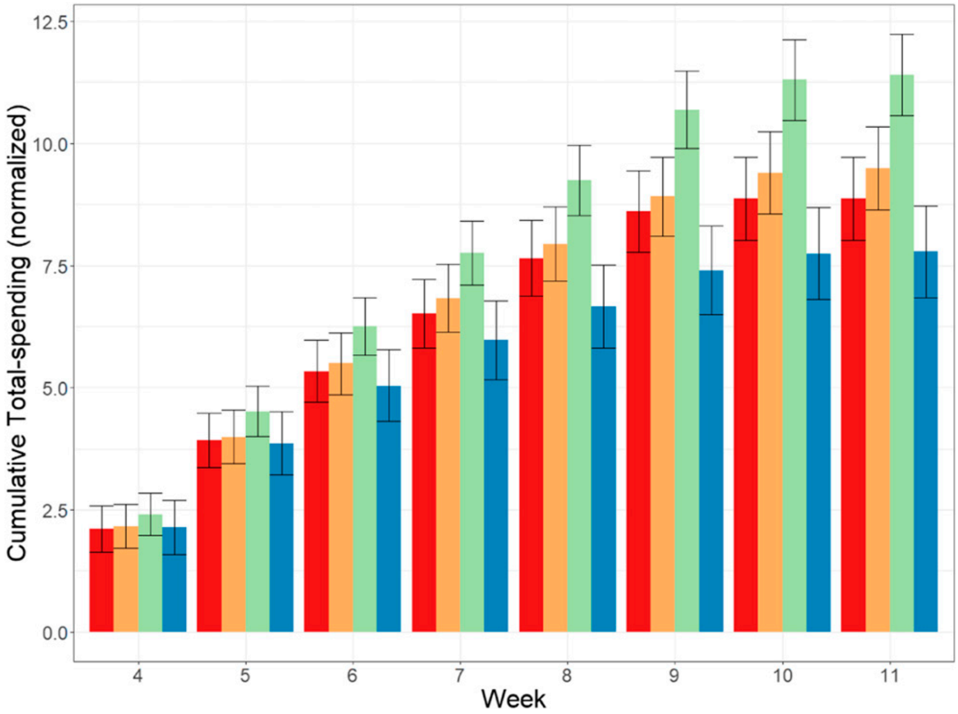
In the previous section, we presented the results of one-way ANOVA tests, without accounting for pre-experiment engagement. To examine how frustrated riders' behavior is affected by the different proactive actions, we next compare pre- and postexperiment engagement. Specifically, we use a difference-in-differences (diff-in-diffs) approach (see, e.g., Angrist

Figure 4. (Color online) Cumulative Total Spending (After Subtracting \$5 for the Riders in Credit Condition)



Note. Confidence interval is reported at the 90% level.

Figure 5. (Color online) Cumulative Total Spending (After Removing the First Three Weeks)



Note. Confidence interval is reported at the 90% level.

and Pischke 2008) to identify the impact of the different interventions. We consider riders' usage during the 28 days before and after the experiment and estimate the following specification:

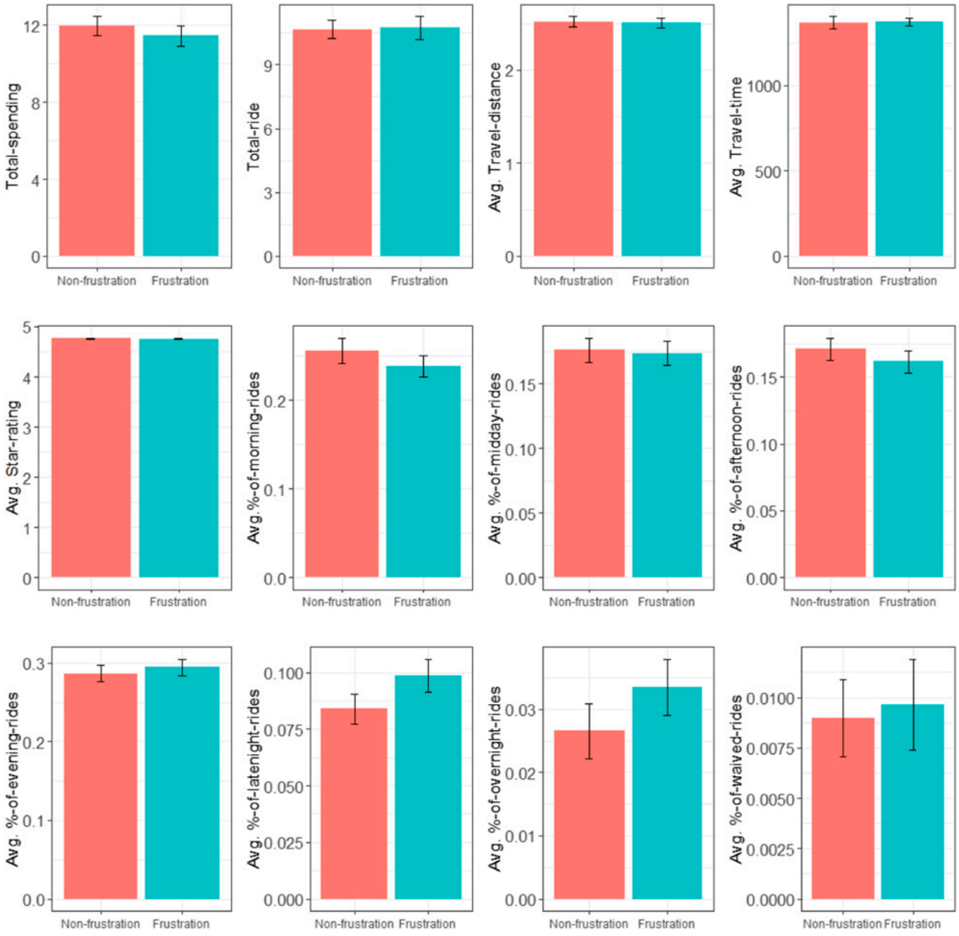
$$\begin{aligned} y_{it} &= \beta_1 \cdot \mathbb{1}\{\text{Comms}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t \\ &\quad + \beta_2 \cdot \mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t \\ &\quad + \beta_3 \cdot \mathbb{1}\{\text{Waived}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t \\ &\quad + \mu_i + \kappa_t + \epsilon_{it}, \end{aligned} \tag{3}$$

where i corresponds to a rider and t to a time period (for this analysis, we aggregate the observations at the day level); and y_{it} denotes the dependent variable for rider i at time t (we consider both total spending and number of rides). The indicators $\mathbb{1}\{\text{Comms}\}_i$, $\mathbb{1}\{\text{Credit}\}_i$, and $\mathbb{1}\{\text{Waived}\}_i$ are binary variables to indicate the condition assigned to rider i (the Control condition is the reference group). The indicator $\mathbb{1}\{\text{After-experiment}\}_t$ is a binary variable for the period after the experiment exposure. Finally, we include individual fixed effects (denoted μ_i) and time fixed effects (κ_t) to capture any time-invariant individual specific effects and unobserved heterogeneity across riders as well as unobserved time-specific demand shocks. The key parameters in Equation (3) are β_1, β_2, and β_3. These parameters capture the potential causal effect of each type of compensation (following the frustration) on the engagement behavior.

As shown in Table 3, the interaction coefficient between the Credit condition and the postexperiment period is positive and statistically significant. This implies that riders in the Credit condition use the service more (and spend more) relative to riders in the Control condition during the postexperiment period. Consequently, this suggests that the effect of the Credit condition in response to the frustration is causal. On the other hand, the interaction coefficients for Comms and Waived are not significant, that is, the three conditions (Control, Comms, and Waived) are not statistically different in terms of postexperiment engagement from each other (at least, based on the data and results from our field experiment). We estimate the same model with both dependent variables (total spending and number of rides) and find consistent results. For robustness purposes, we estimate the same model with a different baseline (i.e., Comms or Waived) and find that the coefficient of the interaction term between the Credit condition and the postexperiment period is positive and statistically significant. In addition, the parallel trend assumption holds (see details in Table A.4 in the appendix), thus suggesting that the interaction coefficient between the Credit condition and the postexperiment period captures a causal effect (see details in Section A.5). It thus confirms that the Credit condition has a positive effect on riders' engagement, whereas the other conditions do not seem to have a significant effect.

As discussed, the fact that proactively providing a credit of $5 for the frustrated ride is different from

Figure 6. (Color online) Conditional Independence Assumptions



Note. Confidence interval is reported at the 90% level.

offering a refund (i.e., the Waived condition) can be explained as follows. A $5-credit compensation incentivizes customers to come back and use the service again in order to spend the money in their account (the next ride can be seen as a free ride, so why not use it). If the firm is providing a high-quality service for this new ride, customers will then adjust their belief on the firms' service quality (and type), and this correction process is likely to enhance the future engagement behavior. On the contrary, although giving a refund for the frustrated ride is a cost-incurring compensation, it does not directly incentivize customers to come back and use the service again. The absence of such a returning incentive may not offer a chance for consumers to adjust their belief on the service quality. Thus, it may lead to a different impact on the engagement behavior.

We next investigate whether our findings vary for different frustration types. We estimate Equation (3) for each segment separately (ETA error and VGR) and report the results in Table 4. Interestingly, we find that the interaction coefficient between the Credit condition and the postexperiment period is positive and statistically significant only for the ETA error. One possible explanation is that riders tend to blame the service provider for high ETA errors but not for large VGR values. In other words, waiting more than anticipated can be perceived as the driver's fault (or an issue with Via's dispatch algorithm). On the other hand, a large VGR, which translates to a long travel time, seems to be more acceptable as riders may attribute the blame to external factors (instead of blaming the service provider). This finding is consistent with previous studies, such as Grewal et al. (2008), who find that locus of responsibility matters, and that compensation is necessary only when the company can be held responsible. A second possible explanations is related to transparency (see, e.g., Buell et al. 2016). For a long ETA error, the rider does not have transparency (even though riders can check the interface and monitor the driver's progress and potential detours, they do not know whether the delay is driven by the driver or by other factors). When riders are in the car, however, they can directly observe the road and traffic

Table 4. Diff-in-Diffs Results for Each Frustration Type (ETA Error vs. VGR) for Experiment 1

	Frustration types			
	ETA error		VGR	
	Total-rides$_{it}$ (1)	Total-spending$_{it}$ (2)	Total-rides$_{it}$ (3)	Total-spending$_{it}$ (4)
$\mathbb{1}\{\text{Comms}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.039	0.127	0.003	0.026
	(0.024)	(0.124)	(0.020)	(0.108)
$\mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.062**	0.308**	0.029	0.153
	(0.025)	(0.127)	(0.018)	(0.095)
$\mathbb{1}\{\text{Waived}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.024	0.018	−0.018	0.056
	(0.024)	(0.123)	(0.025)	(0.139)
Time fixed effects	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes
Observations	97,020	97,020	117,666	117,666
R^2	0.304	0.224	0.303	0.228

Note. All standard errors are clustered at the rider level.
$^{**}p < 0.05$.

conditions. As a result, riders may attribute the blame to the service provider when they do not have transparency into the work that they are doing, and to external factors when they do have transparency.

For robustness purposes, we vary the threshold value for the VGR and consider four values: 2.0, 2.33, 2.66, and 3.0. For each threshold value, we estimate our model and find a consistent result (the results are omitted for conciseness). This shows that our finding (i.e., proactively compensating the frustrated ride with a high VGR is not effective) is robust to different VGR threshold values.

5.4. Heterogeneous Treatment Effects

In the previous section, we found that proactively offering credit to frustrated riders has a positive effect on their future engagement. We next examine the heterogeneity of this treatment effect with respect to past usage by considering several dimensions. Specifically, we exploit riders' pre-experiment usage behavior during the 28 days prior to the experiment and consider: (1) total number of rides, (2) total spending, (3) average travel distance, and (4) tenure. We divide the riders into three groups: high (top 30%), middle (top 30%–70%), and low (bottom 30%), based on each feature during the 28 days prior to the experiment. We then examine how our main finding (i.e., the Credit compensation is more effective relative to the other conditions) varies across segments by estimating Equation (3) for each segment of riders:

1. Total rides and spending. The results are reported in Table A.1. As we can see, riders in the Credit condition show a higher spending level relative to the Control riders (the Comms and Waived conditions do not seem to have an effect) for the high and middle groups. The effects are not statistically significant for the low group. The results for the total number of rides are consistent with the total spending. We thus conclude that the Credit compensation for frustrated riders is more impactful for riders in the high and middle groups. This suggests that the most effective strategy is to compensate frequent riders. Infrequent users are still in the discovery phase of exploring the service and do not show a different engagement pattern across the different conditions (note that the churn rate for services such as ride-sharing is often high). A more detailed discussion can be found in Section A.4 in Appendix A.

2. Average travel distance. The results are reported in Table A.2. As we can see, the interaction coefficient between the Credit condition and the postexperiment period is positive and statistically significant only for the low segment. This result implies that the treatment effect is significant for riders who typically complete short rides.

3. Tenure. The tenure is defined as the number of days since the rider joined the platform (i.e., first ride with Via). The results are presented in Table A.3. As we can see, the coefficient of the interaction term between the Credit condition and the postexperiment period is positive and statistically significant only for the middle and low groups. This result implies that (relatively) more recent riders are more likely to ride (and spend more) when they receive a Credit compensation for the frustration.

5.5. Frustrated vs. Nonfrustrated Riders

Based on the results of Section 4, both ETA error and VGR have a negative effect on riders' engagement. To confirm this finding, we next leverage the experimental data. Specifically, we combine the nonexperimental data with our experimental data by only considering the riders in the Control condition (i.e., riders who experienced a frustration but did not receive any compensation). We highlight that experiencing a frustration is endogenous because it is a rare event that can depend on various riders' features (e.g., rider frequency, morning versus late night users, regular commuters versus occasional users). Since we have a high cross-section variation of the usage behavior across riders, we conduct a propensity score matching (PSM) analysis to define a group of nonfrustrated riders. The details of the analysis are reported next:

1. In our data, we have 961 frustrated riders (from the Control group) and a much larger number of nonfrustrated riders (several tens of thousands from our nonexperimental data set). Thus, the number of nonfrustrated riders is much larger (as expected). In addition, nonfrustrated riders are more likely to be nonfrequent users. Therefore, we filter our sample based on riders' pre-experiment usage. This does not mean that we keep only high users. Instead, we fit a distribution of nonfrustrated riders (from the nonexperimental data) who are similar to the frustrated riders (from the experimental data). For robustness purposes, we do this filtering in several ways. The filter is based on the number of rides in both the pre- and postexperiment periods. Specifically, we restrict our sample to users with a number of pre-experiment rides between four and 150 and a number of postexperiment rides larger than three (we obtain consistent results for several other threshold values). Note that the distributions of frustrated and nonfrustrated riders are similar and statistically indifferent, as shown in Figure 6.

2. As discussed, we filter our sample based on riders' pre-experiment usage. We define the pre-experiment period to be the four weeks prior to the first day of our experiment (which is July 5, 2017). We then compute the pre-experiment engagement during these 28 days for both frustrated and nonfrustrated riders.

3. We next compute a propensity score based on the following equation (using the R package *matchit* with the nearest-neighbor matching method):

$$\mathrm{P}(\mathbb{1}\{\text{Frustration}\}_i)|\mathrm{X}_i) = \frac{\exp(\mathrm{X}_i)}{\exp(1 + \mathrm{X}_i)},$$

where i indicates a rider and $\mathbb{1}\{\text{Frustration}\}_i$ is an indicator that is equal to 1 if rider i is a frustrated rider (and noncompensated) and 0 otherwise. The vector X_i includes several control variables related to riders' pre-experiment engagement, such as total number of rides, total spending, average ride distance, average ETA error, average VGR, whether the ride was free or waived, and request time. Note that the average distance between matched samples is 0.052 (min: 0.005, median: 0.035, and max: 0.987).

4. After computing the propensity score, we match the frustrated riders with nonfrustrated riders based on the nearest neighbor. Nearest-neighbor matching selects the best control match for each frustrated rider. Specifically, in each matching step, we choose the nonfrustrated rider who is not yet matched but is the closest to the frustrated rider.

5. Based on this matching process, we obtain 961 nonfrustrated riders.

6. As shown in Table 6, the conditional independence assumptions are valid, that is, the matched samples are well balanced across all the relevant dimensions.

We then conduct a diff-in-diffs analysis using the matched data set (with the 961 nonfrustrated riders defined earlier):

$$\text{Total-rides}_{it} = \beta_1 \cdot \mathbb{1}\{\text{Frustrated}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t + \mu_i + \nu_t + \epsilon_{it}, \quad (4)$$

where i indicates a rider and t a time unit (in this case, a day). In Equation (4), the indicator $\mathbb{1}\{\text{Frustrated}\}_i$ is equal to 1 if rider i is in the (frustrated) Control condition of the experiment and 0 otherwise; and $\mathbb{1}\{\text{After-experiment}\}_t$ is a binary variable for the postexperiment period. Thus, the baseline group consists of the riders in the nonfrustrated condition (from the matching analysis) and β_1 captures the effect of frustration on engagement. Finally, μ_i and ν_t represent exposure fixed effects (capturing time-invariant heterogeneity at the rider level) and unobserved shocks at time t, respectively.

We extend the specification in Equation (4) by incorporating the users from the Credit condition (i.e., users who experienced a frustration and received a proactive compensation of $5). Specifically, we add the term $\beta_2 \cdot \mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$ to Equation (4). The variable $\mathbb{1}\{\text{Credit}\}_i$ is an indicator for riders in the Credit condition and the parameter β_2 captures the effect of Credit compared with nonfrustrated riders. The results are presented in Table 5.

As we can see, the interaction coefficient between frustrated riders (i.e., $\mathbb{1}\{\text{Frustrated}\}_i$) and the postexperiment period (i.e., $\mathbb{1}\{\text{After-experiment}\}_t$) is negative and statistically significant, hence, implying that the frustration has a negative impact on future engagement. This effect is likely to be causal since we accounted for the endogeneity of frustrations in the matching analysis. This result implies that after the frustration, riders who do not receive a credit compensation will be less likely to complete future rides. Note that we obtained a consistent result when using PSM with a regression instead of a diff-in-diffs (the details are omitted for conciseness).

More interestingly, our analysis can answer the following question: Do riders who experience a frustration combined with a proactive credit compensation (i.e., riders in the Credit condition) exhibit the same (or higher) level of future engagement relative to riders who do not experience a frustration? To answer this question, we look at the coefficient of the Credit indicator variable. This coefficient is not statistically

Table 5. PSM Results with Frustrated and Nonfrustrated Riders

	Dependent variable			
	Total-rides$_{it}$ (1)	Total-spending$_{it}$ (2)	Total-rides$_{it}$ (3)	Total-spending$_{it}$ (4)
$\mathbb{1}\{\text{Frustrated}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	−0.064***	−0.202**	−0.064***	−0.202**
	(0.015)	(0.084)	(0.015)	(0.084)
$\mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$			−0.020	0.011
			(0.014)	(0.079)
Time fixed effects	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes
Observations	109,554	109,554	185,991	185,991
R^2	0.242	0.194	0.274	0.207

Notes. All standard errors are clustered at the rider-level. Columns (1) and (2) present the results of the comparison between frustrated and nonfrustrated riders. Columns (3) and (4) compare the engagement of riders in the Credit condition, frustrated riders in the Control condition, and nonfrustrated riders.

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

significant, implying that the engagement level of riders who experienced a frustration combined with a credit compensation is not different from nonfrustrated riders. This finding bears interesting practical implications in the context of service management.

6. Experiment 2: Washington, DC

We next design and run a second experiment in Washington, DC. Our first goal is to test the robustness of the findings from experiment 1 on a different market (e.g., different traffic patterns, competition intensity, market maturity, and types of riders) during a different time period. In addition, we refine the design of our experiment by exploiting the knowledge gathered in the first experiment and consider two different compensation levels.

6.1. Design and Implementation

To sharpen our insights, we decided to test different levels of monetary compensation. In the first experiment, we only use a $5 credit (which is a typical compensation level for the platform). In the second experiment, we use $5 and also consider a smaller amount by offering a 50% discount on the next ride. The average cost of a ride in Washington, DC, in our data is $3.60, so that a 50% discount on the next ride amounts approximately to $1.80. The discount is capped at $3 to eliminate the incentive of a potential gaming behavior. This additional condition allows us to examine how different credit levels impact the engagement of riders who experienced a frustration. We also vary the threshold value for the ETA error. Instead of using 10 minutes, we decreased the threshold to eight minutes. The goal of this reduction is to adapt the appropriate definition for a frustration depending on the setting and the historical data. Reducing the threshold from 10 to eight minutes was also motivated by testing the robustness of our results to a lower bound (we could afford to do so given the lower number of observations we expected in the Washington, DC, context). We naturally decided to focus on the ETA error given that the VGR frustration was not statistically conclusive in experiment 1. One of the key

Table 6. Diff-in-Diffs Results for Experiment 2

	Dependent variable			
			log	log
	Total-rides$_{it}$	Total-spending$_{it}$	(1 + Total-rides$_{it}$)	(1 + Total-spending$_{it}$)
	(1)	(2)	(3)	(4)
$\mathbb{1}\{\text{Discount}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.039***	0.082	0.023**	0.034*
	(0.015)	(0.054)	(0.009)	(0.018)
$\mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.031*	0.113*	0.018*	0.039*
	(0.016)	(0.060)	(0.010)	(0.020)
Time fixed effects	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes
Observations	53,534	53,534	53,534	53,534
R^2	0.212	0.203	0.223	0.221

Notes. All standard errors are clustered at the rider level. In column (2), the interaction coefficient between Discount and After-experiment becomes significant at the 95% level with a nonclustered standard error.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

findings of experiment 1 is that riders in the Credit condition are more likely to spend more (and complete a larger number of rides) relative to riders in other conditions. Given this finding, we are interested in varying the amount of credit granted. Therefore, in experiment 2, we decided to use the following three conditions: Control, Discount, and Credit. The Control and Credit conditions are defined in the same way as in experiment 1. The Discount condition includes riders who received a 50% discount for their next ride. The text messages can be found in Section B.1 in Appendix B. Note that in the second experiment, we removed the Comms and Waived conditions. This follows from the results observed in experiment 1, which were in clear favor of the Credit condition. Experiment 2 was conducted from September 28 to November 7, 2017, in Washington, DC. This experiment includes a total of 948 subjects divided as follows: Control (308), Discount (342), and Credit (298). As before, by comparing riders' behavior in those three conditions, it will allow us to test the robustness of our findings and to refine our managerial insights. After applying our filter (see Section 3.4), we are left with 923 riders divided as follows: Control (305), Discount (332), and Credit (286). As mentioned, we discarded the Comms and Waived conditions as well as the VGR frustration in this experiment. This was motivated by the results obtained in experiment 1 regarding the lack of effectiveness of such variants. Instead, we added the Discount condition to investigate the impact of different levels of compensation by testing our Hypothesis 2 developed in Section 2. For conciseness, we present only the diff-in-diffs results. The ANOVA and regression results are relegated to Appendix B.

6.2. Empirical Strategy: Difference-in-Differences

As in experiment 1, we use a diff-in-diffs approach to identify the impact of the different compensation conditions. The model specification is given by:

$$\begin{aligned} y_{it} = {} & \beta_1 \cdot \mathbb{1}\{\text{Discount}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t \\ & + \beta_2 \cdot \mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t \\ & + \mu_i + \kappa_t + \epsilon_{it}, \end{aligned} \tag{5}$$

where i corresponds to a rider and t to a day; and y_{it} denotes the dependent variable for rider i at time t (we consider both total spending and number of rides). The indicators $\mathbb{1}\{\text{Discount}\}_i$ and $\mathbb{1}\{\text{Credit}\}_i$ are binary variables to indicate the condition assigned to rider i (as before, the Control condition is the reference group); and $\mathbb{1}\{\text{After-experiment}\}_t$ is a binary variable for the postexperiment period. Finally, we include individual (μ_i) and time (κ_t) fixed effects to capture any time-invariant individual specific effects and unobserved heterogeneity across riders as well as unobserved time-specific demand shocks. As before, the key parameters in Equation (5) are β_1 and β_2. These parameters capture the potential causal effect of each type of compensation (following the frustration) on the engagement behavior. As in experiment 1, we observe that the average pre-experiment behavior is similar for the different conditions, so that the parallel trend assumption holds between conditions (see details in Section B.4).

As shown in Table 6, the interaction coefficients between the Credit and Discount conditions with the postexperiment period are positive and statistically significant (the coefficient for the Discount condition with the total spending becomes statistically significant after applying a logarithmic transformation). This implies that riders in the Credit and Discount conditions use the service more (and spend more) relative to riders in the Control condition during the postexperiment period. Consequently, the effects of the Credit and Discount conditions in response to the frustration are causal. As a result, this confirms that the Credit and Discount conditions have a positive effect on the engagement of frustrated riders. In addition, the interaction coefficient between Discount and After-experiment is not statistically different from the coefficient between Credit and After-experiment (e.g., comparison of these two coefficients in column (1): t-statistics= -1.38). This result implies that offering a \$5 credit is indifferent from a 50% discount (corresponding to an average of \$1.80). Thus, our results support our Hypothesis 2 (see Section 2). In summary, the results presented in Table 6 translate to the following insights:

1. We could replicate the same findings as in experiment 1, that is, riders in the Credit and Discount conditions are more likely to be engaged relative to riders in the Control condition. Note that both experiments (NYC and Washington, DC) are quite different in terms of market size, maturity (Via has been operating for a much longer time in NYC), period of the year, and alternatives for transportation. Still, within each experiment, we could find similar results and managerial insights on the impact of compensating frustrated riders.
2. Even though we used a stricter criterion to define a frustration (by lowering the ETA error threshold from 10 to eight minutes), we could still observe the same main effect.
3. The difference between the Credit and Discount conditions is not statistically significant. This is interesting as the company seeks to determine the optimal compensation level for frustrated riders. Our results suggest that a 50% discount remains nearly as effective as \$5 credit. Note that we perform a pairwise comparisons between conditions and observe a consistent result.

We present additional results in the appendix. Specifically, in Appendix B, Section B.2 shows the

ANOVA results, Section B.3 considers the pre-experiment usage level as a moderator, Section B.4 considers varying the value of T between one and four weeks, and Section B.5 reports the results of several regression analyses. In summary, we find consistent results across three methods: ANOVA, regression, and diff-in-diffs. Our results suggest that the platform does not need to offer a \$5 credit to compensate for the frustration, as a similar effect can be achieved by using a smaller amount. Such a finding is important in practice as the service provider seeks to find the minimal amount that will increase the engagement of frustrated riders. Note that the optimal amount may depend on various factors (seasonality, type of frustration, type of rider, etc.), and identifying the optimal level of compensation is an interesting question left for future research.

7. Experiment 3: Viaversary

Our first two experiments have focused on sending proactive compensation to riders who experienced a frustration. It is clear that sending a reward to riders should increase their engagement. The next question is: Is it more effective to send compensation to a rider who experienced a frustration or to a nonfrustrated rider? Note that the answer is not straightforward as frustrated riders may be disappointed by the service and, hence, potentially decrease their engagement. On the other hand, nonfrustrated riders (who are very likely to have experienced higher service levels than frustrated riders) may react better to promotional offers. This question is the main motivation behind our third experiment.

7.1. Design and Implementation

Since it is not a common business practice to send monetary rewards to random users, we decided to send a reward to riders on their Viaversary date, that is, the calendar date on which they joined the platform (Via). In this experiment, we simply divided the riders into the Control and Credit conditions. As before, riders in the Credit condition received a \$5 credit to celebrate their joining date anniversary (whereas riders in the Control condition were not sent anything). Specifically, this experiment was conducted in NYC from October 30, 2017, to January 1, 2018, and includes a total of 605 subjects divided as follows: Control (177) and Credit (428).[12] To ensure that we select a sample of active users, we restrict the selection process to riders who have used the service within two weeks prior to their Viaversary date. This experiment allows us to test the effectiveness of sending compensation to nonfrustrated riders. The text message can be found in Section C.1 of Appendix C.

After applying our filter (see Section 3.4), we are left with 599 riders divided as follows: Control (175) and Credit (424). We next report the results on the total spending and number of rides during $T = 4$ weeks after being exposed to the experiment.

We conduct a manipulation check to validate that the treatment was applied on average to a less frustrating ride in terms of ETA error. We find that riders in experiment 3 have on average a 54% lower ETA error relative to riders in experiment 1 (which was run in the same city). The difference is statistically significant at the 99% level, t-statistics (4,542) = −7.138, p-value < 0.01.

7.2. Empirical Strategy: Difference-in-Differences

As before, we use a diff-in-diffs approach to identify the impact of the Viaversary experiment. The model specification is given by:

$$y_{it} = \beta_1 \cdot \mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t + \mu_i + \kappa_t + \epsilon_{it}, \tag{6}$$

where i corresponds to a rider and t to a day; and y_{it} denotes the dependent variable for rider i at time t (we consider both total spending and number of rides). The indicator $\mathbb{1}\{\text{Credit}\}_i$ is a binary variable to indicate the condition assigned to rider i (as before, the Control condition is the reference group); and $\mathbb{1}\{\text{After-experiment}\}_t$ is a binary variable for the post-experiment period. Finally, we include individual (μ_i) and time (κ_t) fixed effects to capture any time-invariant individual specific effects and unobserved heterogeneity across riders as well as unobserved time-specific demand shocks. As before, the key parameter in Equation (6) is β_1. This parameter captures the potential causal effect of the Viaversary compensation on the engagement behavior.

As shown in Table 7, the interaction coefficient between the Credit condition and the postexperiment period is not statistically significant. We obtained the same result using a one-way ANOVA test (see Section C.2). This implies that riders in the Credit condition do not use the service more (and do not spend more) relative to riders in the Control condition during the postexperiment period. As a result, the Credit condition does not seem to have a positive effect on the engagement of nonfrustrated riders. We conclude that rewarding riders after a frustration seems more effective than rewarding riders for an arbitrary milestone with the company.

The practical implication of this finding can be communicated as follows. Given a limited budget of promotions, it seems more effective to allocate promotions to riders who have experienced a frustration (as the results from experiment 3 suggest that sending promotions to nonfrustrated riders do not seem to have a significant effect).

8. Conclusion

In this paper, we investigated whether a service provider should proactively compensate users who experienced a frustration (i.e., a low level of service). When a user experiences a low level of service, the future engagement of this particular user is at risk. A possible strategy for the service provider is to proactively send compensation following the frustration. The questions are then the following: Is it effective to do so? If yes, what is the potential impact on the engagement behavior? How do different actions (e.g., sending credit versus waiving the charge) compare? For which types of frustration and which groups of users does compensation work best?

To answer these questions, we partnered with the ride-sharing platform, Via. We designed and ran three field experiments to study the impact of compensating riders who had experienced a frustration. Motivated by historical data, we considered two types of frustration: long waiting times and long travel times. Using a difference-in-differences approach, we find that sending compensation to frustrated riders (i) is profitable and boosts their engagement behavior, (ii) works well for long waiting times but not for long travel times, (iii) seems more effective than sending the same offer to nonfrustrated riders, and (iv) has an impact that is moderated by past usage. We also observed that the best strategy is to send credit for future usage.

We believe that our results are generalizable to settings where customers frequently use the service. Examples of this type of industries are online platforms such as food delivery services, ride-sharing, and e-commerce. In such settings, the high frequency of usage can be leveraged in order to provide an opportunity for frustrated customers to use the service again and correct for their frustrated experience. Another important feature present in our setting is the rarity of service failures (i.e., frustrations). One reason why our proactive $5 credit compensation was effective is the fact that service failures are relatively rare in our setting. If the service failures are frequent, riders' belief adjustment may not be as effective as referenced in prior studies that found that repeated apologies have a negative impact (see, e.g., Schweitzer et al. 2006, Ho 2012).

The research presented in this paper advances our understanding of how firms should proactively respond to service failures. In the context of an online ride-sharing platform, we tested several types of apologies: providing a credit of $5, offering a refund, and sending an apologetic message. We found that offering a $5 credit was the only compensation type that was successful at enhancing customer engagement. Our results can be explained by several potential mechanisms, such as costly versus cheap apologies, opportunity to correct for firms' beliefs, and prospect theory, just to name a few. A great avenue for future research would be to rigorously disentangle between the different potential mechanisms in order to sharpen further the impact of this research.

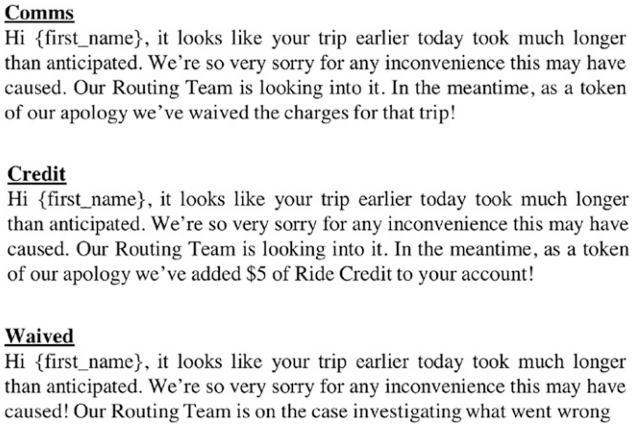
This paper is the first to rigorously investigate the impact of proactively sending compensation to frustrated customers in the ride-sharing market. Our results allow us to draw practical insights for proactive campaigns related to service quality. Besides boosting engagement behavior, this type of compensation leads to additional benefits in terms of customer satisfaction. When receiving such compensation, users are often pleasantly surprised and feel that the service provider is looking after them (a sample of the text messages that users sent to Via after receiving compensation can be found in Appendix D). As we can see, users appreciate the reward, so that this practice can make the difference in a competitive industry. Several interesting extensions are left for future research. As observed, our main effect depends on the type of frustration. For each service industry, one can consider a similar setting under different frustration types with various service levels. In addition, the exact reward amount may be optimized at the rider/time/quality levels by developing customized data-driven campaigns.

Table 7. Diff-in-Diffs Results for Experiment 3

	Dependent variable			
	Total-rides$_{it}$	Total-spending$_{it}$	log (1 + Total-rides$_{it}$)	log (1 + Total-spending$_{it}$)
	(1)	(2)	(3)	(4)
$\mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	−0.019 (0.021)	−0.027 (0.129)	−0.008 (0.011)	0.004 (0.026)
Time fixed effects	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes
Observations	34,742	34,742	34,742	34,742
R^2	0.243	0.143	0.236	0.180

Note. All standard errors are clustered at the rider level.

Figure A.1. Examples of Text Messages Sent to Riders in Our Field Experiment (for the VGR Category)



Comms
Hi {first_name}, it looks like your trip earlier today took much longer than anticipated. We're so very sorry for any inconvenience this may have caused. Our Routing Team is looking into it. In the meantime, as a token of our apology we've waived the charges for that trip!

Credit
Hi {first_name}, it looks like your trip earlier today took much longer than anticipated. We're so very sorry for any inconvenience this may have caused. Our Routing Team is looking into it. In the meantime, as a token of our apology we've added $5 of Ride Credit to your account!

Waived
Hi {first_name}, it looks like your trip earlier today took much longer than anticipated. We're so very sorry for any inconvenience this may have caused! Our Routing Team is on the case investigating what went wrong

Acknowledgments

The authors thank several people at Via's New York City office for insightful discussions and feedback. In particular, they are thankful to Saar Golde, Ori Klein, Alex Lavoie, and Gabrielle McCaig. The authors also thank Gad Allon, Alix Barasch, Anindya Ghose, John Horton, Masakazu Ishihara, Eitan Muller, Sungsik Park, Vishal Singh, Olivier Toubia, Russell Winer, and Nathan Yang for their valuable feedback, which has helped improve this paper.

Appendix A. Additional Results for Experiment 1

A.1. Text Messages for Riders in the VGR Category

Figure A.1 provides example text messages sent to riders in our field experiment.

A.2. Results for Heterogenous Treatment Effects

Tables A.1–A.3 report results for heterogeneous treatment effects for total rides and spending, average travel distance, and tenure, respectively.

A.3. ANOVA Tests for Each Frustration Type

We examine how the main effects are moderated by the frustration type. We run one-way ANOVA tests for each segment separately (ETA error and VGR). The upper (respectively, lower) panel in Figure A.2 reports the average total spending (respectively, number of rides) for the ETA error and VGR segments.[13] As we found in the diff-in-diffs analysis, the main effect (i.e., riders in the Credit condition are more likely to be engaged relative to riders in the other conditions) is replicated for the ETA error but not for the VGR.[14] Note that for the VGR segment, the numerical value of the Credit condition is still higher relative to other conditions, but the differences are not statistically significant. For the ETA error segment, however, the value is higher, and the differences are statistically significant. As mentioned in the paper, one possible explanation is that riders tend to blame the service provider for high ETA errors but not for large VGR values.

A.4. Pre-Experiment Usage as a Moderator

We analyze how the main effects are moderated by different levels of pre-experiment usage as is common in several papers (see, e.g., Serpa and Krishnan 2018). Our motivation behind this analysis is twofold. As discussed in Section 3.2, a frustration (long ETA error or high VGR) is a rare event. As a result, frequent riders are more likely to be exposed to our field experiment. To address this issue, we examine how the main effects are moderated by pre-experiment usage. Note that we deal with this issue more thoroughly in the subsequent sections by comparing the pre- and postexperiment engagement levels and by controlling for this factor in the regression analysis.

We divide the riders from our experiment into three groups, high, middle, and low, based on their usage prior to the experiment. To remain consistent, we compute the total spending and number of rides for each rider during the four weeks before the experiment. Based on these variables, we define the top 30% of riders as the high group, the bottom 30% as the low group, and the remaining riders are assigned to the middle group. For robustness

Table A.1. Heterogenous Treatment Effects: Total Number of Rides and Spending

	Dependent variable					
	Total-rides$_{it}$			Total-spending$_{it}$		
	High (1)	Middle (2)	Low (3)	High (4)	Middle (5)	Low (6)
$\mathbb{1}\{\text{Comms}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.030 (0.039)	0.030 (0.019)	0.003 (0.018)	0.079 (0.182)	0.102 (0.120)	0.058 (0.094)
$\mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.062* (0.035)	0.056*** (0.018)	0.014 (0.018)	0.336** (0.164)	0.266** (0.111)	0.052 (0.089)
$\mathbb{1}\{\text{Waived}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.004 (0.042)	0.027 (0.021)	−0.019 (0.018)	−0.022 (0.196)	0.193 (0.136)	−0.070 (0.100)
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	69,216	85,624	65,968	73,584	74,144	73,080
R^2	0.207	0.130	0.243	0.114	0.111	0.198

Note. All standard errors are clustered at the rider level.
$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.2. Heterogenous Treatment Effects: Average Travel Distance

	Dependent variable					
	Total-rides$_{it}$			Total-spending$_{it}$		
	High (1)	Middle (2)	Low (3)	High (4)	Middle (5)	Low (6)
$\mathbb{1}\{\text{Comms}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.012	0.018	0.039	0.146	0.134	−0.009
	(0.026)	(0.023)	(0.032)	(0.159)	(0.131)	(0.138)
$\mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.034	0.021	0.083***	0.185	0.178	0.286**
	(0.025)	(0.021)	(0.030)	(0.144)	(0.121)	(0.128)
$\mathbb{1}\{\text{Waived}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	−0.016	0.005	0.026	−0.030	0.094	0.031
	(0.031)	(0.023)	(0.034)	(0.177)	(0.138)	(0.157)
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	66,248	88,312	66,248	66,248	88,312	66,248
R^2	0.298	0.308	0.296	0.224	0.220	0.227

Note. All standard errors are clustered at the rider level.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

purposes, we consider a time window of four, five, or six weeks to divide the riders into groups and find consistent results. We also find the same qualitative results when using a continuous variable.

The results are presented in Figure A.3. We find that the difference in average total spending between the Credit and Control conditions is statistically significant only for the high and middle groups.[15] In addition, for the total number of rides, this finding holds only for the middle group. This result implies that the pre-experiment usage level does affect the impact of compensating frustrated riders. Specifically, we obtain a significant effect for the high and middle groups but not for the low group—hence confirming the diff-in-diffs results. This suggests the following managerial insights:

1. Infrequent users who experienced a frustration are not affected by receiving a promotion aimed to compensate their frustration. These infrequent users are still in the discovery phase of exploring the service and do not show a different engagement pattern across the different conditions.

2. Riders in the Credit condition spend more relative to riders in other conditions. Regarding the total spending, the difference between the Credit condition and each other condition for the high group is statistically significant at the 95% level (for the middle group, only the difference between the Credit and Control conditions is significant at the 95% level). For the total number of rides, the middle group shows a significant difference between the Credit and each other condition.

3. For both high and middle groups, it is profitable to offer a $5 credit following the frustration, that is, the additional spending between riders in the Credit and Control conditions exceeds $5 (actually it even exceeds $10, meaning that the return on investment is high).

4. For riders who are very frequent (top 30%), we obtain an average of 11.13% additional spending between the Credit and Control conditions. For riders who are in the middle group, we obtain an average of 15.32% additional spending between the Credit and Control conditions. Thus, the relative effect is the most significant for riders in the middle group.

A.5. Varying the Time Window and Considering Pre-experiment Usage

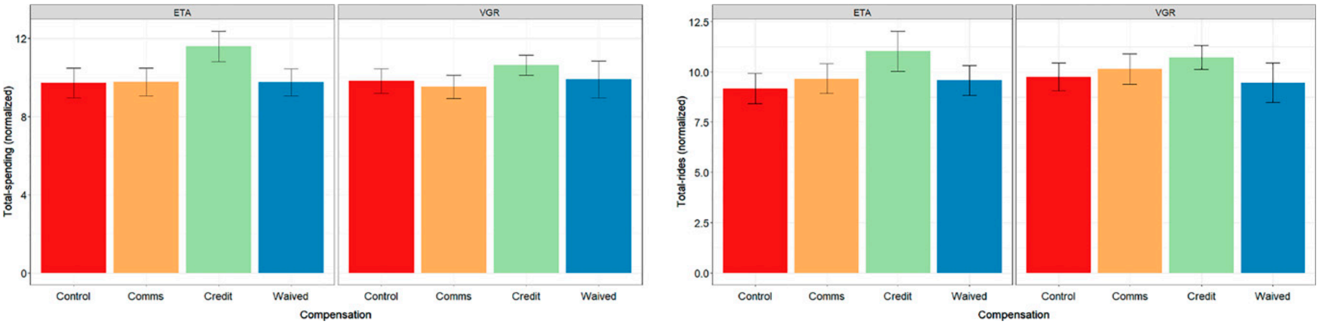
We have shown that riders in the Credit condition are more likely to be engaged relative to riders in the other conditions. We next investigate the pre-experiment

Table A.3. Heterogenous Treatment Effects: Tenure

	Dependent variable					
	Total-rides$_{it}$			Total-spending$_{it}$		
	High (1)	Middle (2)	Low (3)	High (4)	Middle (5)	Low (6)
$\mathbb{1}\{\text{Comms}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.015	0.020	0.033	0.066	0.061	0.178
	(0.030)	(0.023)	(0.030)	(0.148)	(0.122)	(0.166)
$\mathbb{1}\{\text{Credit}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	0.035	0.040*	0.058**	0.171	0.246**	0.201
	(0.029)	(0.021)	(0.028)	(0.142)	(0.108)	(0.154)
$\mathbb{1}\{\text{Waived}\}_i \times \mathbb{1}\{\text{After-experiment}\}_t$	−0.047	0.015	0.034	−0.131	0.082	0.105
	(0.035)	(0.024)	(0.031)	(0.179)	(0.127)	(0.180)
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Rider fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	59,584	101,080	60,144	59,584	101,080	60,144
R^2	0.274	0.288	0.346	0.213	0.222	0.232

Note. All standard errors are clustered at the rider level.
*$p < 0.1$; **$p < 0.05$.

Figure A.2. (Color online) Average Total Spending and Number of Rides by Frustration Type for Experiment 1



behavior of riders in each condition. More precisely, we compare riders' usage between the different conditions during both pre- and postexperiment periods.

The upper part of Figure A.4 presents the total spending over time for different values of T. The upper right panel presents the cumulative total spending during the four weeks following the experiment exposure. The x-axis in the right panels indicates time points (in weeks) starting from the date on which riders were exposed to the experiment ($x = 0$) until four weeks after the exposure time ($x = 4$). The upper left panel reports the cumulative total spending during the pre-experiment period (here, the x-axis indicates the time points from the four weeks prior to the experiment until the date on which riders were exposed to the experiment). The lower right (respectively, left) panel reports the cumulative total number of rides during the postexperiment (respectively, pre-experiment) period. Interestingly, one can see from Figure A.4 that the cumulative total spending during the pre-experiment period is not statistically different for riders in the Credit and Control conditions. For example, riders in the Credit condition spent on average 2.99% more relative to the Control condition, but this difference is not statistically significant (i.e., $F(3, 3939) = 1.515$, $p = 0.21$). This pattern is replicated for each time point during the pre-experiment period.[16]

On the other hand, riders in those two conditions show a different engagement behavior after being exposed to the experiment. Starting from the first week after the experiment, riders in the Credit condition are more likely to spend (and to complete rides) relative to riders in the Control condition. In addition, this gap increases over time so that four weeks after being exposed to the experiment, riders in the Credit condition spent on average 11.98% more relative to the Control condition. This difference is statistically significant ($F(3, 3939) = 5.43$, $p < 0.01$). The same pattern is observed across all four weeks after the experiment.[17] As we can see from the lower panels of Figure A.4, the number of rides shows the same consistent pattern.[18]

A.6. Regression Analysis

To complement our analysis, we next run a regression analysis to examine how different compensation types affect riders' engagement. We focus on the total spending and number of rides during the first four weeks after being exposed to the experiment and estimate the following regression equation:

$$\begin{aligned} y_i = {} & \alpha + \beta_1 \text{Comms}_i + \beta_2 \text{Credit}_i + \beta_3 \text{Waived}_i \\ & + \gamma_1 \text{Pre-experiment-Rides}_i \\ & + \gamma_2 \text{Pre-experiment-Rides}_i^2 + \gamma_3 \text{ETA error}_i \\ & + \mu_i + \epsilon_i, \end{aligned} \tag{A.1}$$

where i corresponds to a rider, y_i denotes the dependent variable (for conciseness, we only report the results for the total number of rides), and $\text{Comms}_i, \text{Credit}_i$, and Waived_i represent binary variables for each condition (the Control condition is the reference group). Note that we control for the riders' pre-experiment usage by including

Figure A.3. (Color online) Average Total Spending and Number of Rides by Pre-experiment Usage for Experiment 1

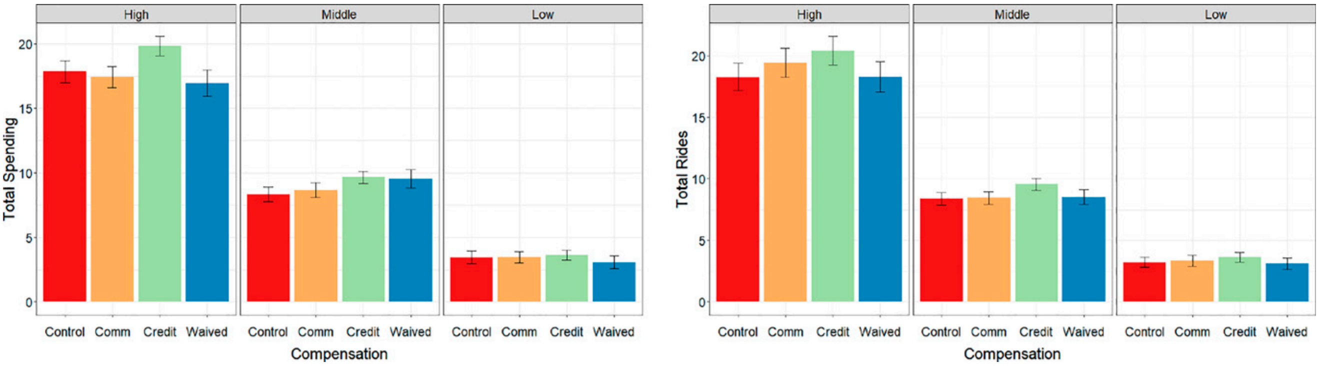
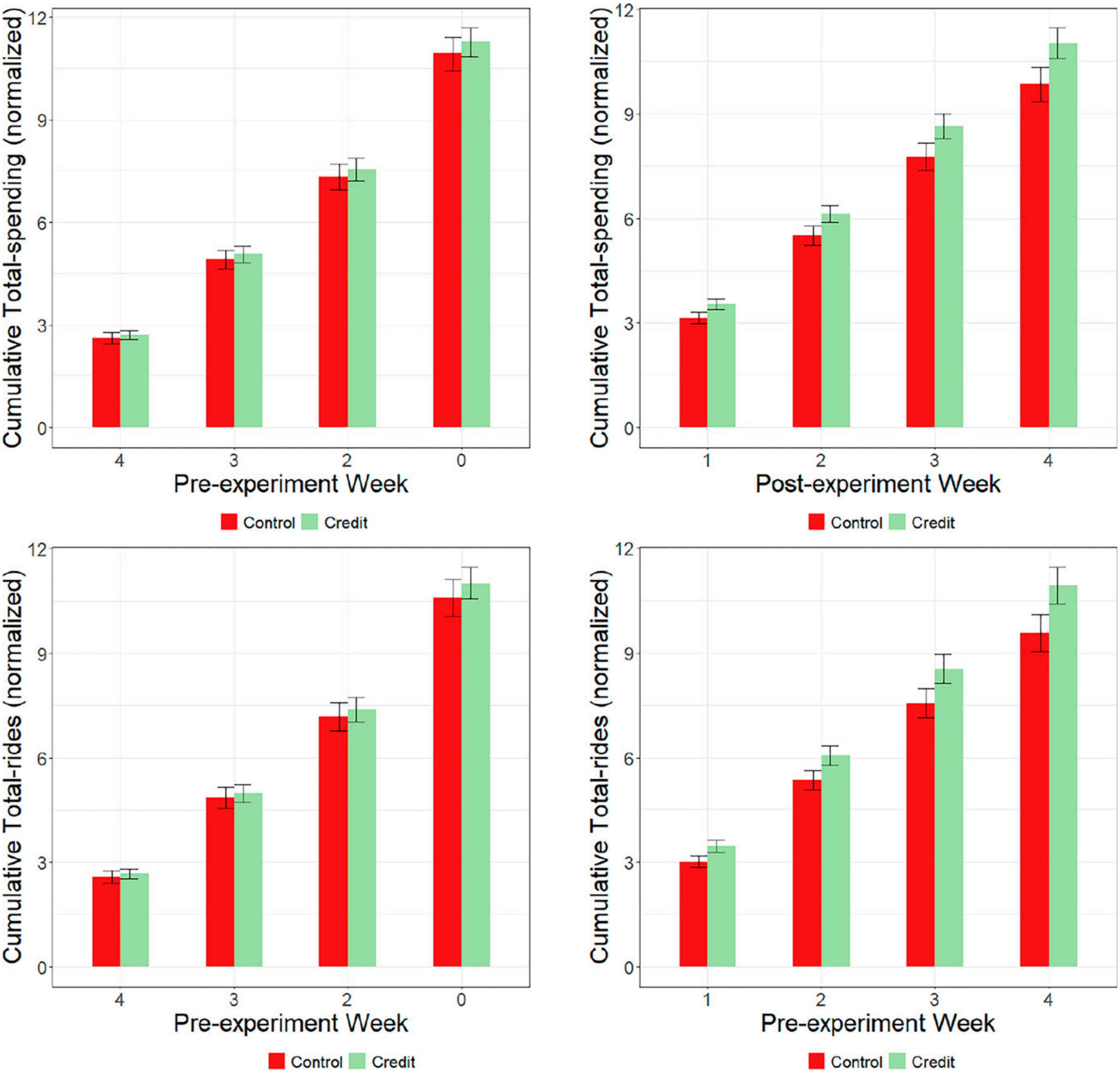


Figure A.4. (Color online) Cumulative Spending and Number of Rides for Experiment 1



Note. Week 0 denotes before the experiment exposure.

the total number of rides during the pre-experiment period (of four weeks) as well as a quadratic term to capture a potential nonlinear effect. This allows us to study how the main effect is moderated by pre-experiment usage. We also include a dummy variable for the ETA error segment, which indicates whether rider i belongs to the ETA error or VGR segment. Finally, we include day exposure fixed effects, which vary at the rider level, by controlling for the date on which rider i was exposed to the experiment (μ_i). Such a variable helps control for unobserved individual heterogeneity. The last term, ϵ_i, is a stochastic error term. The results of the ordinary least squares (OLS) regression are reported in Table A.4. Consistent with our previous findings, the coefficient of the Credit variable is positive and statistically significant. However, the variables for Comms and Waived are not significant. For robustness purposes, we also consider the regression equation with a log transformation in the dependent and independent variables with nonbinary values (this allows us to correct for the skewness of the distribution); see the last two columns of Table A.4. Finally, since the dependent variable can only take positive integer values, we also run a Poisson regression with the same specification as in (A.1) (based on the assumption that the error term ϵ_i follows a Poisson distribution). We find that the results of the Poisson regression are consistent with the OLS regression.

Table A.4. Regression Results for Experiment 1

	Dependent variable		
	Total-rides$_i$	log(1 + Total-rides$_i$)	
	(1)	(2)	(3)
Comms	0.553	0.017	0.006
	(0.419)	(0.035)	(0.034)
Credit	1.179***	0.085**	0.087***
	(0.396)	(0.033)	(0.032)
Waived	0.366	0.039	0.028
	(0.477)	(0.040)	(0.038)
Pre-experiment-rides	0.621***	0.078***	
	(0.018)	(0.002)	
Pre-experiment-rides^2	0.002***	−0.0004***	
	(0.0002)	(0.00002)	
log(Pre-experiment-rides)			0.727***
			(0.011)
ETA	0.132	0.004	0.009
	(0.302)	(0.025)	(0.024)
Exposure fixed effects	Yes	Yes	Yes
Constant	6.493***	1.457***	0.748***
	(1.180)	(0.100)	(0.097)
Observations	3,943	3,943	3,943
R^2	0.596	0.503	0.546

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

We next run a regression with the interaction terms between compensation conditions and frustration types (ETA error versus VGR) to examine the difference in engagement from the two types of frustration. As shown in Table A.5, the results are consistent with the ANOVA tests. In particular, the interaction coefficient between the ETA error and the Credit condition is statistically significant, whereas none of the interaction coefficients between the ETA error segment and the other conditions are significant. As a robustness test, we run the same regression using the Credit condition as the baseline and report the results in Table A.6. We find that the coefficients of Control, Comms, and Waived are negative and statistically significant, implying that riders in the Credit condition are more likely to complete rides relative to riders in other conditions.

Table A.5. Regression Results for Experiment 1 by Frustration Type

	Dependent variable	
	Total-rides	log(1 + Total-rides)
	(1)	(2)
Comms	0.045	−0.001
	(0.563)	(0.048)
Credit	0.656	0.026
	(0.503)	(0.042)
Waived	0.322	0.067
	(0.752)	(0.063)
ETA	−0.656	−0.055
	(0.597)	(0.050)
Comms × ETA	1.126	0.041
	(0.830)	(0.070)
Credit × ETA	1.318*	0.156**
	(0.800)	(0.068)
Waived × ETA	0.370	−0.018
	(0.983)	(0.083)
Pre-experiment-rides	0.621***	0.078***
	(0.018)	(0.002)
Pre-experiment-rides^2	0.002***	−0.0004***
	(0.0002)	(0.00002)
Exposure fixed effects	Yes	Yes
Constant	6.857***	1.491***
	(1.197)	(0.101)
Observations	3,943	3,943
R^2	0.596	0.504

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A.6. Regression Results for Experiment 1 for ETA Error (Baseline: Credit)

	Dependent variable	
	ETA error only (baseline: Credit)	
	Total-rides$_i$	log(1 + Total-rides$_i$)
	(1)	(2)
Control	−1.869***	−0.183***
	(0.716)	(0.056)
Comms	−1.370**	−0.139***
	(0.689)	(0.054)
Waived	−1.450**	−0.128**
	(0.707)	(0.055)
Pre-experiment-rides		0.074***
		(0.002)
Pre-experiment-rides^2		−0.0004***
		(0.00003)
Exposure fixed effects	No	Yes
Constant	11.037***	1.566***
	(0.496)	(0.134)
Observations	1,764	1,764
R^2	0.004	0.484

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

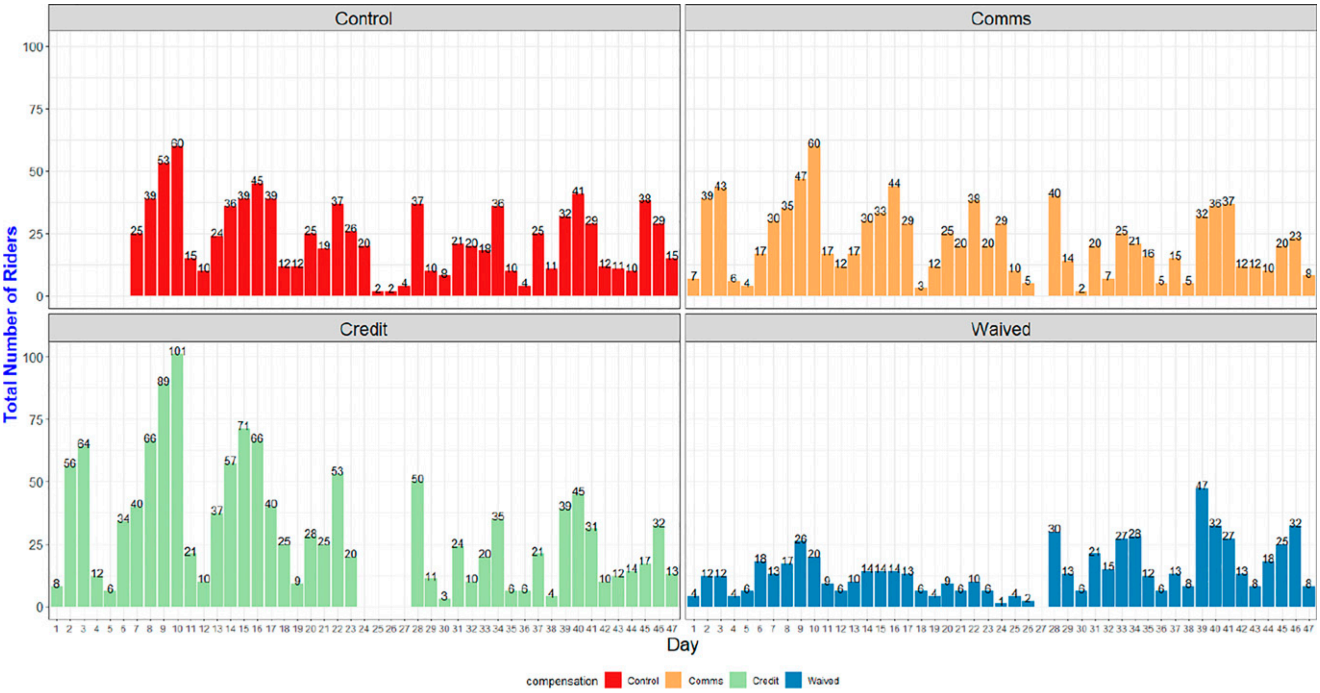
Last, we run a separate regression analysis for each group of riders (depending on their pre-experiment usage). Consistent with the one-way ANOVA tests, we create three groups: high, middle, and low, based on the top 30%, top 30%–70%, and bottom 30% of the distribution of the pre-experiment usage (i.e., the total spending). We also used the number of rides instead of the spending and observed the same qualitative results. As we can see from Table A.7, the coefficient of the Credit condition is significant only for the high and middle groups but not for the low group. This result confirms once again that (i) the effect of the Credit condition on the frustration is moderated by pre-experiment usage, and that (ii) this effect is only present for frequent riders. Note that there is no statistical difference between the high and middle groups in terms of the magnitude of the Credit condition coefficient.

A.7. Randomization Check

We next provide the following additional details related to the randomization procedure:

1. Explaining the reasons why the number of riders in each condition is different.

Figure A.5. (Color online) Total Number of Riders in the Different Conditions over Time



2. Running additional analyses to showcase the robustness of our results to the fact that the number of riders is different in each condition.

3. Reporting several plots to confirm that users in the different conditions are well balanced across several important dimensions. This provides reassurance that our assignment was randomized.

Table A.7. Regression Results for Experiment 1 by Pre-experiment Usage

	Dependent variable		
	Total-rides$_i$		
	High (1)	Middle (2)	Low (3)
Comms	1.238	0.552	−0.058
	(0.999)	(0.558)	(0.514)
Credit	2.076**	1.488***	0.509
	(0.924)	(0.527)	(0.483)
Waived	0.266	0.631	−0.425
	(1.128)	(0.638)	(0.581)
Pre-experiment-rides	0.252***	0.649*	0.390
	(0.062)	(0.387)	(0.646)
Pre-experiment-rides^2	0.005***	0.009	0.100
	(0.0005)	(0.016)	(0.109)
ETA	−0.408	0.392	0.388
	(0.721)	(0.407)	(0.375)
Constant	11.542***	1.715	2.111**
	(1.620)	(2.156)	(0.880)
Observations	1,236	1,529	1,178
R^2	0.475	0.154	0.051

Notes. We do not include exposure fixed effects in this regression. Indeed, since we split the sample into three groups, exposure fixed effects will capture all the variation in the dependent variable.

**p* < 0.1; ***p* < 0.05; ****p* < 0.01.

First, we report the total number of riders in each condition and frustration type: for ETA error: Control: 410, Comms: 478, Credit: 445, Waived: 431; for VGR: Control: 551, Comms: 514, Credit: 896, Waived: 218.

See Figure A.5 for the assignment in each day of our first field experiment. As we can see, our assignment suffered from two technical difficulties: (A) the Control condition (for

Table A.8. First and Second Robustness Tests

	Dependent variable	
	Total-Rides	
	Removing the first week (1)	Random sampling (649 riders in each group) (2)
Comms	0.406	0.539
	(0.419)	(0.533)
Credit	1.220***	1.070**
	(0.400)	(0.518)
Waived	0.509	0.319
	(0.464)	(0.514)
Pre-total-ride	0.609***	0.708***
	(0.057)	(0.083)
Pre-total-ride^2	0.002**	0.0003
	(0.001)	(0.002)
ETA	0.129	−0.203
	(0.311)	(0.376)
Time fixed effects	Yes	Yes
Constant	3.716***	2.381
	(0.851)	(2.649)
Observations	3,591	2,596
R^2	0.586	0.527

***p* < 0.05; ****p* < 0.01.

Table A.9. Third Robustness Test

	Dependent variable		
	Total-Rides		
	ETA (1)	VGR (2)	ETA/VGR with lowest no. (3)
Comms	0.985	0.417	0.469
	(1.033)	(0.834)	(0.713)
Credit	2.273**	−0.127	1.214*
	(1.054)	(0.854)	(0.718)
Waived	0.528	0.340	0.640
	(1.054)	(0.763)	(0.672)
Pre-total-ride	0.497***	0.696***	0.534***
	(0.059)	(0.059)	(0.057)
Pre-total-ride^2	0.004***	0.001	0.003***
	(0.001)	(0.001)	(0.001)
ETA			0.663
			(0.484)
Time fixed effects	Yes	Yes	Yes
Constant	3.151***	1.716*	2.516***
	(1.193)	(1.020)	(0.931)
Observations	705	864	1,437
R^2	0.659	0.594	0.624

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

both ETA error and VGR) did not get assigned any rider during the first week of the experiment, and (B) the Waived condition (only for VGR) did not work properly for the first 28 days (users who were supposed to be assigned to the Waived group were assigned to the Credit condition instead).

To account for these real-world technical issues, we run a series of robustness tests. First, we remove the observations from the first week of our experiment and re-estimate all our econometrics models. This fully addresses issue (A). Second, we subsample our data to ensure that we have the same number of users in each condition and re-estimate our models. Specifically, we select the minimum number of users from each condition and randomly sample the same number of users from the other conditions. Third, we consider using only the data starting from day 29 (so that the experiment length is reduced) and re-estimate our models. This will address both issues (A) and (B). We find the same qualitative results across all three robustness tests, hence strengthening the validity of our results.

Table A.10. Adding Nonfrustrated Riders

	Dependent variable Total-rides$_{it}$
$\mathrm{I(Control)}_i \times \mathrm{I(After\text{-}experiment)}_t$	−0.065***
	(0.015)
$\mathrm{I(Comms)}_i \times \mathrm{I(After\text{-}experiment)}_t$	−0.043***
	(0.015)
$\mathrm{I(Credit)}_i \times \mathrm{I(After\text{-}experiment)}_t$	−0.022
	(0.014)
$\mathrm{I(Waived)}_i \times \mathrm{I(After\text{-}experiment)}_t$	−0.059***
	(0.016)
Time fixed effects	Yes
Rider fixed effects	Yes
Observations	269,720
R^2	0.278

***$p < 0.01$.

After day 28 of the experiment, we have the following number of users in each condition, which appears to me more balanced: for ETA error: Control: 169, Comms: 179, Credit: 186, Waived: 171; for VGR: Control: 248, Comms: 181, Credit: 217, Waived: 218.

As mentioned, we conduct three robustness tests: (1) removing all the observations from the first week, (2) using a random sample with the same number of users in each condition (based on the size of the group with the lowest number of riders), and (3) analyzing riders who were exposed to the experiment only after the fourth week.
Table A.8 shows the results of the first two analyses. After removing the riders who were exposed to the experiment in the first week, we obtain that our main finding still holds: riders in the Credit condition are more likely to ride relative to riders in the Control condition. Similarly, when using a random sample (where each condition includes the same number of 649 riders), we obtain consistent results.

Next, Table A.9 presents the results of the third robustness test (i.e., considering the data only after the fourth week). The first column shows the result for the riders who experienced an ETA error frustration, whereas the second column uses VGR riders. Once again, the Credit condition has a positive and statistically significant coefficient for the ETA error, whereas this effect is found to be insignificant for VGR. Finally, we combine a random sample from the ETA error and VGR conditions. For example, we randomly select 169 riders from each compensation group in the ETA error condition (because the Control condition of ETA error has the lowest number of riders) and randomly select 181 riders from each compensation group in the VGR condition. We then combine both samples. The third column reports the result of this test and shows that the Credit condition has a positive (and significant) impact on engagement relative to other conditions.

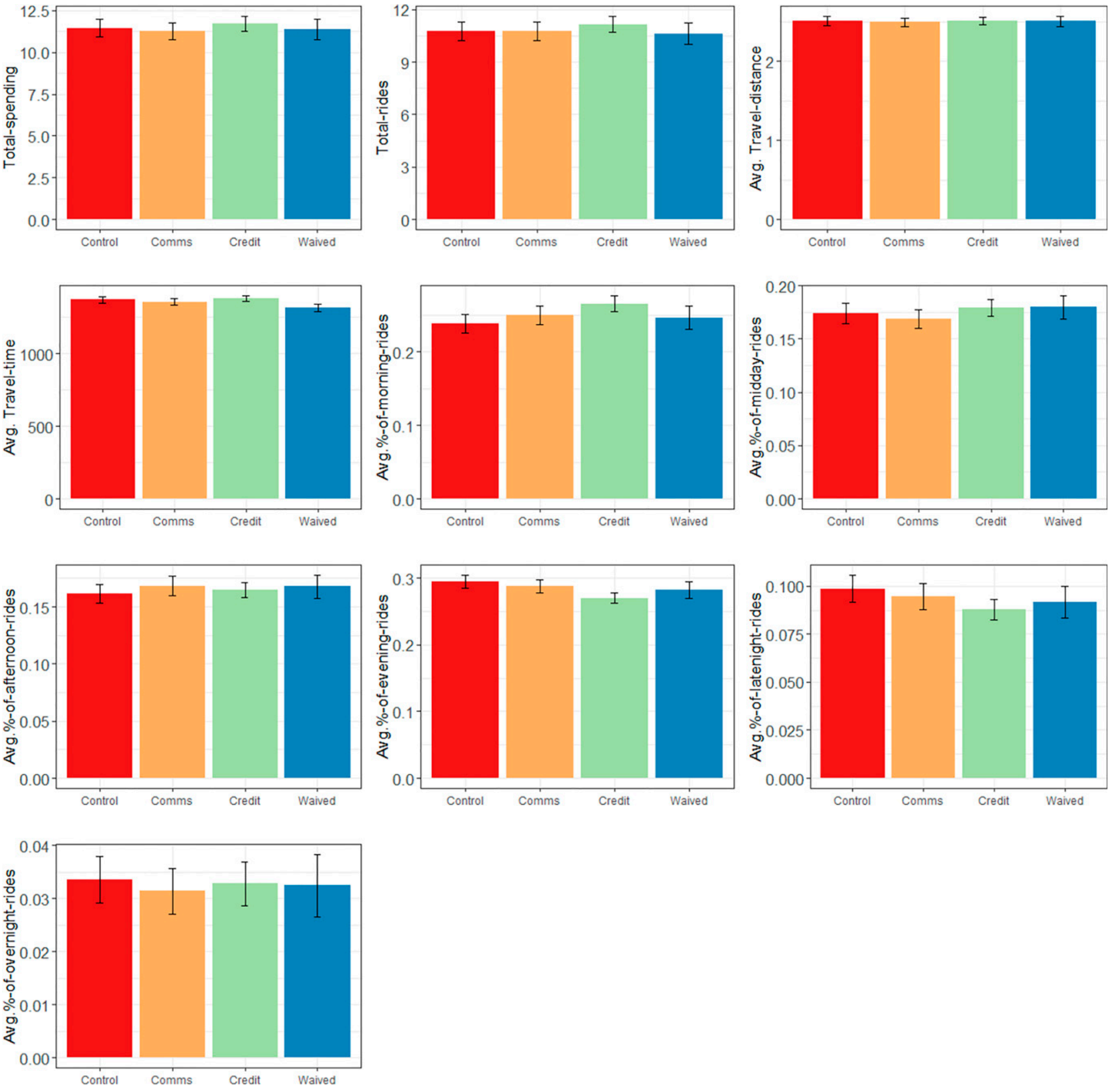
Last, we include nonfrustrated riders (defined in the matching analysis from Section 5.5) and conduct a diff-in-diffs analysis (note that the baseline group is now the nonfrustrated riders, so that the interpretation of the coefficients is relative to nonfrustrated riders). Specifically, we estimate the following specification:

$$\begin{aligned}\text{Total-Rides}_{it} = {} & \beta_1 \cdot \mathrm{I(Control)}_i \times \mathrm{I(After\text{-}experiment)}_t \\ & + \beta_2 \cdot \mathrm{I(Comms)}_i \times \mathrm{I(After\text{-}experiment)}_t \\ & + \beta_3 \cdot \mathrm{I(Credit)}_i \times \mathrm{I(After\text{-}experiment)}_t \\ & + \beta_4 \cdot \mathrm{I(Waived)}_i \times \mathrm{I(After\text{-}experiment)}_t \\ & + \mu_i + \nu_t + \epsilon_{it},\end{aligned}$$

where all the variables are defined in the same way as before.

As shown in Table A.10, riders in the Control, Comms, and Waived conditions are less likely to complete rides in the postexperiment period. This confirms our finding that only the Credit condition affects frustrated riders' engagement. Once again, the Credit condition seems to be indifferent from nonfrustrated riders.

Figure A.6. (Color online) Summary Statistics of Experiment 1 (Pre-experiment)



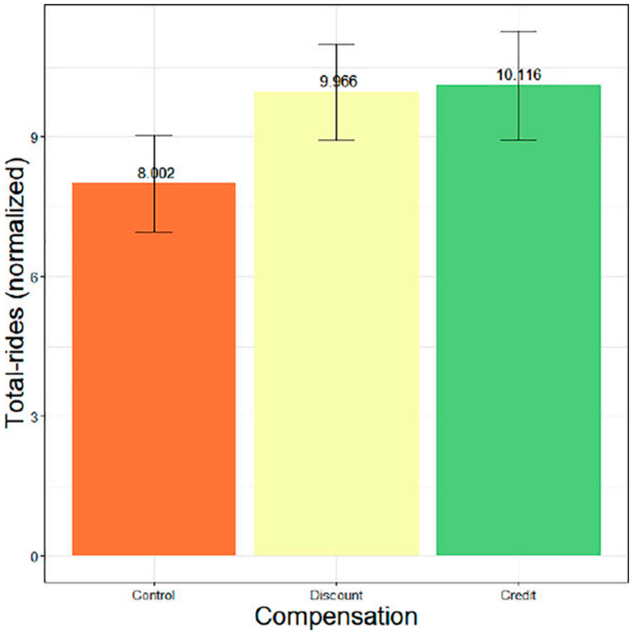
Notes. This figure provides the results of several balancedness tests to showcase the validity of the randomization in our first field experiment. The confidence interval is reported at the 90% level.

A.8. Balancedness Tests

We next present several balancedness tests to showcase the validity of our randomization. We compare the riders in the four conditions (i.e., Control, Comms, Credit, and Waived) along several dimensions. Specifically, we investigate several rider level variables during the 28 days prior to the experiment. We first compute the total spending and number of rides (we normalize these variables by dividing their real values by a constant to avoid revealing sensitive information). We next compute rides' features such as average travel distance, average travel time, and average star rating. Moreover, we examine the time of the rides within the day based on computing the percentage of rides completed in each time slot: morning (6AM–10AM), midday (10AM–2PM), afternoon (2–5PM), evening (5–9PM), late night (9PM–12AM), and overnight (12–6AM).

As shown in Figure A.6, the pre-experiment usage behavior in different compensation groups is balanced (i.e., statistically insignificant), hence implying that our

Figure B.1. (Color online) Average Total Number of Rides for Experiment 2



randomization was well executed. We also compare the summary statistics of the average ETA error in the 28 days prior to the experiment (normalized based on dividing each rider's ETA error by the maximum ETA error value) and find no systematic differences between conditions (omitted for conciseness).

Appendix B. Experiment 2

B.1. Text Messages for Experiment 2

The text message for the Credit condition is the same as in experiment 1. For the Discount condition, we used the following text message: "Hi {first_name} it looks like your trip earlier today took much longer than anticipated. We are so very sorry for any inconvenience this may have caused. Our Routing Team is looking into it. As a token of our apology, your next ride this week is 50% off!"

B.2. ANOVA Results

Figure B.1 reports the average number of rides for each condition during the first four weeks after being exposed to the experiment. We will consider smaller time windows in the sequel. The results of Figure B.1 imply the following:

- Riders in the Credit condition took 26.42% extra rides relative to riders in the Control condition.
- Riders in the Discount condition took 24.57% extra rides relative to riders in the Control condition.
- The Credit condition is not statistically different from the Discount condition.

For the total rides, the result of the one-way ANOVA is significant for all three conditions ($F(2,920) = 3.22$, $p < .05$). In addition, the post hoc comparisons among the different conditions are significant at the 90% level (see Table B.1).[19] These results thus confirm the diff-in-diffs results.

Table B.1. Pairwise Comparisons Between Conditions (for Experiment 2)

	Total-rides	
	Difference	*p*-value
Discount: Control	1.963	0.031
Credit: Control	2.114	0.025
Credit: Discount	0.151	0.871

B.3. Pre-Experiment Usage as a Moderator

We next analyze how the main effect is moderated by pre-experiment usage. We divide the riders into two groups: high and low, based on their pre-experiment usage. Since we have a smaller number of riders in experiment 2, we only use two groups instead of three. To remain consistent, we compute the total spending and number of rides for each rider during the four weeks prior to the experiment. We then define the high and low groups by using a threshold from the top 75%. The results are presented in Figure B.2.

We find that the difference in average number of rides between the Credit and Control conditions is statistically significant only for the high group. For the total number of rides: high group: $F(2,625) = 6.079$, $p < .01$, and low group: $F(2,292) = 0.447$, $p = 0.64$ (see Table B.2 for more details on the pairwise comparisons).

As before, we infer that the level of pre-experiment usage does affect the impact of compensating frustrated riders. Specifically, we have:

1. Infrequent users who experienced a frustration are not affected (statistically) by receiving a promotion aimed to compensate their frustration.
2. For the high group, it is profitable to offer either a $5 credit or a 50% discount for the next ride following the frustration.
3. For the most frequent riders (more than three rides during the past four weeks), we obtain an average of 31.54% (respectively, 42.35%) additional rides between the Credit (respectively, Discount) and Control conditions.

B.4. Varying the Time Window and Considering Pre-Experiment Usage

As in experiment 1, we compare riders' engagement during pre- and postexperiment periods.

The right panel in Figure B.3 presents the total number of rides during the four weeks following the experiment exposure. The *x*-axis in the right panel indicates time points (in weeks) between one week and four weeks after the experiment exposure. The left panel reports the total number of rides during the pre-experiment period (here, the *x*-axis indicates time points starting from four weeks prior to the experiment until the date on which riders were exposed to the experiment).

Interestingly, we can see from Figure B.3 that the total number of rides during the pre-experiment period is not statistically different for riders in the Credit (or Discount) and Control conditions. For example, riders in the Credit (respectively, Discount) condition completed on average 8.27% (respectively, 1.68%) more rides relative to riders in the Control condition, but this difference is not statistically significant ($F(2,920) = 0.425$, $p = 0.654$). This pattern is

Figure B.2. (Color online) Average Total Number of Rides by Pre-experiment Usage for Experiment 2

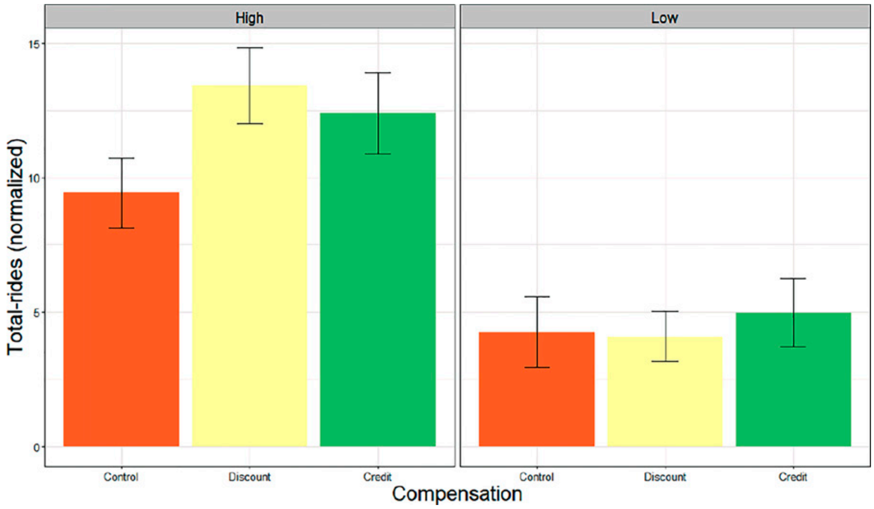
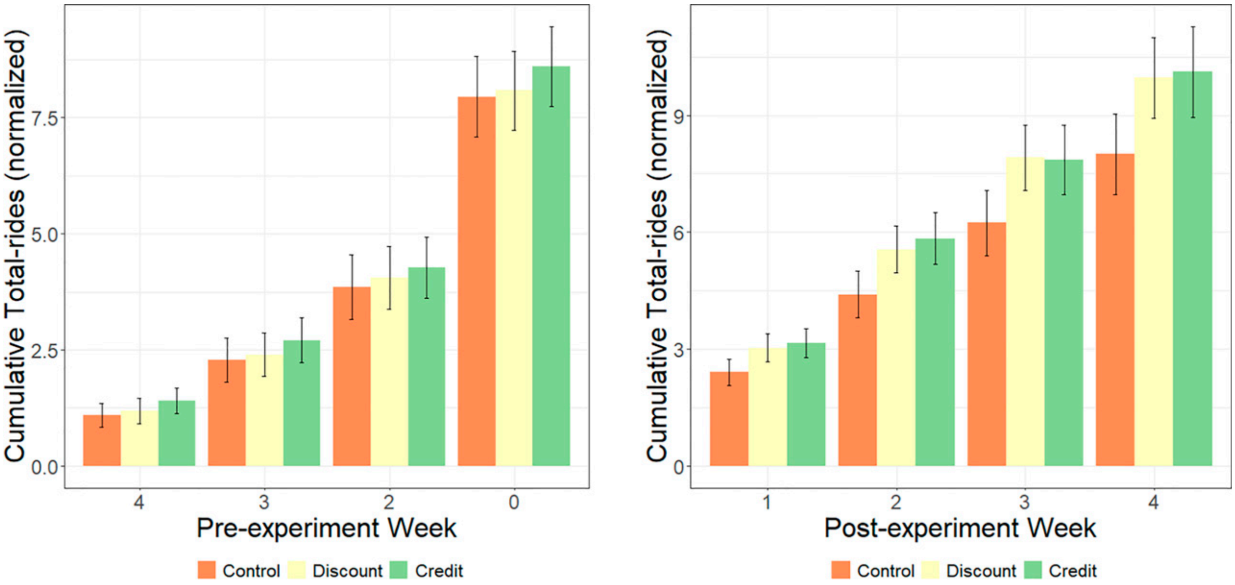


Table B.2. Pairwise Comparisons Between Conditions for High and Low Usage Groups (for Experiment 2)

	Total-rides			
	High		Low	
	Difference	*p*-value	Difference	*p*-value
Discount: Control	3.992	0.002	−0.158	0.986
Credit: Control	2.973	0.038	0.721	0.771
Credit: Discount	−1.019	0.684	0.879	0.631

Figure B.3. (Color online) Cumulative Number of Rides for Experiment 2



Note. Week 0 denotes before the experiment exposure.

Table B.3. Regression Results for Experiment 2

	Dependent variable		
	Total-rides$_i$	log(1 + Total-rides$_i$)	
	(1)	(2)	(3)
Discount	1.193**	0.229***	0.273***
	(0.478)	(0.072)	(0.071)
Credit	1.185**	0.188**	0.191**
	(0.509)	(0.077)	(0.076)
Pre-experiment-rides	0.410***	0.066***	
	(0.036)	(0.005)	
Pre-experiment-rides^2	−0.002***	−0.0005***	
	(0.001)	(0.0001)	
log(Pre-experiment-rides)			0.494***
			(0.028)
Exposure fixed effects	Yes	Yes	Yes
Constant	2.251*	0.815***	0.436**
	(1.215)	(0.184)	(0.185)
Observations	923	923	923
R^2	0.349	0.291	0.306

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

replicated for each point in time during the pre-experiment period.[20] On the other hand, riders in those three conditions show different engagement patterns after being exposed to the experiment. Starting from the first postexperiment week, riders in the Credit and Discount conditions are more likely to spend (and to complete rides) relative to riders in the Control condition. This gap increases over time so that four weeks after being exposed to the experiment, riders in the Credit (respectively, Discount) condition completed on average 26.43% (respectively, 24.57%) more rides relative to riders in the Control condition ($F(2, 920) = 3.22$, $p < .05$); see Table B.1 for more details on the pairwise comparisons. The same pattern is observed across each one of the four weeks after the experiment.[21]

Table B.4. Regression Results for Experiment 2 by Pre-experiment Usage

	Dependent variable		
	Total-rides		log(1 + Total-rides)
	High	Low	Low
	(1)	(2)	(3)
Discount	1.833***	−0.045	0.060
	(0.632)	(0.604)	(0.124)
Credit	1.367**	0.429	0.135
	(0.640)	(0.650)	(0.133)
Pre-experiment-rides	0.359***	0.811	
	(0.045)	(0.498)	
Pre-experiment-rides^2	−0.001*		
	(0.001)		
log(Pre-experiment-rides)			0.305**
			(0.147)
Constant	2.537***	1.435	0.716***
	(0.536)	(0.875)	(0.107)
Observations	628	295	295
R^2	0.293	0.012	0.018

Notes. We do not include exposure fixed effects in this regression. Indeed, since we split the sample into three groups, exposure fixed effects will capture all the variation in the dependent variable.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

B.5. Regression Analysis

We investigate whether our main findings continue to hold after controlling for several factors that may affect the engagement behavior. To this end, we run regression analyses. As before, we consider the first four weeks after being exposed to the experiment and estimate the following regression specification:

$$\begin{aligned} y_i = \alpha &+ \beta_1 \text{Discount}_i + \beta_2 \text{Credit}_i \\ &+ \gamma_1 \text{Pre-experiment-Rides}_i \\ &+ \gamma_2 \text{Pre-experiment-Rides}_i^2 + \mu_i + \epsilon_i, \end{aligned} \tag{B.1}$$

where i corresponds to a rider, y_i denotes the dependent variable (for conciseness, we report only the results for the total number of rides), and Discount$_i$, Credit$_i$ represent binary variables for each experiment condition. We control for riders' pre-experiment usage by including the total number of rides during the pre-experiment period (of four weeks) as well as a quadratic term to capture a potential nonlinear effect. Finally, we include exposure fixed effects, which vary at the rider level, by capturing the date on which rider i was exposed to the experiment (μ_i).

The results of the OLS regression are reported in Table B.3. Consistent with the findings from the one-way ANOVA tests, the coefficients of the Credit and Discount variables are positive and statistically significant. For robustness purposes, we also consider the regression equation with a log transformation in the dependent and independent variables with nonbinary values (see the last two columns of Table B.3). Finally, since the dependent variable can only take positive integer values, we run a Poisson regression with the same specification as in (B.1). We find that the results of the Poisson regression are consistent with the OLS regression. As before, the pre-experiment usage level shows a concave pattern.

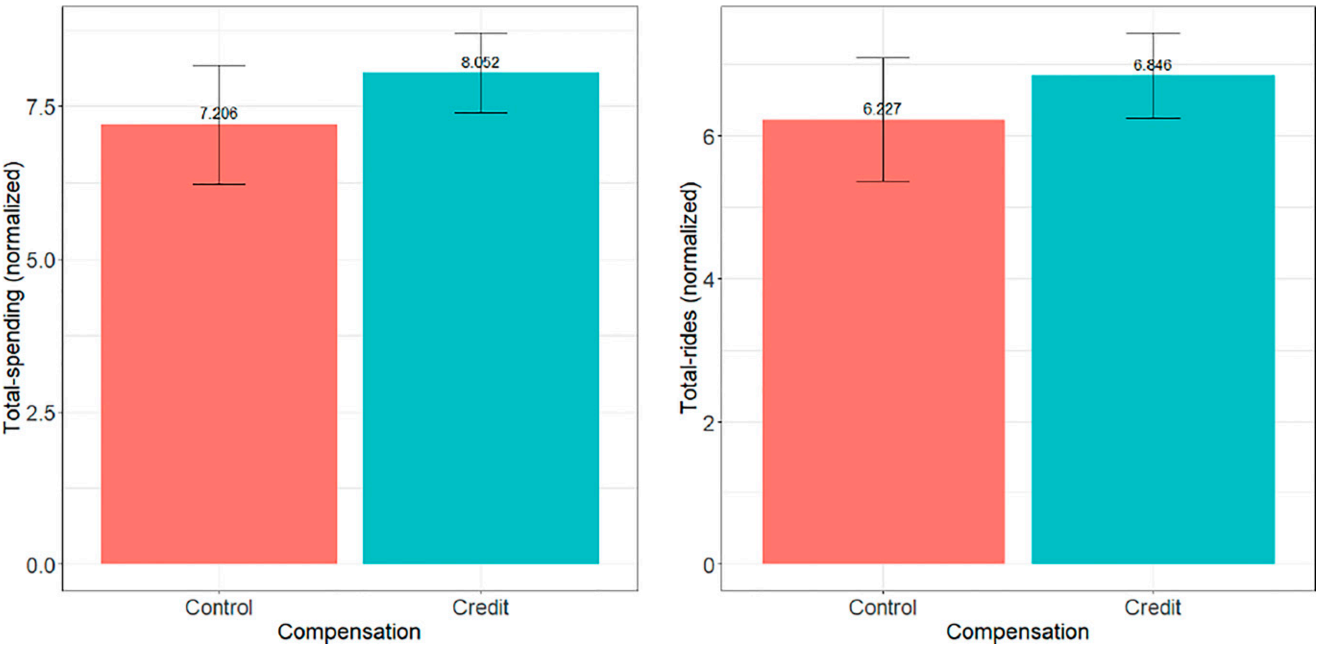
We next run a separate regression analysis for each group of riders (depending on their pre-experiment usage). Consistent with the one-way ANOVA test, we create two groups: high and low, based on the top 50% (i.e., median split) of the distribution of the pre-experiment number of rides.[22] As we can see from Table B.4, the coefficients for the Credit and Discount conditions are significant only for the high group. This result confirms once again that (i) the effect of the Credit and Discount conditions is moderated by pre-experiment usage, and (ii) this effect is only present for frequent riders.

Appendix C. Experiment 3

C.1. Text Messages for Experiment 3

The text message for the Credit condition is the following: "Hi {first_name} happy Via-versary! It's been exactly {years} year since you signed up—we've added $5 to your account to thank you for being part of the Via community!"

Figure C.1. (Color online) ANOVA Results for Experiment 3



C.2. ANOVA Results

As we can see from Figure C.1, riders in the Credit condition take on average 9.94% more rides relative to riders in the Control condition during the first four weeks after being exposed to the experiment. However, this difference is not statistically significant ($t(597) = -.97$, $p = .33$). The same result applies to the total spending. This implies that offering a compensation to nonfrustrated riders (in this case, riders who celebrate their joining date anniversary) does not seem to have a significant impact on their engagement behavior. For robustness purposes, we vary the time window from four weeks to one week by increments of one week and observe consistent results. We next conduct the same analysis by splitting the riders into high and low groups based on their pre-experiment usage. We divide the riders into two groups by using the median split (we also vary this number up to the top 25% by increments of 5%). As before, we observe that the Control and Credit conditions are not statistically different from each other neither for the high group nor for the low group (the results are very similar to Figure C.1 and are omitted for conciseness).

Appendix D. Text Messages

“T\hat’s a really good approach to taking care of customers and ensuring satisfaction and loyalty . . . Thank you!”

“Wow! Thank you! You folks are really great! You have repeatedly earned my loyalty and gratitude by the way you conduct your business. Please keep it up. And again, thank you.”

“Thank you! This is why I continue to do business with you. Excellent customer service.”

“Via is Awesome! Thank you very much for that consideration. That is very considerate of you and your staff. Much appreciated.”

“How nice, thank you. Via is the best!! Only service I use.”

“Wow! This is awesome customer service. Thanks for taking the initiative and reaching out to me.”

“You are an amazing company—I rave about Via every chance I get and here is just another example!”

“Wow! That is really sweet. I really appreciate your customer service and LOVE Via. Thank you and Merry Christmas!”

“Awwwwwww thanks so much. Now I am going to keep recommending Via.”

“Definitely makes me want to take Via more frequently.”

“Thank you! Am impressed that you could notice this and then compensate!”

“You guys once again prove how awesome you are. I actually just recommended you to a friend I met at the bar.”

Endnotes

[1] Customer complaints is a topic of active media coverage, see, for example, https://www.nytimes.com/2017/06/15/smarter-living/consumer-complaint-writing-letter.html.

[2] See https://www.dominos.co.nz/inside-dominos/technology/delivery-guarantee.

[3] See https://www.dealnews.com/features/Amazon-and-Walmart-Will-Offer-Store-Credit-for-Late-Christmas-Deliveries/944691.html.

[4] See https://www.macrumors.com/2016/09/15/best-buy-delays-iphone-7-plus-orders/.

[5] See https://wap.ceo.ca/@newswire/navya-partners-with-via-to-introduce-a-revolutionary.

[6] We cannot reveal the exact details of our data set due to confidentiality reasons. However, such information has no impact on our analysis and our key findings.

[7] For instance, this set of riders includes riders who subscribe to the ViaPass service, which is a package that allows riders to use a large number of monthly rides for a fixed charge. Since such riders are not charged on a per-ride basis, we decided not to include them in our analysis.

[8] For the VGR frustration, we also restrict the actual ride time to be at least 20 minutes (to avoid short rides for which the VGR is not an appropriate frustration measure).

[9] In our data set, the average price of a ride in NYC is $5.62, and the vast majority of rides cost exactly $5.

[10] Similarly, for the total number of rides, the result of the one-way ANOVA is statistically significant for all four conditions ($F(3,3939) = 3.98$, $p < 0.01$), and the pairwise comparison among the different conditions is statistically significant at the 95% level.

[11] We thank one of the anonymous reviewer for suggesting this great test that strengthens our findings.

[12] We acknowledge the smaller sample size of observations in this experiment. Repeating this type of experiment in a larger scale is left for future research.

[13] Total spending: ETA error segment: $F(3,1760) = 4.41$, $p < 0.01$ and VGR segment: $F(3,2175) = 2.04$, $p = .11$. Total number of rides: ETA error segment: $F(3,1760) = 2.624$, $p < 0.05$ and VGR segment: $F(3,2175) = 1.498$, $p = 0.213$.

[14] For the total number of rides, only the difference between the Credit and Control conditions is statistically significant. However, we found significant effects between the Credit and each other condition in the regression analysis (see details in Table A.5).

[15] Total spending: high group: $F(3,1178) = 6.773$, $p < 0.01$; middle group: $F(3,1570) = 3.499$, $p = 0.02$; and low group: $F(3,1183) = .627$, $p > 0.1$. Total number of rides: high group: $F(3,1145) = 2.124$, $p = 0.096$; middle group: $F(3,1478) = 3.384$, $p = 0.02$; and low group: $F(3,1308) = .83$, $p = 0.48$.

[16] For three weeks: $F(3,3939) = 0.502$, $p = 0.68$; for two weeks: $F(3,3939) = 0.925$, $p = 0.428$; and for one week: $F(3,3939) = 1.371$, $p = 0.25$.

[17] For three weeks: $F(3,3939) = 5.17$, $p < .01$; for two weeks: $F(3,3939) = 5.19$, $p < .01$; and for one week: $F(3,3939) = 6.14$, $p < .01$.

[18] Postexperiment data: for four weeks: $F(3,3939) = 4.25$, $p < .01$; for three weeks: $F(3,3939) = 3.629$, $p = .01$; for two weeks: $F(3,3939) = 3.715$, $p = .01$; and for one week: $F(3,3939) = 5.092$, $p < 0.01$. Pre-experiment data: for four weeks: $F(3,3939) = 0.253$, $p = 0.859$; for three weeks: $F(3,3939) = 0.372$, $p = 0.77$; for two weeks: $F(3,3939) = 0.48$, $p = 0.70$; and for one week: $F(3,3939) = 0.721$, $p = 0.54$.

[19] Similarly, for the total number of rides, the result of the one-way ANOVA is also significant ($F(2,920) = 3.04$, $p < 0.05$).

[20] For three weeks: $F(2,920) = 0.249$, $p = 0.78$; for two weeks: $F(2,920) = 0.558$, $p = 0.572$; and for one week: $F(2,920) = 0.918$, $p = 0.4$.

[21] For three weeks: $F(2,920) = 3.994$, $p < 0.05$; for two weeks: $F(2,920) = 3.994$, $p < 0.05$; and for one week: $F(2,920) = 3.369$, $p < 0.05$. The results of the pairwise comparisons are same as the results for four weeks, meaning that the differences are significant between Credit and Control and Discount and Control but not between Credit and Discount.

[22] We also checked the results when using the total spending and observed the same qualitative results.

References

Aaker J, Fournier S, Brasel SA (2004) When good brands do bad. *J. Consumer Res.* 31(1):1–16.

Abeler J, Calaki J, Andree K, Basek C (2010) The power of apology. *Econom. Lett.* 107(2):233–235.

Anderson SW, Baggett LS, Widener SK (2009) The impact of service operations failures on customer satisfaction: Evidence on how failures and their source affect what matters to customers. *Manufacturing Service Oper. Management* 11(1):52–69.

Anderson EW, Fornell C, Lehmann DR (1994) Customer satisfaction, market share, and profitability: Findings from Sweden. *J. Marketing* 58(3):53–66.

Andreyeva T, Long MW, Brownell KD (2010) The impact of food prices on consumption: A systematic review of research on the price elasticity of demand for food. *Amer. J. Public Health* 100(2):216–222.

Angrist JD, Pischke JS (2008) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, NJ).

Berry LL, Parasuraman A (2004) *Marketing Services: Competing Through Quality* (Simon and Schuster, New York).

Bitner MJ (1990) Evaluating service encounters: The effects of physical surroundings and employee responses. *J. Marketing* 54(2):69–82.

Bitner MJ, Booms BH, Tetreault MS (1990) The service encounter: Diagnosing favorable and unfavorable incidents. *J. Marketing* 54(1):71–84.

Bolton RN, Grewal D, Levy M (2007) Six strategies for competing through service: An agenda for future research. *J. Retailing* 83(1):1–4.

Buell RW, Kim T, Tsay CJ (2016) Creating reciprocal value through operational transparency. *Management Sci.* 63(6):1673–1695.

Chen SFS, Monroe KB, Lou YC (1998) The effects of framing price promotion messages on consumers' perceptions and purchase intentions. *J. Retailing* 74(3):353–372.

Cohen P, Hahn R, Hall J, Levitt S, Metcalfe R (2016) Using big data to estimate consumer surplus: The case of Uber. Technical report, National Bureau of Economic Research, Cambridge, MA.

Cohen MC, Leung NHZ, Panchamgam K, Perakis G, Smith A (2017) The impact of linear optimization on promotion planning. *Oper. Res.* 65(2):446–468.

Craighead CW, Karwan KR, Miller JL (2004) The effects of severity of failure and customer loyalty on service recovery strategies. *Production Oper. Management* 13(4):307–321.

Cui R, Zhang DJ, Bassamboo A (2019) Learning from inventory availability information: Evidence from field experiments on Amazon. *Management Sci.* 65(3):1216–1235.

Davis S, Inman JJ, McAlister L (1992) Promotion has a negative effect on brand evaluations—Or does it? Additional disconfirming evidence. *J. Marketing Res.* 29(1):143–148.

De Cremer D, Pillutla MM, Folmer CR (2011) How important is an apology to you? Forecasting errors in evaluating the value of apologies. *Psych. Sci.* 22(1):45–48.

Fisher M, Gallino S, Li J (2018) Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management Sci.* 64(6):2496–2514.

Gallino S, Moreno A (2018) The value of fit information in online retail: Evidence from a randomized field experiment. *Manufacturing Service Oper. Management* 20(4):767–787.

Grewal D, Levy M (2009) Emerging issues in retailing research. *J. Retailing* 85(4):522–526.

Grewal D, Roggeveen AL, Tsiros M (2008) The effect of compensation on repurchase intentions in service recovery. *J. Retailing* 84(4):424–434.

Halperin B, Ho B, List JA, Muir I (2019) Toward an understanding of the economics of apologies: Evidence from a large-scale natural field experiment. NBER Working Paper No. w25676, National Bureau of Economic Research, Cambridge, MA.

Ho B (2012) Apologies as signals: With evidence from a trust game. *Management Sci.* 58(1):141–158.

Hoch SJ, Kim BD, Montgomery AL, Rossi PE (1995) Determinants of store-level price elasticity. *J. Marketing Res.* 32(1):17–29.

Hoffman KD, Kelley SW, Chung BC (2003) A CIT investigation of servicescape failures and associated recovery strategies. *J. Services Marketing* 17(4):322–340.

Kalwani MU, Yim CK (1992) Consumer price and promotion expectations: An experimental study. *J. Marketing Res.* 29(1):90–100.

Kelley SW, Hoffman KD, Davis MA (1993) A typology of retail failures and recoveries. *J. Retailing* 69(4):429–452.

Kohavi R, Deng A, Frasca B, Walker T, Xu Y, Pohlmann N (2013) Online controlled experiments at large scale. *Proc. 19th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM), 1168–1176.

Lehtimäki AV, Monroe KB, Somervuori O (2019) The influence of regular price level (low, medium, or high) and framing of discount (monetary or percentage) on perceived attractiveness of discount amount. *J. Revenue Pricing Management* 18(1):76–85.

Maxwell SE, Delaney HD, Kelley K (2017) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, vol. 1. (Routledge, England, UK).

McKechnie S, Devlin J, Ennew C, Smith A (2012) Effects of discount framing in comparative price advertising. *Eur. J. Marketing* 46(11):1501–1522.

Mittal V, Kamakura WA (2001) Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *J. Marketing Res.* 38(1):131–142.

Nevo A, Wolfram C (2002) Why do manufacturers issue coupons? An empirical analysis of breakfast cereals. *RAND J. Econom.* 33(2):319–339.

Ohtsubo Y, Matsunaga M, Himichi T, Suzuki K, Shibata E, Hori R, Umemura T, Ohira H (2020) Costly group apology communicates a group's sincere "intention". *Soc. Neurosci.* 15(2):244–254.

Ohtsubo Y, Watanabe E, Kim J, Kulas JT, Muluk H, Nazar G, Wang F, Zhang J (2012) Are costly apologies universally perceived as being sincere? A test of the costly apology-perceived sincerity relationship in seven countries. *J. Evolutionary Psych.* 10(4):187–204.

Parasuraman A, Berry LL, Zeithaml VA (1991) Understanding customer expectations of service. *Sloan Management Rev.* 32(3):39–48.

Parasuraman A, Zeithaml VA, Berry LL (1985) A conceptual model of service quality and its implications for future research. *J. Marketing* 49(4):41–50.

Reimers I, Xie C (2019) Do coupons expand or cannibalize revenue? Evidence from an e-market. *Management Sci.* 65(1):286–300.

Roggeveen AL, Tsiros M, Grewal D (2012) Understanding the co-creation effect: When does collaborating with customers provide a lift to service recovery? *J. Acad. Marketing Sci.* 40(6):771–790.

Schweitzer ME, Hershey JC, Bradlow ET (2006) Promises and lies: Restoring violated trust. *Organ. Behav. Human Decision Processes* 101(1):1–19.

Serpa JC, Krishnan H (2018) The impact of supply chains on firm-level productivity. *Management Sci.* 64(2):511–532.

Singh J, Teng N, Netessine S (2019) Philanthropic campaigns and customer behavior: Field experiments on an online taxi booking platform. *Management Sci.* 65(2):913–932.

Skarlicki DP, Folger R, Gee J (2004) When social accounts backfire: The exacerbating effects of a polite message or an apology on reactions to an unfair outcome 1. *J. Appl. Soc. Psych.* 34(2):322–341.

Smith AK, Bolton RN (1998) An experimental investigation of customer reactions to service failure and recovery encounters: Paradox or peril? *J. Service Res.* 1(1):65–81.

Smith AK, Bolton RN, Wagner J (1999) A model of customer satisfaction with service encounters involving failure and recovery. *J. Marketing Res.* 36(3):356–372.

Taylor S (1994) Waiting for service: The relationship between delays and evaluations of service. *J. Marketing* 58(2):56–69.

Tsiros M, Mittal V, Ross WT Jr (2004) The role of attributions in customer satisfaction: A reexamination. *J. Consumer Res.* 31(2): 476–483.

Walster E, Berscheid E, Walster GW (1973) New directions in equity research. *J. Personality Soc. Psych.* 25(2):151–176.

Weiner B (1985) An attributional theory of achievement motivation and emotion. *Psych. Rev.* 92(4):548–573.

Winer RS (1986) A reference price model of brand choice for frequently purchased products. *J. Consumer Res.* 13(2):250–256.

Wu M, Teunter RH, Zhu SX (2019) Online marketing: When to offer a refund for advanced sales. *Internat. J. Res. Marketing* 36(3):471–491.

Zahorik AJ, Rust RT (1992) Modeling the impact of service quality on profitability: A review. *Adv. Services Marketing Management* 1(1):247–276.

Zeithaml VA, Berry LL, Parasuraman A (1996) The behavioral consequences of service quality. *J. Marketing* 60(2):31–46.

Zhang DJ, Dai H, Dong L, Qi F, Zhang N, Liu X, Liu Z (2017) How does dynamic pricing affect customer behavior on retailing platforms? Evidence from a large randomized experiment on Alibaba. Working paper, Olin Business School, Washington University in St. Louis.