Frustration-based Promotions: Field Experiments in Ride-Sharing

Maxime C. Cohen
NYU Stern School of Business, New York, NY 10012, maxime.cohen@stern.nyu.edu

Michael D. Fiszer
Via Transportation Inc., New York, NY 10013, michael@ridewithvia.com

Baek Jung Kim
NYU Stern School of Business, New York, NY 10012, bkim2@stern.nyu.edu

The service industry has become increasingly competitive. One of the main drivers for increasing profits and market share is quality of service. When a consumer encounters a bad experience, or a *frustration*, s/he may be tempted to stop using the service (or at least to decrease its usage) and switch to the competitor(s). In collaboration with one of the leading ride-sharing platforms, Via, our goal is to understand the benefits of proactively compensating customers who have experienced a frustration. Motivated by historical data, we considered two types of frustrations: long waiting times and long travel times. In this paper, we examine whether a firm should proactively send compensation to users who have experienced a frustration. To this end, we designed and ran three field experiments to investigate how different types of compensation affect the engagement behavior of riders who experienced a frustration. We conducted ANOVA tests, regression analyses, and used a difference-in-differences approach to support our findings. We observed that sending a compensation to frustrated riders (i) is profitable and boosts their engagement behavior (relative to not sending a compensation), (ii) works well for long waiting times but not for long travel times, (iii) is more effective than sending the same offer to non- frustrated riders, and (iv) has an impact that is moderated by past usage frequency. We also observed that the best strategy is to send credit for future usage (as opposed to waiving the charge or sending an apologetic message). We performed several robustness checks to showcase the consistency of our results in two markets, in two different time periods, and on several metrics.

Key words: Ride-sharing, Promotions, Field Experiments, Service Operations

1. Introduction

In an economy where customers have access to a large amount of information and can compare alternative services, how can a company keep a customer from straying to competing firms? For services with repeat customers, such as hotels, car rentals, 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 at any given moment. Online platforms report product reviews and comparisons posted by past users. Furthermore, in competitive markets, prices are roughly the same, and therefore the main distinction lies in other factors, such as quality of

service or a positive feeling towards the brand. Therefore, to increase their market share, businesses are constantly seeking 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 or text, by calling the customer service hotline, or even 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 business 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 his/her 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 some 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 depending on the context, the relationship with the customer, and their competitive advantage.

It may not be easy to detect the various potential frustrations experienced by consumers 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 that 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

[1] Customer complaints is a topic of active media coverage. A large number of resources even provide guidance on how and when to complain, when to use email and when to use social media, how to request the right level of compensation, and so forth (see, e.g., https://www.nytimes.com/2017/06/15/smarter-living/consumer-complaint-writing-letter.html, http://www.getrichslowly.org/2011/08/23/disatisfied-customer-make-an-effective-complaint/, https://www.thebalance.com/how-to-handle-customer-complaints-2888421).

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

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

pre-ordered the iPhone 7 or iPhone 7 Plus in September 2016. Several disappointed customers who experienced delays in shipments of their smartphones were proactively offered a $100 discount on their next purchase at Best Buy.[4]

How can we design the process of sending targeted proactive compensations to customers experiencing 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 ride-sharing context, customers typically interact with several online platforms to request on-demand transportation services. In recent years, this means of transportation seriously disrupted the industry. In the U.S., several companies compete for market share by offering this type of service: the traditional taxi service, Uber, Lyft, Via, and Gett as well as several other firms. When a customer is ordering a ride, s/he 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 point, boards the vehicle, and is dropped off at her/his destination. In shared services, such as Via, Lyft Line, and UberPOOL, the vehicle's route may be modified to a small or large extent so as to stop to pick up and drop off other passengers. Customers may thus experience several types of frustrations, such as delays/long waiting times, a higher than expected number of stops, significant detours, limited space in the car, poor service interaction with the driver, etc.

In this paper, we collaborate with one of the leading ride-sharing platforms, Via (some background on the company can be found in Section 3.1). We designed and ran three field experiments to study the impact of proactively compensating riders who experienced a frustration.

1.1. Contributions

Given the popularity of ride-hailing and 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 contributions can be summarized as follows.

- **Discussing two frustration types and three engagement metrics in the context of 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 2.1). To measure the riders' engagement behavior, we compute the total spending, the total number of rides, and the time interval between rides (using different time windows).
- **Analyzing the impact of different ways of compensating frustrated riders.** Our first field experiment (run in NYC) uses four different compensation conditions: Control, Comms,

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

Credit, and Waived (details on each condition can be found in Section 2.2). We observed that the Credit condition is significantly different from the other conditions. Namely, proactively offering a $5 credit to frustrated riders boosts their engagement behavior relative to not offering a compensation. In addition, we find that offering a $5 credit to frustrated riders is revenue positive for the service provider. 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., a positive impact on rider engagement, is significant for long waiting times but not for long travel times. Next, we study how the main effect is moderated by the pre-experiment usage. We find that the effect is significant for frequent and intermediate riders but not for new riders. Furthermore, the relative impact is the highest for intermediate riders.
- **Testing the robustness of our findings.** We run a second experiment in a different market (Washington D.C.) at a different time period to test the robustness of our findings. We observe similar qualitative insights on the impact of compensating frustrated riders. In addition, we consider a new condition (called Discount) in which frustrated riders receive a 50% discount on their next ride. This allows us to (partially) test the price sensitivity of our compensation campaign.
- **Compensating frustrated riders versus non-frustrated riders.** Instead of sending a compensation to riders who experienced a frustration, in our third experiment, we consider sending a reward to "random" users. In particular, 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. We find that it is more effective to allocate the compensations to riders who experienced a frustration (relative to riders who did not).

1.2. Related Literature

This paper is related to several streams of literature: pricing decisions in ride-hailing and field experiments in online platforms as well as several topics from the marketing literature (targeted promotions, impact of service quality on consumers, and consumer psychology research).

Pricing and operational decisions in the ride-hailing market: The recent popularity of ride-hailing platforms triggered a great interest in studying pricing decisions in this context. Several papers consider the problem of designing the right incentives on prices and wages to coordinate supply and demand for on-demand platforms (see, e.g., Tang et al. 2016, Taylor 2016, Hu and Zhou 2017, Cohen and Zhang 2017, Bimpikis et al. 2016, Benjaafar et al. 2017, 2018). Our work also

considers price incentives in the context of ride-hailing (more specifically, ride-sharing) but admits several key differences. First, our study is not related to coordinating supply and demand. Instead, we focus on sending targeted compensations to riders who have experienced poor service quality. Second, the focus of our paper is empirical as our results are based on analyzing the data obtained from three field experiments. Third, our key metrics are related to rider engagement behavior (such metrics have received limited attention in the academic literature on ride-hailing).

In Chen and Sheldon (2016), the authors analyze Uber data from 25 million trips and show empirically that dynamic wages (owing to surge pricing) can entice drivers to work longer. In Hu and Zhou (2017), the authors study the pricing of an on-demand platform and demonstrate the good performance of a flat-commission contract. More precisely, it is shown that under concave supply curves, the optimal flat-commission contract achieves at least 75% of the optimal profit. Two recent papers, Banerjee et al. (2015) and Cachon et al. (2017), compare the impact of static versus dynamic prices and wages. By assuming that the payout ratio is exogenously given and that customers are heterogeneous, Banerjee et al. (2015) show the good performance of static pricing. Under different modeling assumptions (endogenous payout ratio and homogeneous valuations), Cachon et al. (2017) found that dynamic pricing performs well. Finally, a number of recent papers study empirically the supply side of the ride-hailing market by investigating drivers' data. Examples of such papers include Hall et al. (2017) and Chen et al. (2017).

As mentioned, our paper focuses on the ride-sharing market in which several riders can be assigned to the same vehicle. Although the concept of ride-sharing has existed for decades and can potentially provide several societal benefits, such as reducing travel costs, congestion, and emissions, its adoption remained limited for some time (see, for example, Furuhata et al. 2013). However, the recent ubiquity of digital and mobile technology, and the popularity of peer-to-peer services, have led to unprecedented growth in recent years. In the U.S. market, several players share the ride-sharing market.

Field experiments in online platforms: A recent trend for online platforms is to design and run a multitude of micro-experiments with the goal of generating high-quality data. A typical platform can decide to run a series of carefully designed experiments (often called *A/B tests*) to validate some intuitions on users' behavior. For example, are users more likely to react to promotions sent in the morning or in the evening? To answer such a question, the firm can design a small experiment and randomly send promotions to two samples of users. Then, by testing the statistical significance of the results, the platform can gain some important knowledge. Today, Microsoft and several other leading companies, including Amazon, Booking.com, Facebook, and Google, each conduct more than 10,000 online controlled experiments annually, with many tests engaging millions of users (Kohavi and Thomke 2017). For more details on this topic, we refer the

reader to the recent article by Kohavi and Thomke (2017) and 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., Singh et al. 2017, Fisher et al. 2017, Zhang et al. 2017, Gallino and Moreno 2015).

Marketing: This paper is related to two streams of marketing research. The first steam focuses 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). In particular, this type of studies investigate the relationship between service quality and profits (see, e.g., Zahorik and Rust 1992). Other papers examine how customers react to service failures or dissatisfaction (Berry and Parasuraman 2004, Bolton 1998, Smith and Bolton 1998). For example, Smith and Bolton (1998) examine how customers' dissatisfaction from service failures affects their cumulative assessment. The authors show that appropriate compensation can enhance customer satisfaction and increase retention, suggesting that firms have incentives to proactively compensate for service failures. The second stream is related to targeted (email and mobile) offers (see, e.g., Arora et al. 2008, Fong et al. 2015, Andrews et al. 2015). Typically, users are targeted according to demographics and past purchase behavior with the goal of maximizing conversion rates. In this paper, however, the mechanism to select targeted users depends on the quality level experienced by the rider. More precisely, we formally define frustrating events (see Section 2) that trigger a user to receive compensation.

Structure of the paper. Section 2 presents the design and implementation of the three field experiments we conducted. Section 3 describes our data and discusses three key metrics related to the riders' engagement behavior. Section 4 reports the results of our experiments, including ANOVA tests, regression analyses, and a difference-in-differences approach. Finally, our conclusions are presented in Section 5.

2. Field Experiments: Design and Implementation

In this section, we present the field experiments we designed and conducted together with our industry partner. First, we run an experiment in New York City (NYC). Second, we conduct an additional experiment in Washington D.C. This allows us to test the robustness of our findings in a different market and at a different time period and refine our findings. Finally, we run a third experiment to draw complementary insights and to strengthen our managerial implications.

2.1. General Scope

As mentioned, our goal is to proactively send promotions to riders who experienced low quality of service (henceforth, a frustration). To this end, 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 considered to be among

the most important ones for customer satisfaction. We selected these two metrics after analyzing rider feedback originating in the past year (2017). When a rider places a ride request by specifying the pick-up and drop-off locations, s/he typically receives 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 2 minutes (among other options). Subsequently, the rider can decide whether or not 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/when 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 2 minutes, but the driver arrived for pick-up after 6 minutes, the ETA error is equal to 4 minutes. In such a situation, the rider experiences a frustration as s/he had to wait longer than anticipated.

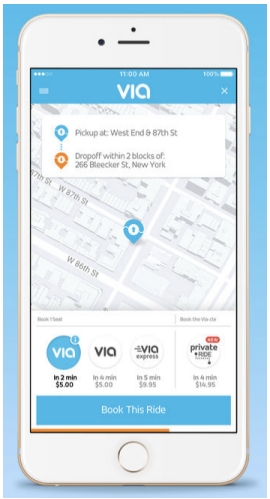

Figure 1 **Screenshot of Via's interface (source: App Store).**

The second frustration metric is related to the 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/or dropped-off by the driver (i.e., the number of stops). The travel time can also be affected, for example, by the traffic and weather conditions. Note that factors such as the number of stops or the itinerary can be controlled by Via, whereas factors such as the weather are not. Our goal is to capture an aggregated measure that is normalized for uncontrolled factors. For the purpose of our field experiments, we propose to consider a metric called *VGR*, which stands for Via Google ratio. For each ride in real-time, 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 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. This VGR metric allows us to normalize for the 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.

Next, we 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 in the sequel). As we will discuss in Section 3.1, Via provides the vast majority of their rides within very good service levels on both dimensions (i.e., short waiting times and low VGRs).

2.2. Experiment 1: New York City

Our objective in the first field experiment is to investigate the impact of different 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 from the perspective of riders. We decided to use the threshold value of 10 minutes as a large number of rider complaints were initiated under such circumstances. Note that such events correspond to the 0.34% worst values in our dataset (see details in Section 3.1). Similarly, we set the threshold value for the VGR to be 2 (i.e., 100% higher relative to the travel time estimated by Google Maps).

Specifically, our first experiment was conducted between July 5th and August 25th 2017, in New York City. During this period, we monitored a subset of riders who experienced an unexpectedly long ETA error (in our case, more than 10 minutes) and riders who experienced large VGRs (more than 2).[5] If a rider experiences either an ETA error greater than 10 minutes or a VGR larger than 2, we classify such an 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 for 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 future rides.[6] Last, the *Waived* condition includes the riders who received a waived ride (i.e., the charges for the ride were refunded to the rider). Each rider was sent the appropriate promotion via a text message

[5] 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).

[6] In our dataset, the average price of a ride in NYC is $7.10, so the $5 almost corresponds to a free ride.

(the sample of text messages sent to riders who experienced a long ETA error can be found in Figure 2).

Overall, our experiment includes a total of 3,982 riders who were 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/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 Section 4.1.

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!

Figure 2 Examples of text messages sent to riders in our field experiment (for the ETA error category).

2.3. Experiment 2: Washington D.C.

We next design and run a similar experiment in Washington D.C. Our first goal is to test the robustness of the findings from Experiment 1 on a different market (e.g., different traffic patterns, different competition intensity, different market maturity, and different riders) during a different time period. In addition, we refine the design of our experiment by exploiting the knowledge gathered in the first experiment. First, we decided to test different levels of monetary compensations. In the first experiment, we use only a $5 credit (which is a typical level for promotions/compensations). In the second experiment, we use $5 as well as a smaller amount (50% discount on the next ride, which amounts approximately to $1.80). This additional condition allows us to examine how different credit levels impact the engagement of riders who have experienced a frustration. Second, we

vary the threshold value for the ETA error. Instead of using 10 minutes, we decreased the threshold to 8 minutes. The goal of this reduction is to infer the appropriate definition for a frustration. Third, we decided to focus on the ETA error only as the VGR frustration was not statistically conclusive from the results of Experiment 1. As we will elaborate in Section 4.1, one of the key findings of Experiment 1 is that the riders in the Credit condition are more likely to spend more (and complete a larger number of rides) relative to the riders in the 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. 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. Specifically, the Washington D.C. experiment was conducted from September 28th to November 7th, 2017. This experiment includes a total of 948 subjects divided as follows: Control (308), Discount (342), and Credit (298). As before, by comparing the behavior of riders in those three conditions, it will allow us to test the robustness of our findings and to refine our managerial insights. The results of this experiment are presented in Section 4.2.

2.4. Experiment 3: Viaversary

Our first two experiments focused on sending compensations to riders who experienced a frustration. It is clear that sending a reward to a rider should increase her/his engagement on the platform. The question is the following: Given a budget of rewards, is it more effective to send a compensation to a frustrated rider or to a random 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, random riders (who are very likely to have experienced higher service level than frustrated riders) may react better to promotional offers. This question is the main motivation behind our third experiment.

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 Control and Credit conditions. As before, the riders in the Credit condition received a $5 credit to celebrate their joining date anniversary (whereas the riders in the Control condition were not sent anything). Specifically, this experiment was conducted in NYC from October 30th 2017, to January 1st, 2018, and includes a total of 605 subjects divided as follows: Control (177) and Credit (428). To ensure that we select a sample of active users, we restrict the selection process to riders who have used

the service within 2 weeks prior to their Viaversary date. This experiment allows us to compare the effectiveness of sending a compensation to frustrated riders versus non-frustrated riders. The results of this experiment are presented in Section 4.3.

3. Data and Key Metrics

In this section, we first provide some background on our industry partner. Then, we describe our data and report the summary statistics of the key variables we consider. Finally, we discuss three key metrics that capture the engagement behavior of the riders exposed to our experiments.

3.1. Industry Partner and Data Description

We start by reporting a brief overview of our industry partner, Via Transportation Inc., or simply Via.[7] Founded in June 2012, Via is a privately held transportation company based in New York City 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. In NYC, Chicago, and Washington D.C., more than 25 million rides have been served through the Via platform. As of January 2018, Via is providing over 1.5 million rides monthly,[8] and it was reported that Via had raised more than $375 million in financing.

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 about $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. They're the difference between Via being a piece of software, which if you have a bad experience you delete and never use again, and a company that you have a personal connection and brand loyalty to through thick and thin. 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 when they arise, provide context, and have considerable discretion to compensate proactively.

To guide the empirical analysis presented in this paper, we use a large historical dataset. More precisely, our dataset includes all the rides completed in New York City between May 1 and

[7] https://ridewithvia.com/

[8] https://www.prnewswire.com/news-releases/navya-partners-with-via-to-introduce-a-revolutionary-new-app-enabling-safe-and-secure-autonomous-rides-with-unprecedented-user-controls-300581722.html

December 31, 2017. Each observation in our dataset 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 the rider ID, the exact times and locations (of both pick-up and drop-off), the distance traveled, the proposed ETA, the ETA error, trip duration metrics, and the pre-tax price paid.

Our dataset includes several million rides completed by a large number of different riders.[9] In this paper, the two main relevant features are the ETA error and the VGR, which translate to two different types of frustrations. 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 above, Via still strives to improve the experience for customers that 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 Experiment 1). Note that ETA errors greater than 10 minutes occur in less than 0.34% of the rides in our dataset. Even though experiencing a 10 minutes 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 (for Experiment 1). The occurrences where the VGR is greater than 2 are also very rare in our dataset.[10]

3.2. Key Metrics

Our goal is to identify and use the appropriate metrics to accurately measure riders' engagement with the platform. Typically, capturing engagement behavior depends highly 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 1-week window or a 1-month period?). To measure riders' engagement behavior, we consider the following three metrics: (i) the total spending (in \$) during the first T weeks after being exposed to the experiment, (ii) the total number of rides completed during the first T weeks after the experiment, and (iii) the average time interval between rides within the first T weeks after the experiment. Note that for robustness purposes, 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.

[9] We cannot reveal the exact details of our dataset due to confidentiality reasons. However, such information has no impact on the analysis and on the key findings presented in this paper.

[10] Due to confidentiality, we cannot reveal the summary statistics of the VGR (note that this does not have any impact on the results or the findings presented in this paper).

The first two metrics (total spending and total number of rides) within the first T weeks after being exposed to the experiment allow us to investigate how rider usage behavior varies depending on the condition—namely, how different promotions can compensate for the frustration experienced by the rider. We next compare these two metrics to their corresponding values prior to the experiment. This provides a way to measure how the post-experiment engagement is different relative to the pre-experiment behavior. Finally, the last metric (average time between rides) allows us to infer the impact of the different conditions on the frequency of usage, which is an important metric for online platforms. We next discuss in greater detail these engagement metrics.

Total spending (in $) and total number of rides:

$$S_T^j = \frac{1}{N_j} \sum_{i \in j} \sum_{t=1}^{T} s_i^j(t), \tag{1}$$

$$R_T^j = \frac{1}{N_j} \sum_{i \in j} \sum_{t=1}^{T} r_i^j(t), \tag{2}$$

where i corresponds to a rider, j to a compensation condition (i.e., Control, Comms, Credit, or Waived), and t to a week. As mentioned before, 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), the quantity $s_i^j(t)$ 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. As a result, S_T^j denotes the average cumulative spending of riders in condition j during T weeks. Similarly, in equation (2), the quantity $r_i^j(t)$ represents the the number of rides completed by rider i from condition j in week t. Therefore, R_T^j denotes the average total number of rides completed by riders in condition j during T weeks.

By comparing S_T^j and R_T^j across the different conditions, we can understand how different conditions (i.e., compensation methods) affect the engagement of a rider who experienced a frustration. To complement our analysis, we will also consider the total spending and the total number of rides prior to the experiment. We will then compare the quantities S_T^j and R_T^j for the different conditions before and after the experiment. These results, however, do not provide any indication whether the frustration affects the engagement behavior. To identify the causal effect of the frustration on the engagement behavior, we will construct a sample of non-frustrated riders (this analysis is presented in Sections 4.1.6 and 4.2.5).

Average time interval between rides:

Last, we consider an additional engagement metric based on the average time intervals between rides (i.e., the frequency of using the service). Similar to the total spending and the total number of rides, we believe that the time interval between rides is a good way to capture the engagement behavior. To measure the impact of the experiment, we compute the average time interval between

rides during the first T weeks after the experiment (i.e., experiencing a frustration). The average time interval between successive rides can be written as

$$\bar{D}_T^j = \frac{1}{N_j} \sum_{i \in j} \Big[\frac{1}{T_i} \sum_{m=1}^{T_i} \Big(d_{i,m}^j - d_{i,m-1}^j \Big) \Big], \tag{3}$$

where i corresponds to a rider, j to a compensation condition (i.e., Control, Comms, Credit, or Waived), and the index m denotes the m-th ride completed by rider i (within T weeks after being exposed to the experiment). T_i represents the total number of rides completed by rider i within T weeks. The term $d_{i,m}^j$ denotes the date of the m-th ride completed by rider i from condition j after being exposed to the experiment ($d_{i,0}^j$ represents the date on which rider i was exposed to the experiment). As a result, $\bar{D}_T^j$ corresponds to the average time interval between rides for the riders in condition j during T weeks. Note that in equation (3), we consider only the riders who rode at least once after being exposed to the experiment (i.e., we exclude the riders who did not ride during the T weeks following the experiment exposure). Interestingly, we observed that the proportion of riders who rode less than once are very similar in each condition. It is worth mentioning that for the Credit condition, the first ride after the experiment may be problematic. Since such a ride includes a \$5 discount (i.e., free or almost free), the riders in the Credit condition have a strong incentive to complete their first ride. To mitigate this effect, we consider both including and excluding the first ride observation in our analysis (see details in Section 4.1.3).

3.3. 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.[11] To address this issue, we eliminate the top 1% of the 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 the riders who spent the most (similarly, we eliminate the top 1% of the observations for each key metric). To ensure the robustness of our results, we vary this threshold from 1% to 5% by increments of 1%. We observed consistent results and patterns under each of these thresholds. Consequently, in the next section, we report the results obtained when we discard the top 1% of the observations based on the distribution of each key metric.

[11] For instance, this set of riders includes some 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.

4. Field Experiments: Results

In this section, we report the results of the three field experiments we conducted. For each experiment, we report the results for each condition and each key metric. Then, we refine these results by applying several segmentations. Finally, we complement our findings by performing different robustness checks, a regression analysis, and a difference-in-differences approach.

4.1. Experiment 1: New York City

As mentioned in Section 2.2, this experiment includes a total of 3,982 riders. After applying our filter (see Section 3.3), we are left with 3,947 riders divided as follows: Control (963), Comms (987), Credit (1,344), and Waived (653). We next report the results for each key metric.

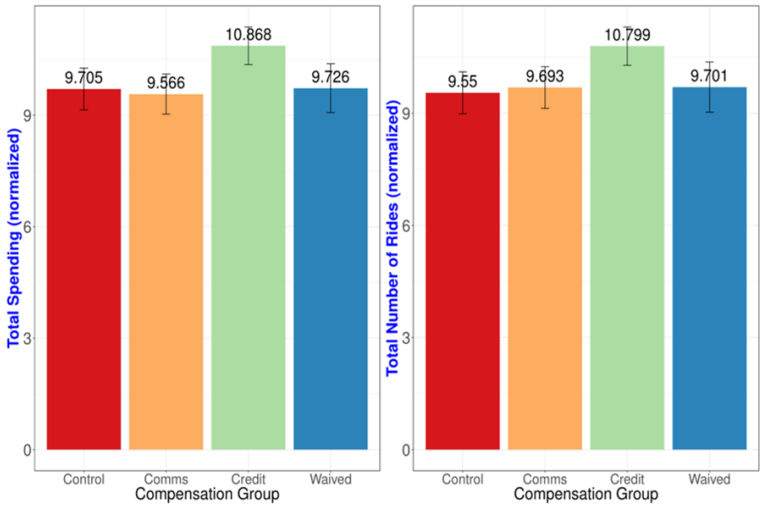

Figure 3 **Average total spending and total number of rides by compensation group for Experiment 1.**

4.1.1. Total spending and total number of rides. The left (resp. right) panel in Figure 3 reports the average total spending S_T^j (resp. total number of rides R_T^j) for each condition using $T = 4$.[12] We will explore smaller time windows in the sequel. The results are presented in Figure 3.[13] We then can infer the following (by taking the average across all the samples):

- Riders in the Credit condition spent 11.99% more (and took 13.07% extra rides) relative to riders in the Control condition.
- Riders in the Credit condition spent 13.61% extra (and took 11.40% extra rides) relative to riders in the Comms condition.

[12] For each figure, we also include the confidence interval corresponding to each average value.

[13] We normalize all the numbers presented in all the figures so as not to reveal sensitive business information. However, the normalization does not affect the relative differences between the different conditions, which is the main focus of this paper.

- Riders in the Credit condition spent 11.75% extra (and took 11.32% extra rides) relative to riders in the Waived condition.

Note that for the total spending, the result of the one-way ANOVA (see, e.g., Maxwell and Delaney 2004) is significant for all four conditions ($F(3, 3939) = 5.43,\ p<.01$). In addition, the post-hoc comparisons among the different conditions using the Bonferroni procedure is significant at the 95% confidence level. Similarly, for the total number of rides, the result of the one-way ANOVA is also significant for all four conditions ($F(3, 3939) = 4.82,\ p<.01$), and the post-hoc comparison among the different conditions is also significant at the 95% confidence level (with the exception of the comparison between the Credit and Control conditions, which is significant at the 90% level).[14]

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 a compensation to the frustrated riders. Our results suggest that providing a $5 credit toward a future ride is significantly more effective than sending a text message to the rider (i.e., Comms) or waiving the charge. Such a finding bears the following important practical implications:

1. The service provider can compensate for poor experiences by proactively sending a compensation.
2. It is more effective to credit a rider's account than waiving the charge. This follows from the fact that providing a credit value offers an opportunity for the rider to use the service again. Since on average the quality of service is high, the rider is very likely to experience a high quality of service (low ETA error and small VGR), and as a result, this will correct for the previous frustrating experience. On the other hand, by sending an apologetic text message or by waving the charge, it can be considered as an instant treatment, without providing an incentive for the rider to try the service again. In addition, providing credit is better than waiving the charge from a pure revenue perspective.
3. The Credit condition is significantly different relative to each of the other conditions. Nevertheless, Control, Comms, and Waived are not statistically different from each other.
4. Offering a $5 credit to frustrated riders is revenue positive. Our results suggest that riders in the Credit condition will spend on average an extra 11.99% relative to riders in the Control condition. In addition, subtracting the $5 investment, this marketing campaign is revenue positive.

[14] Regarding the pairwise comparison of the total spending: Control (M=9.7, SD=8.85) is statistically significant relative to Comms (M=9.57, SD=8.63), Credit (M=10.87, SD=9.49), and Waived (M=9.73, SD=8.53). Regarding the pairwise comparison of the total number of rides: Control (M=9.55, SD=8.9) is statistically significant relative to Comms (M=9.7, SD=9.03), Credit (M=10.8, SD=9.77), and Waived (M=9.7, SD=8.9).

We next analyze how the main effects we just described are affected by the different types of frustration by computing the two metrics for each segment separately (ETA error and VGR). The upper (resp. lower) panel in Figure 4 reports the average total spending (resp. total number of rides) for the ETA error and VGR segments. Interestingly, 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.[15] Note that for the VGR segment, the numerical value of the Credit condition is still higher relative to the 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. One possible explanation may be that riders tend to blame the service provider for high ETA errors but not for large VGR values. In other words, it seems that waiting more than anticipated seems to be 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 partially explains why we obtained a stronger signal for the ETA error segment.

Next, we analyze how the main effects are moderated by different levels of pre-experiment usage. Our motivation behind this analysis is twofold. First, as discussed in Section 3.1, 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 (i.e., selection bias). To address this issue, we examine how the main effects are moderated by the pre-experiment usage.[16] Second, we expect that the frequency of usage impacts the way that riders perceive the frustration. In particular, a long-term relationship between the service provider and riders would soften (or strengthen?) the impact of a frustrating experience on the engagement.

Specifically, we divide the riders from our experiment into three groups: high, medium, and low, based on their usage prior to the experiment. To remain consistent with our key metrics (S_T^j and R_T^j), we compute the total spending and the total number of rides for each rider during the four weeks prior to the experiment. Based on these variables, we define the top 30% of the riders as the high group, the bottom 30% as the low group, and the remaining riders are assigned to the medium group. For robustness purposes, we vary this threshold from the top 10% to the top 50%, and the results are consistent across the different threshold values. In addition, we consider different combinations between the total spending and the total number of rides to segment the riders according to their pre-experiment usage (under all the variations we considered, we obtained

[15] Regarding the total spending for the ETA error segment: $F(3,1760) = 4.41,\ p<0.01$, and for the VGR segment: $F(3,2175) = 2.04,\ p = .11$. Regarding the total number of rides for the ETA error segment: $F(3,1763) = 3.82,\ p<0.01$, and for the VGR segment: $F(3,2176) = 1.32,\ p = 0.27$.

[16] Note that we deal with this issue more thoroughly in the subsequent sections by comparing the pre- and post-experiment engagement levels and by controlling this factor in the regression analysis.

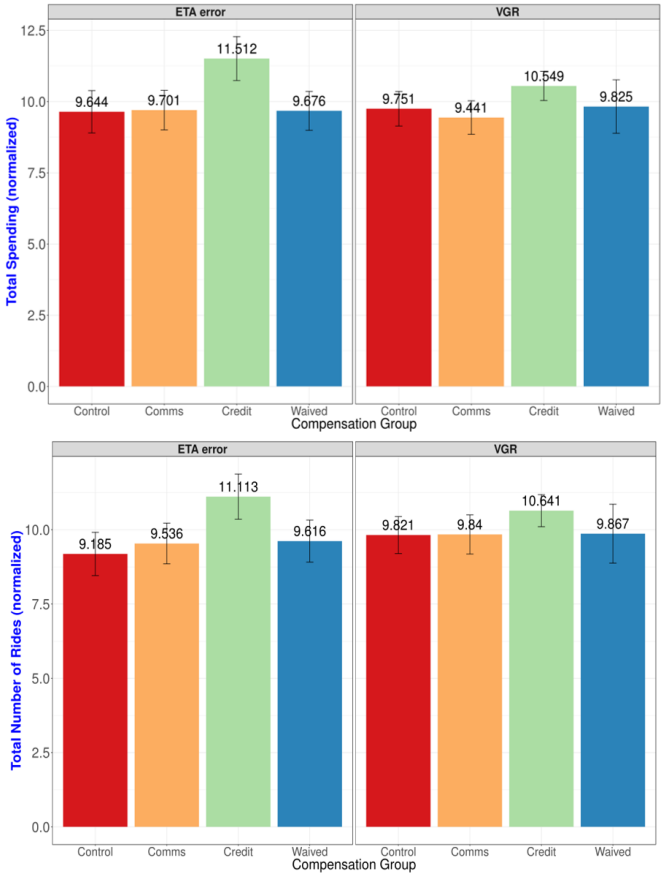

Figure 4 **Average total spending and total number of rides by compensation group and by type of frustration for Experiment 1.**

consistent results). Finally, we observed consistent results when using time windows of 4, 5, and 6 weeks to divide the riders into the three groups. The results are presented in Figure 5. We obtained that the difference in the average total spending between the Credit and Control conditions is statistically significant only for the high and medium groups.[17] In addition, for the total number of rides, this finding holds only for the medium group. This result implies that the level of pre-experiment usage does affect the impact of compensating frustrated riders. Specifically, we obtain a significant effect for the high and medium groups but not for the low group. This suggests the following managerial insights:

1. New riders who have experienced a frustration are not affected by receiving a promotion aimed to compensate for their frustration. New riders 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 typically high).

[17] Regarding the total spending for the high group: $F(3,1119) = 6.16,\ p<0.01$, for the medium group: $F(3,1490) = 3.98,\ p<0.01$, and for the low group: $F(3,1117) = .20,\ p>0.1$. Regarding the total number of rides for the high group: $F(3,1055) = 1.77,\ p = 0.15$, for the medium group: $F(3,1400) = 4.6,\ p<0.01$, and for the low group: $F(3,1156) = .19,\ p = 0.19$.

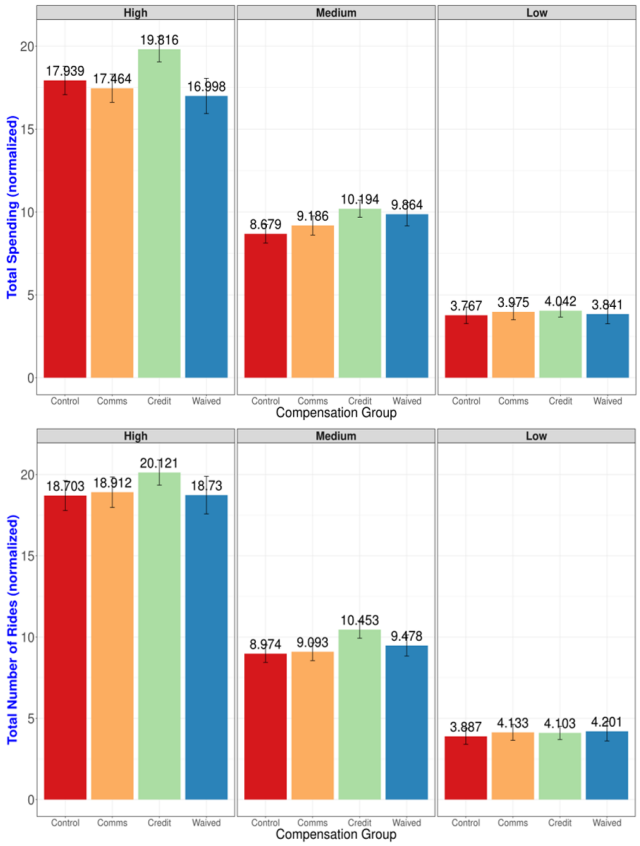

Figure 5 **Average total spending and total number of rides by compensation group and by pre-experiment usage for Experiment 1.**

2. As before, riders in the Credit condition spend (and ride) more relative to riders in the other conditions. Regarding the total spending, the difference between the Credit condition for the high group is statistically significant at the 95% level with respect to each of the other conditions (for the middle group, only the difference between the Credit and Control conditions is significant at the 95% level).
3. For both the high and medium groups, it is profitable to offer a $5 credit following the frustration, that is, the additional spending between the riders in the Credit and Control conditions is larger than $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 10.46% additional spending between the Credit and Control conditions. For riders who are in the medium group, we obtain an average of 17.46% additional spending between the Credit and Control conditions. Therefore, the relative effect is the most significant for the riders in the medium group. This suggests that the most effective strategy is to compensate the riders in the medium group. Indeed, the riders in the low group are still exploring/discovering the service, and the riders in the high group are already loyal and use the service very often. As a result, the service

provider should prioritize making sure that the riders in the medium group experience a high quality of service or compensating them for the rare instances where they do not.

4.1.2. Varying the time window and considering the pre-experiment usage. So far, we have shown that riders in the Credit condition are more likely to be engaged with the service relative to riders in the other conditions. We next investigate the pre-experiment behavior of riders in each condition. More precisely, we compare riders' usage between the different conditions both during the pre-experiment and post-experiment periods. The upper part of Figure 6 presents the

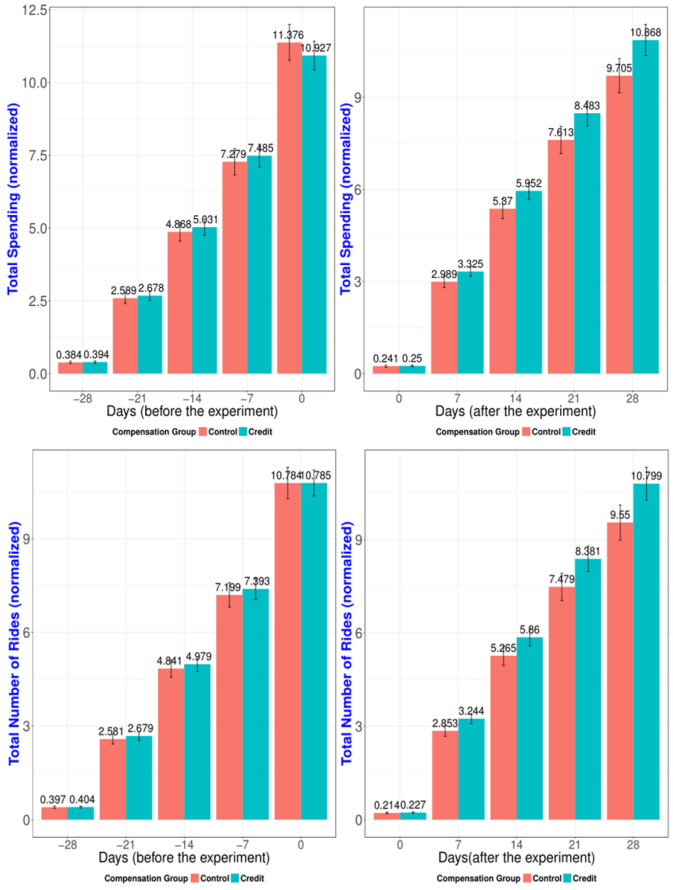

Figure 6 Cumulative spending and number of rides over time by compensation group for Experiment 1 (week 0 denotes the experiment exposure).

cumulative total spending over time (i.e., S_T^j for different values of T). The upper right panel presents the cumulative total spending during the 4 weeks following the exposure to the experiment. The x-axis in the right panel indicates time points starting from the date when the riders were exposed to the experiment ($x = 0$) until 4 weeks after the exposure time ($x = 28$). The upper left panel reports the cumulative total spending during the pre-experiment period (here, the x-axis indicates the time points from the 4 weeks prior to the experiment until the date on which the

riders were exposed to the experiment). The lower right (resp. left) panel reports the cumulative total number of rides during the post-experiment (resp. pre-experiment) period. Interestingly, one can see from Figure 6 that the cumulative total spending during the pre-experiment period is not statistically different for the riders in the Credit and Control conditions. For example, the riders in the Credit condition spent on average 2.99% more relative to the Control condition during the 4 weeks prior to the experiment. Moreover, these two numbers are not statistically different (i.e., $F(3,3939)=0.24,\ p=0.87$). This pattern is replicated for each time point during the pre-experiment period.[18]

On the other hand, the riders in those two conditions show different engagement behaviors after being exposed to the experiment. Starting from the first week after the experiment, riders in the Credit condition are more likely to spend more (and to complete a larger number of rides) relative to riders in the Control condition. In addition, this gap increases over time so that 4 weeks after being exposed to the experiment, riders in the Credit condition spent on average 11.99% more relative to riders in the Control condition. This difference is statistically significant ($F(3,3939)=5.43,\ p<.01$, and the pairwise comparisons are reported in Figure 3). The same pattern is observed across all four weeks after the experiment.[19] Note that it takes about 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. Furthermore, a return on investment in 18 days in this context is considered a high level of performance. Similarly, as we can see from the lower panels in Figure 3, the total number of rides over time (R_T^j) shows the same consistent pattern.[20]

The previous analysis strengthens our findings. This confirms that the service provider should proactively send an appropriate compensation to riders who experienced a frustration. Otherwise, such riders may potentially decrease their engagement level (relative to the Control condition) due to the unpleasant experience. This finding suggests that by proactively sending a monetary compensation, the firm can mitigate the adverse effect of frustrated riders who decrease their engagement level.

4.1.3. Average time interval between rides. Next, we analyze the time interval between rides (i.e., $\bar{D}_T^j$) using $T=4$. One can see from Figure 7 that riders in the Control condition take on

[18] For 3 weeks: $F(3,3939)=0.37,\ p=0.77$, for 2 weeks: $F(3,3939)=0.75,\ p=0.52$, and for 1 week: $F(3,3939)=1.17,\ p=0.32$.

[19] For 3 weeks: $F(3,3939)=5.49,\ p<.01$, for 2 weeks: $F(3,3939)=5.04,\ p<.01$, and for 1 week: $F(3,3939)=5.14,\ p<.01$.

[20] For the post-experiment data: For 4 weeks: $F(3,3939)=4.82,\ p<.01$, for 3 weeks: $F(3,3939)=4.39,\ p<.01$, for 2 weeks: $F(3,3939)=4.04,\ p<.01$, and for 1 week: $F(3,3939)=4.77,\ p<.01$. For the pre-experiment data: For 4 weeks: $F(3,3939)=0.60,\ p=0.62$, for 3 weeks: $F(3,3939)=0.58,\ p=0.63$, for 2 weeks: $F(3,3939)=0.34,\ p=0.80$, and for 1 week: $F(3,3939)=0.05,\ p=0.99$

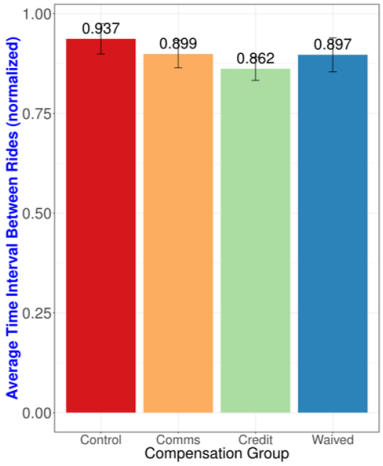

Figure 7 Average time interval between rides for Experiment 1.

average 8.65% more time between rides relative to riders in the Credit condition. The difference between the Credit and Control conditions is statistically significant at the 90% confidence level (one-way ANOVA test: $F(3, 3203) = 2.37$, $p = .07$, and the post-hoc pairwise comparison with t-test using the Bonferroni correction: For Control (M=0.94, SD=0.64) and for Credit (M=0.86, SD=0.586) are statistically different at the 90 % confidence-level). Note that on average, the riders in the Comms and Waived conditions take a longer time relative to the Credit condition but the differences are statistically insignificant.[21] Recall that the first ride after exposure may be problematic for riders in the Credit condition (as they can use the $5 credit). Hence, we next perform two robustness checks. First, we conduct the same analysis while excluding the first ride for all the conditions. Second, we exclude the first ride only for the Credit condition (and include it for all the other conditions). In both cases, we observe the same pattern, that is, riders in the Credit condition have a shorter interval between rides relative to the other conditions, suggesting that our results are robust. In summary, the results presented in Figure 7 present patterns consistent with our previous findings. In particular, riders in the Credit condition are more likely to be engaged with the service relative to riders in the other conditions.

4.1.4. Regression analysis. In the previous sections, we presented the results from several one-way ANOVA tests. Our analysis suggests that riders in the Credit condition are more likely to be engaged relative to riders in the other conditions (Control, Comms, and Waived). In addition, we investigated whether this main finding continues to hold after controlling for some of the potential factors that may affect the engagement behavior. To complement our study, we next run a regression analysis to examine how different types of compensations affect riders' engagement.

[21] Post-hoc pairwise comparison: Comms (M=0.9, SD=0.59) and Waived (M=0.89, SD=0.59) are not statistically different from the Credit condition.

To keep the presentation short, we focus on the first two key engagement metrics, as our dependent variables (S_T^j and R_T^j) within the first 4 weeks after being exposed to the experiment (i.e., $T=4$). Concretely, we consider the following regression equation:

$$\begin{aligned} y_i = \alpha + \beta_1 \text{Comms}_i + \beta_2 \text{Credit}_i + \beta_3 \text{Waived}_i + \\ + \gamma_1 \text{Pre-experiment-Rides}_i + \gamma_2 \text{Pre-experiment-Rides}_i^2 + \gamma_3 \text{ETA-error}_i + \mu_i + \epsilon_i, \end{aligned} \tag{4}$$

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 R_T^j); and $\text{Comms}_i, \text{Credit}_i$, and Waived_i represent binary variables for each experiment condition (the Control condition is the reference group). Note that we control for the riders' pre-experiment usage by including the total number of rides during the pre-experiment period (of 4 weeks) as well as a quadratic term to capture the potential non-linear effect. This allows us to study how the main effect is moderated by the pre-experiment engagement behavior. We also include a dummy variable for the ETA error segment, which indicates whether rider i belongs to the ETA or VGR segment. Finally, we include date/time fixed effects, which vary at the rider level, by capturing the date on which rider i was exposed to the experiment (variable μ_i). Such a variable helps controlling for the individual heterogeneity, which is similar to the individual fixed effects.[22] The last term, ϵ_i, is an IID error term which is assumed to follow a normal distribution.

The results of the OLS regression are reported in Table 1. Consistent with our findings from the one-way ANOVA tests, the coefficient of the Credit indicator is positive and statistically significant, which means that riders in the Credit condition complete a larger number of rides relative to riders in the Control condition. However, the variables for Comms and Waived are not significant, meaning that riders in the Comms and Waived conditions are not statistically different relative to riders in the Control condition. For robustness purposes, we also consider the regression equation with a log transformation in the dependent and independent variables that have non-binary values (this allows us to correct for the skewness of the distribution) (see the last two columns of Table 1). Finally, since the dependent variable, y_i, can take only positive integer values, we also run a Poisson regression with the same specification as in (4) (based on the assumption that the error term ϵ_i follows a Poisson distribution). We observed that the results of the Poisson regression are consistent with the OLS regression. An additional interesting finding is related to the non-linear effect of the pre-experiment usage. One can see that the quadratic coefficient is negative so that it implies a concave effect (i.e., diminishing marginal differences). This confirms the finding that riders in the medium group are potentially the ones with the highest relative impact (see the discussion in Section 4.1.1).

[22] For robustness purposes, we also considered different ways of capturing the individual fixed effects, such as the days of the week on which rider i was exposed to the experiment. Note that we obtained consistent qualitative results when we removed the fixed effects from the regression equation.

Table 1 Regression results for Experiment 1.

OLS with pooled samples	(1)	(2)	(3)
	Model1	Model2	Model3
Variable	DV: Total Rides	DV: log(Total Rides)	DV: log(Total Rides)
Comms	0.343	0.0131	0.00583
	(0.381)	(0.0343)	(0.0339)
Credit	**1.204*** **	**0.0855*** **	**0.0896*** **
	(0.372)	**(0.0317)**	**(0.0315)**
Waived	0.406	0.0435	0.0372
	(0.436)	(0.0392)	(0.0387)
Pre-experiment-Rides	**0.981*** **	**0.107*** **	
	(0.0419)	**(0.00315)**	
Pre-experiment-Rides ^{2}	**-0.00598*** **	**-0.00109*** **	
	(0.00101)	**(7.02e-05)**	
ETA error	0.224	0.00810	0.00964
	(0.286)	(0.0246)	(0.0242)
Log(Pre-experiment-Rides)			**0.863*** **
			(0.0140)
Date FE	Yes	Yes	Yes
Constant	3.321**	1.233***	0.294***
	(1.376)	(0.0967)	(0.102)
Observations	3,947	3,947	3,947
R-squared	0.539	0.534	0.549

Next, we run a separate regression analysis for each segment (ETA error versus VGR) to examine the difference in the engagement behavior induced by the two different types of frustration. As shown in Table 2, the results are consistent with the one-way ANOVA test. In particular, the ETA error segment has a significant coefficient for the Credit condition only, whereas the VGR segment shows that none of the coefficients related to the compensation conditions are significant.

Table 2 Regression results for Experiment 1 by frustration type.

OLS by frustration type	(1)	(2)
	ETA only	VGR only
VARIABLES	DV: Total Rides	DV: Total Ride
Comms	0.776	-0.0616
	(0.596)	(0.527)
Credit	**1.957*** **	0.634
	(0.607)	(0.475)
Waived	0.493	0.959
	(0.613)	(0.714)
Pre-experiment-Rides	**1.033*** **	**0.912*** **
	(0.0374)	**(0.0354)**
Pre-experiment-Rides ^{2}	**-0.00754*** **	**-0.00407*** **
	(0.000688)	**(0.000692)**
Date FE	Yes	Yes
Constant	2.647*	4.634***
	(1.506)	(1.682)
Observations	1,767	2,180
R-squared	0.530	0.563

Last, we 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 three groups: high, medium, and low, based on the top 30%, top 30%-70%, and the bottom 30% of the distribution

of the pre-experiment usage (i.e., the total number of rides).[23] Interestingly, as one can see from Table 3, the coefficient for the Credit condition is significant only for the high and medium 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 the pre-experiment usage and that (ii) this effect is present only for the experienced riders who regularly use the service (i.e., high and medium groups). Note that there is no statistical difference between the high and middle groups in terms of the magnitude of the Credit condition coefficient (i.e., the Credit coefficient in the high group is not statistically different from the Credit coefficient in the middle group).

Table 3 Regression results for Experiment 1 by pre-experiment usage.

OLS by Pre-experiment usage	**(1)**	**(2)**	**(3)**
	Top 30%	**Top 30~70%**	**Bottom 30%**
VARIABLES	**DV: Total Rides**	**DV: Total Rides**	**DV: Total Rides**
Comms	0.149	0.370	0.449
	(0.969)	(0.618)	(0.556)
Credit	**1.558***	**1.802*****	0.362
	(0.892)	**(0.584)**	(0.521)
Waived	-0.287	0.418	0.239
	(1.098)	(0.709)	(0.624)
Pre-experiment-Rides	**0.980*****	**1.122****	**1.913*****
	(0.240)	**(0.490)**	**(0.687)**
Pre-experiment-Rides^2	-0.00490	-0.0128	-0.127
	(0.00329)	(0.0177)	(0.0771)
ETA error	0.265	0.295	0.337
	(0.698)	(0.452)	(0.401)
Date FE	Yes	Yes	Yes
Constant	-0.00207	-0.498	-0.317
	(4.183)	(3.233)	(1.424)
Observations	1,059	1,404	1,160
R-squared	0.199	0.118	0.044

We next summarize the results of the regression analysis. Even after controlling for the potential factors that may influence the engagement behavior, we still observe that riders in the Credit condition are more likely to complete a larger number of rides (and to spend more) relative to riders in the Control condition. This effect is moderated by the type of frustration (ETA error versus VGR). In particular, our analysis suggests that only the ETA error segment shows a statistically significant effect. This effect is also moderated by the pre-experiment usage behavior. We find that only riders in the high and middle groups (i.e., all riders except the bottom 30% in terms of usage levels) show a statistically significant effect. In general, the pre-experiment usage behavior depicts a concave pattern such that riders in the middle group potentially show the highest relative impact.

[23] We also checked the case where we use the total spending instead of the total number of rides and observed the same qualitative results.

4.1.5. Difference-in-differences. In the previous section, we conveyed that the findings from the regression analysis are consistent with the results of the one-way ANOVA tests. By comparing the pre-experiment and post-experiment engagement behavior, we next use a *difference-in-differences* approach (see, e.g., Angrist and Pischke 2008) to identify the causality of the compensation effects from the different conditions. Consider the following model specification:

$$\begin{aligned} y_{it} =& \alpha + \beta_1 \text{Comms}_i + \beta_2 \text{Credit}_i + \beta_3 \text{Waived}_i + \gamma \text{After-experiment}_t + \\ &+ \delta_1 \text{Comms}_i \cdot \text{After-experiment}_t + \delta_2 \text{Credit}_i \cdot \text{After-experiment}_t + \\ &+ \delta_3 \text{Waived}_i \cdot \text{After-experiment}_t + \epsilon_{it}, \end{aligned} \tag{5}$$

where i corresponds to a rider and t to a time period (for this analysis, we aggregate the observations at the day level). y_{it} denotes the dependent variable for rider i at time t (we consider both the total spending and the total number of rides). $\text{Comms}_i, \text{Credit}_i$, and Waived_i are binary variables to indicate the condition assigned to rider i (as before, the Control condition is the reference group). $\text{After-experiment}_t$ is a binary variable for the time period after being exposed to the experiment. The key parameters in equation (5) are δ_1, δ_2, and δ_3. These parameters capture the potential causal effect of each type of compensation (following the frustration) on the engagement behavior.

Table 4 Difference-in-differences results for Experiment 1.

	(1)	(2)
	Model1	Model2
Diff-in-Diffs	DV: Total Rides	DV: Total Spending
Comms	-0.00991	**-0.106*** **
	(0.00700)	**(0.0405)**
Credit	**0.0165** **	**0.0832** **
	(0.00652)	**(0.0378)**
Waived	5.76e-06	-0.0178
	(0.00783)	(0.0453)
After-experiment	**-0.0410*** **	**-0.207*** **
	(0.00703)	**(0.0407)**
Comms*After-experiment	0.0145	0.0652
	(0.00990)	(0.0573)
Credit*After-experiment	**0.0416*** **	**0.233*** **
	(0.00922)	**(0.0534)**
Waived*After-experiment	0.00701	0.0748
	(0.0111)	(0.0641)
constant	**0.486*** **	**2.658*** **
	(0.00497)	**(0.0288)**
Observations	228,694	228,694
R-squared	0.001	0.001

As shown in Table 4, the coefficient of the interaction term between the Credit condition and the After-experiment 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 post-experiment period. Consequently, 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, there is no difference between the Control, Comms, and Waived conditions

in terms of the engagement during the post-experiment period. We run the same model with both dependent variables (total spending and total number of rides), and the results were consistent. Therefore, this confirms that the Credit compensation does affect (positively) the engagement of riders who have experienced a frustration.

4.1.6. Non-frustration. So far, we have focused on analyzing the effectiveness of several types of compensation sent to riders who experienced a frustration. Our implicit assumption was that when a rider experiences a frustration (long ETA error or high VGR), it negatively affects his/her frequency of using the service. In this section, our goal is to validate this assumption (note that our results thus far can be used to explain a causal effect of the frustration on the engagement). This will then justify the fact that the service provider should consider sending an appropriate compensation to the frustrated riders. To address this question, we need to show how the engagement behavior of frustrated riders is different from the engagement of non-frustrated riders. Consequently, we need to select a sample of non-frustrated riders and compare this sample to the frustrated riders from the Control condition (i.e., frustrated riders who did not receive any compensation).

We next describe our procedure to select a sample of non-frustrated riders. We naturally want to select riders who are similar to the riders from the Control condition of our field experiment. We look at all the data from July 5, 2017 (the starting date of our field experiment), and record the maximum value of the ETA error during T weeks for each rider (we vary the value of T between 1 and 6 and also consider the average instead of the maximum to check the robustness of the procedure). We label the date on which the maximum ETA error occurred as the "exposure" date. We then remove all riders that have a value longer than 10 minutes (as such riders clearly belong to the frustrated category). For all the remaining riders, we compute the total spending and the total number of rides during the 4 weeks preceding the exposure date. Next, we select riders with a total spending (or total number of rides) during the pre-experiment period similar to that of the riders from the Control condition. We now have two distinct groups of riders. Both groups are similar in terms of engagement behavior, but one of the groups includes frustrated riders while the other group has only non-frustrated riders. At this point, we compute the average total spending and the total number of rides during the 4 weeks after the exposure date. For robustness purposes, we define the non-frustrated group based on either the total spending or the total number of rides during the pre-experiment period (both approaches show a consistent pattern). Note that the initial set of non-frustrated riders is much larger. Therefore, we use a random sampling method to select 1,000 riders at random (we obtained consistent results under different random sets). Specifically, we have a total of 963 and 1,000 riders in the frustrated and non-frustrated groups, respectively.

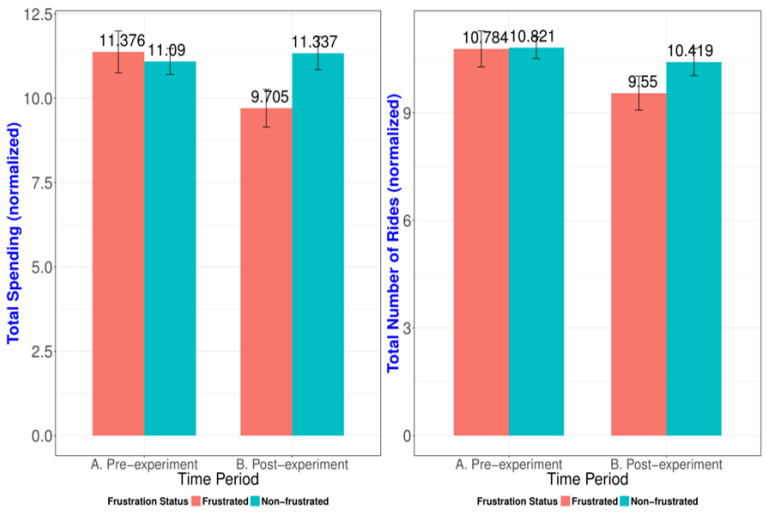

Figure 8 Average total spending and total number of rides by frustration status for Experiment 1.

The left (resp. right) panel in Figure 8 reports the average total spending (resp. total number of rides) for both frustrated and non-frustrated riders during the pre-experiment and the post-experiment periods. As expected, in the pre-experiment period, the frustrated group is not statistically different from the non-frustrated group. The frustrated riders spend on average 4.98% more relative to the non-frustrated riders ($t(1956) = 0.77,\ p = 0.44$, i.e., not statistically significant). However, after they experience a frustration, the frustrated riders spend 12.85% less relative to the non-frustrated group ($t(1956) = -4.31,\ p<.01$, i.e., significant at the 99% level). The same pattern holds for the average total number of rides.[24] This allows us to conclude that experiencing a frustration negatively affects the engagement behavior.

4.2. Experiment 2: Washington D.C.

As explained before, we decided to run a second field experiment to check the robustness of our results in a different market. As mentioned in Section 2.3, this experiment includes a total of 948 riders. After applying our filter (see Section 3.3), we are left with 923 riders divided as follows: Control (305), Discount (332), and Credit (286). We next report the results for each key metric. As mentioned, we discarded the Comms and Waived conditions as well as the VGR frustration in this experiment. This was motivated by the robust and clear 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 on frustrated riders.

[24] Regarding the pre-experiment period, the frustrated riders complete on average 1.98% fewer rides relative to the non-frustrated riders; $t(1960) = -0.83,\ p = 0.41$ (not statistically significant). Regarding the post-experiment period, the frustrated riders complete on average 7.80% fewer rides relative to the non-frustrated riders; $t(1960) = -2.38,\ p<.05$ (significant at the 95% level).

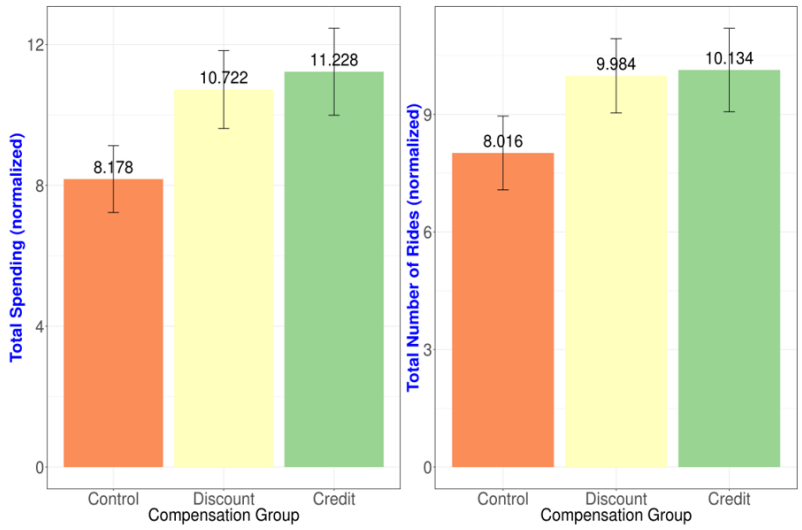

Figure 9 **Average total spending and total number of rides by compensation group for Experiment 2.**

4.2.1. Total spending and total number of rides. The left (resp. right) panel in Figure 9 reports the average total spending S_T^j (resp. total number of rides R_T^j) for each condition within the first 4 weeks after being exposed to the experiment (i.e., $T = 4$). We will explore smaller time windows in the sequel. The results presented in Figure 9 imply the following (by taking the average across all the samples):

- Riders in the Credit condition spent 37.30% more (and took 26.42% extra rides) relative to riders in the Control condition.
- Riders in the Discount condition spent 31.11% more (and took 24.55% extra rides) relative to riders in the Control condition.
- The Credit condition is not statistically different from the Discount condition.

Note that for the total spending, the result of the one-way ANOVA is significant for all three conditions ($F(2, 920) = 5.86,\ p<.01$). In addition, the post-hoc comparisons among the different conditions using the Bonferroni procedure is significant at the 90% level.[25] Similarly, for the total number of rides, the result of the one-way ANOVA is also significant for all three conditions ($F(2, 920) = 3.22,\ p<.05$), and the post-hoc comparison among different conditions using the Bonferroni procedure is also significant at the 90% confidence level.[26] The results presented in Figure 9 translate to the following insights:

[25] The pairwise comparisons between Credit (M=11.228, SD=12.658) and Control (M=8.178, SD=10.041) and between Discount (M=10.722, SD=12.217) and Control are statistically significant but not between Credit and Discount.

[26] The pairwise comparisons between Credit (M=10.134, SD=12.006) and Control (M=8.016, SD=10.973) and between Discount (M=9.984,SD=11.493) and Control are statistically significant but not between Credit and Discount.

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 with the service relative to riders in the Control condition. Note that both experiments (NYC and Washington D.C.) 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 configuration of the 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 8 minutes), we could still observe the same main effect.
3. The difference between Credit and Discount conditions is not statistically significant. This result is interesting as the company seeks to determine the optimal level of compensation for frustrated riders. Our results suggest that a 50% discount remains nearly as effective as a $5 credit.
4. Offering a $5 credit (resp. 50% discount for the next ride) to frustrated riders is revenue positive. Our results suggest that riders in the Credit (resp. Discount) condition will spend on average an extra 37.30% (resp. 31.11%) relative to riders in the Control condition. In addition, by subtracting the $5 (resp. $1.80) investment, this marketing campaign is revenue positive.[27]

Next, we analyze how this main effect is moderated by different levels of pre-experiment usage. Similar to the analysis conducted in Experiment 1, we divide the riders into two groups: high and low, based on their usage prior to the experiment. Since we have a smaller number of riders in Experiment 2, we only use two groups instead of three. To remain consistent with our key metrics (S_T^j and R_T^j), we compute the total spending and the total number of rides for each rider during the four weeks prior to the experiment.[28] Based on these variables, we define the high and low groups by using a threshold from the top 50% to the top 75% (the results are consistent across different threshold values). The results are presented in Figure 10.

We obtained that the difference in the average total spending (and the total number of rides) between Credit and Control conditions is statistically significant only for the high group and not for the low group.[29] As before, we infer that the level of pre-experiment usage does affect the impact of compensating frustrated riders. Specifically, we have:

[27] The average cost of a ride in Washington D.C. in our data is $3.60.

[28] We also perform robustness checks by using further periods in the past.

[29] Regarding the total spending for the high group: $F(2,626) = 9.38,\ p<.01$, and for the low group: $F(2,291) = 1.41,\ p = 0.24$ (High group pairwise comparisons: Credit (M=13.339, SD=13.229) and Control (M=9.464, SD=10.520) are statistically different, and Discount (M=14.317, SD= 13.031) and Control are statistically different at the 90% confidence level). Regarding the total number of rides for the high group: $F(2,625) = 6.08,\ p<.01$, and for the low group: $F(2,292) = 0.45,\ p>.1$ (High group pairwise comparisons: Credit (M=11.723, SD=14.120) and Control (M=9.289, SD=12.697) are statistically different, and Discount (M=11.989, SD=13.906) and Control are statistically different at the 95% confidence level).

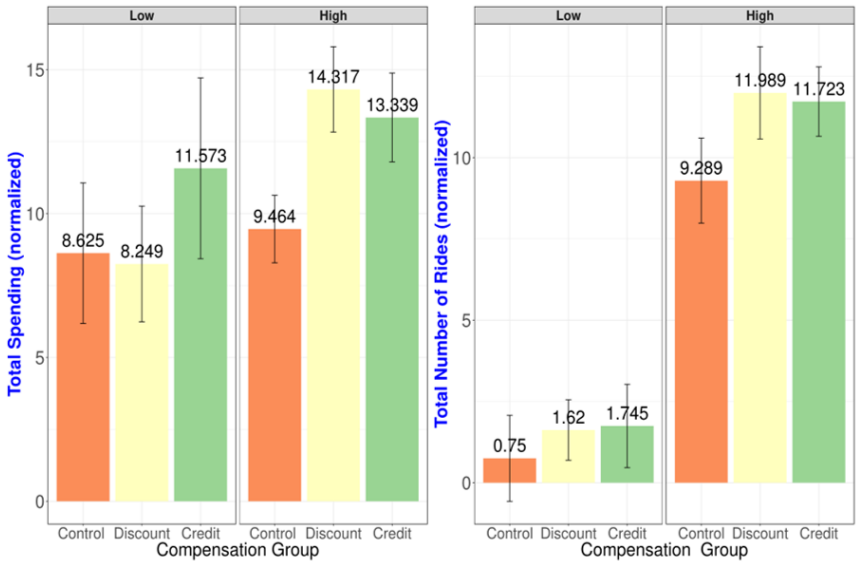

Figure 10 **Average total spending and total number of rides by compensation group and by pre-experiment usage for Experiment 2.**

1. New riders 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 3 rides during the past 4 weeks), we obtain an average of 40.94% (resp. 50.92%) additional spending between the Credit (resp. Discount) and the Control conditions. This suggests that the most effective strategy is to compensate riders who ride more frequently (and potentially avoiding new riders).

4.2.2. Varying the time window and considering the pre-experiment usage. As in Experiment 1, we next compare rider engagement during both the pre-experiment and post-experiment periods. The upper part of Figure 11 presents the cumulative total spending over time (i.e., S_T^j for different values of T). The upper right panel presents the cumulative number of rides during the 4 weeks following the exposure to the experiment. The x-axis in the right panel indicates time points starting from the date on which the riders were exposed to the experiment ($x = 0$) until 4 weeks after the exposure time ($x = 28$). The upper left panel reports the cumulative total spending during the pre-experiment period (here, the x-axis indicates time points from the 4 weeks prior to the experiment until the date on which the riders were exposed to the experiment). The lower right (resp. left) panel reports the cumulative total number of rides during the post-experiment (resp. pre-experiment) period. Interestingly, one can see from Figure 11 that the cumulative total spending during the pre-experiment period is not statistically different for the riders in the Credit (or Discount) and Control conditions. For example, the riders in the Credit

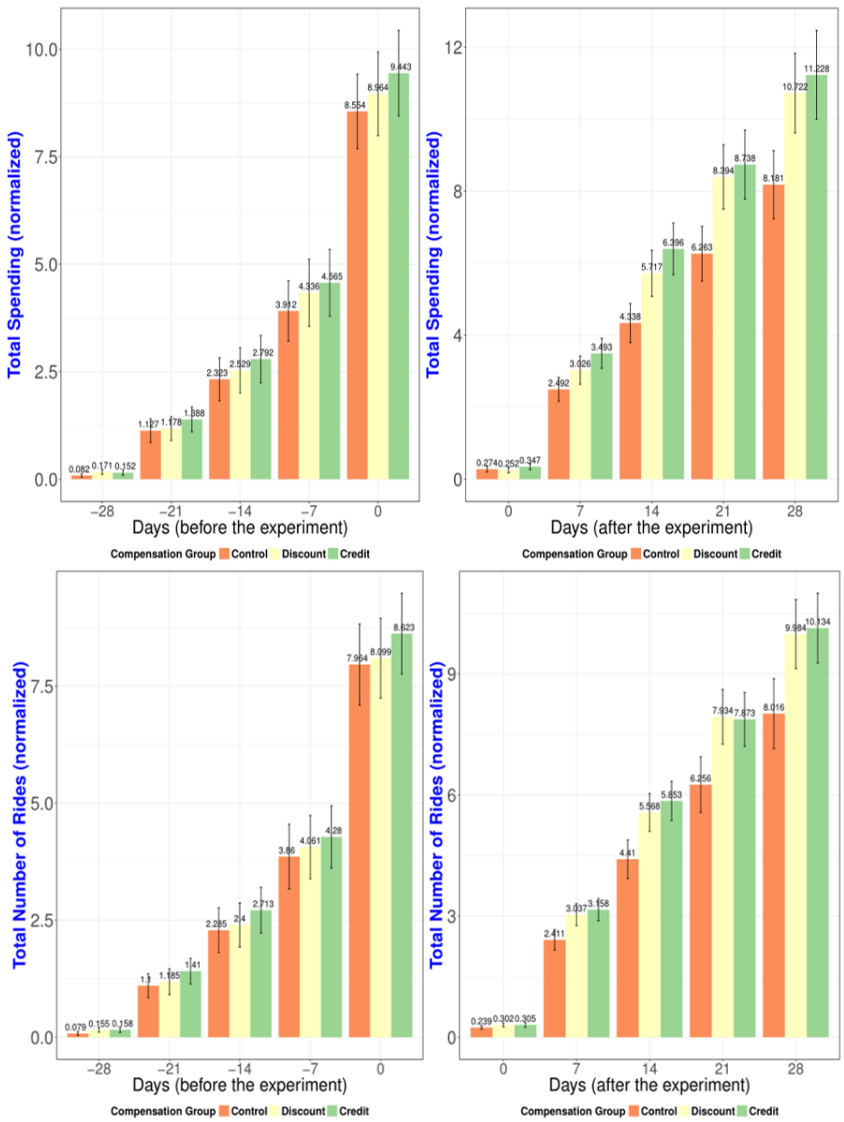

Figure 11 **Cumulative spending and number of rides over time by compensation group for Experiment 2 (week 0 denotes the experiment exposure).**

(resp. Discount) condition spent on average 10.38% (resp. 4.80%) more relative to the riders in the Control condition during the 4 weeks prior to the experiment Moreover, these two numbers are not statistically different (i.e., $F(2,920) = 0.57,\ p = 0.56$). This pattern is replicated for each time point during the pre-experiment period.[30]

On the other hand, riders in those three conditions show different engagement behaviors after being exposed to the experiment. Starting from the first week after the experiment, riders in the Credit and Discount conditions are more likely to spend more (and to complete a larger number of rides) relative to riders in the Control condition. In addition, this gap increases over time so that 4

[30] For 3 weeks: $F(2,920) = 0.50,\ p = 0.60$, for 2 weeks: $F(2,920) = 0.52,\ p = 0.60$, and for 1 weeks: $F(2,920) = 0.50,\ p = 0.61$.

weeks after being exposed to the experiment, riders in the Credit (resp. Discount) condition spent on average 37.30% (resp. 31.11%) extra relative to riders in the Control condition ($F(2,920) = 5.86$, $p<.01$, and the pairwise comparisons are significant between Credit (M=11.228, SD=12.658) and Control (M=8.178, SD=10.041) and Discount (M=10.722, SD=12.217) and Control but not between Credit and Discount). The same pattern is observed across each one of the 4 weeks after the experiment.[31] Note that it takes about 27 (resp. 17) days to earn the \$5 back (resp. the \$1.80). Similarly, as we can see from the lower panels in Figure 11, the cumulative total number of rides over time (R_T^j) shows the same consistent pattern.[32]

The previous analysis allows us to refine our findings. First, it confirms that a similar result as in Experiment 1 is replicated in a different market during a different time period. Second, it seems that the service provider does not need to offer a \$5 credit to compensate for the frustration, as a very similar effect is achieved with a smaller amount (in our experiment a 50% discount on the next ride). 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 riders, etc.) and identifying the optimal level of compensation is an interesting question for future research.

Finally, the last metric (average time interval between rides) shows the same pattern as in Experiment 1 but was not statistically significant in Experiment 2. For conciseness, we omit these results.

4.2.3. Regression analysis. Next, we investigate whether our main findings continues to hold even after controlling for some of the potential factors that may affect the engagement behavior. To this end, we run several regression analyses to examine how different types of promotions affect rider engagement. As before, we focus on the first two key engagement metrics, as our dependent variables (S_T^j and R_T^j) within the first 4 weeks after being exposed to the experiment (i.e., $T = 4$). The regression model specification is given by:

$$\begin{aligned} y_i = \alpha + \beta_1 \text{Discount}_i + \beta_2 \text{Credit}_i + \\ + \gamma_1 \text{Preexperiment-Rides}_i + \gamma_2 \text{Preexperiment-Rides}_i^2 + \mu_i + \epsilon_i, \end{aligned} \tag{6}$$

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 R_T^j); and Discount_i, Credit_i represent binary variables for each experiment condition (the Control condition is the reference group). Note that we control for

[31] For 3 weeks: $F(2,920) = 4.65$, $p<.01$, for 2 weeks: $F(2,920) = 5.72$, $p<.01$, and for 1 week: $F(2,920) = 5.05$, $p<.01$, and the results of the pairwise comparisons are similar to the results for 4 weeks.

[32] For the post-experiment data: For 3 weeks: $F(2,920) = 3.33$, $p<.05$, for 2 weeks: $F(2,920) = 3.99$, $p<.05$, and for 1 week: $F(2,920) = 3.77$, $p<.5$. For the pre-experiment data: For 4 weeks: $F(2,920) = 0.42$, $p = 0.65$, for 3 weeks: $F(2,920) = 0.25$, $p = 0.78$, for 2 weeks: $F(2,920) = 0.56$, $p = 0.57$ and for 1 week: $F(2,920) = 0.92$, $p = 0.4$.

the riders' pre-experiment usage by including the total number of rides during the pre-experiment period (of 4 weeks) as well as a quadratic term to capture the potential non-linear effect. Finally, we include date/time fixed effects, which vary at the rider level, by capturing the date on which rider i was exposed to the experiment (variable μ_i).

Table 5 Regression results for Experiment 2.

OLS with pooled samples	(1) Model1 DV: Total Rides	(2) Model2 DV: log(Total Rides)	(3) Model3 DV: log(Total Rides)
Discount	0.909*	0.199***	0.245***
	(0.468)	(0.0696)	(0.0681)
Credit	0.897*	0.160**	0.162**
	(0.507)	(0.0733)	(0.0730)
Preexperiment-Rides	0.444***	0.0682***	
	(0.0453)	(0.00734)	
Preexperiment-Rides ^{2}	-0.00197***	-0.000480***	
	(0.000634)	(0.000129)	
Log (Preexperiment-Rides)			0.617***
			(0.0297)
Date FE	Yes	Yes	Yes
Constant	1.778	0.732***	0.0204
	(1.349)	(0.224)	(0.235)
Observations	923	923	923
R-squared	0.353	0.306	0.325

The results of the OLS regression are reported in Table 5. Consistent with the findings from the one-way ANOVA tests, the coefficient of the Credit and Discount indicators are positive and statistically significant, which means that the riders in the Credit and Discount conditions complete a larger number of rides relative to the riders in the Control condition. For robustness purposes, we also consider the regression equation with a log transformation in the dependent and independent variables which have a non-binary value (see the last two columns of Table 5). Finally, since the dependent variable y_i can take only positive integer values, we run a Poisson regression with the same specification as in (6) (based on the assumption that the error term ϵ_i follows a Poisson distribution). We observed that the results of the Poisson regression are consistent with the OLS regression. As before, the pre-experiment level of usage shows a concave pattern.

Next, we 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 usage (i.e., the total number of rides).[33] Interestingly, as one can see from Table 6, the coefficient for the Credit and Discount conditions are significant only for the high group but not for the low group. This result confirms once again that (i) the effect of the Credit and Discount conditions on the frustration is moderated by the pre-experiment usage and that (ii) this effect is present only for the experienced riders who regularly use the service.

[33] We also checked the results when we use the total spending instead of the total number of rides and observed the same qualitative results.

Table 6 Regression results by pre-experiment usage for Experiment 2.

OLS by pre-experiment usage Variables	(1) Frequent Riders (more than 3 rides) DV: Total Rides	(2) Non-frequent Riders (less than 3 rides) DV: Total Rides	(3) Non-frequent Riders (less than 3 rides) DV: Total Rides
Discount	**1.833*** **	-0.0451	-0.0451
	(0.620)	(0.601)	(0.601)
Credit	**1.367****	0.429	0.429
	(0.642)	(0.692)	(0.692)
Preexperiment-Rides	**0.359*** **	0.811	
	(0.0443)	(0.503)	
Preexperiment-Rides ^{2}	**-0.00109****		0.270
	(0.000477)		(0.168)
Constant	2.537***	1.435*	1.976***
	(0.449)	(0.815)	(0.581)
Observations	628	295	295
R-squared	0.293	0.012	0.012

In summary, even after controlling for the potential factors that may influence the engagement behavior, we still observe that the riders in the Credit and Discount conditions are more likely to complete a larger number of rides (and to spend more) relative to the riders in the Control condition. This effect is moderated by the pre-experiment usage behavior. We find that only the riders in the high group (i.e., the riders who ride more often) show a statistically significant effect.

4.2.4. Difference-in-differences. Similar to the analysis presented for Experiment 1, we use a difference-in-differences approach to identify the causal effect of the different compensation conditions. The model specification is given by:

$$\begin{aligned} y_{it} =& \alpha + \beta_1 \text{Discount}_i + \beta_2 \text{Credit}_i + \gamma \text{After-experiment}_t + \\ &+ \delta_1 \text{Discount}_i \cdot \text{After-experiment}_t + \delta_2 \text{Credit}_i \cdot \text{After-experiment}_t + \epsilon_{it}, \end{aligned} \tag{7}$$

where i corresponds to a rider and t to a day. y_{it} denotes the dependent variable for rider i at time t (we consider both the total spending and the total number of rides). Discount_i and Credit_i are binary variables to indicate the condition assigned to rider i (as before, the Control condition is the reference group). $\text{After-experiment}_t$ is a binary variable for the time period after being exposed to the experiment. As before, the key parameters in equation (5) are δ_1, δ_2, and δ_3. These parameters capture the potential causal effects of each type of compensation (following the frustration) on the engagement behavior.

As shown in Table 7, the coefficients of the interaction between the Credit condition and the After-experiment as well as between the Discount and the After-experiment are positive and statistically significant. This implies that the riders in the Credit and Discount conditions use the service more (and spend more) relative to the riders in the Control condition during the post-experiment period. Consequently, the effect of the Credit and Discount conditions in response to the frustration are causal. As a result, this confirms that the Credit and Discount compensations does affect (positively) the engagement of riders who experienced a frustration.

Table 7 Difference-in-Differences results for Experiment 2.

Diff-in-Diffs	(1) Model1 DV: Total Rides	(1) Model2 DV: Total Spending
Discount	0.00288	0.00896
	(0.00762)	(0.0265)
Credit	0.0141*	0.0630**
	(0.00790)	(0.0275)
After-Experiment	0.00113	0.0541**
	(0.00778)	(0.0271)
Discount*After-Experiment	0.0392***	0.0817**
	(0.0108)	(0.0375)
Credit*After-Experiment	**0.0312*** **	**0.113*** **
	(0.0112)	**(0.0390)**
Constant	0.170***	0.548***
	(0.00550)	(0.0192)
Observations	53,534	53,534
R-squared	0.001	0.002

4.2.5. Non-frustration. We next investigate how non-frustrated riders are different from frustrated riders in terms of engagement behavior before and after being exposed to the experiment. We use the same procedure as in Section 4.1.5 to select a sample of non-frustrated riders, whereas the frustrated riders are represented by the riders in the Control condition. Specifically, we have a total of 305 and 285 riders in the frustrated and non-frustrated groups, respectively. For robustness purposes, we define the non-frustrated group based on either the total spending or the total number of rides during the pre-experiment period (both approaches show a consistent pattern). As before, we select the sample of non-frustrated riders by using the maximum ETA error during T weeks (we also use the average ETA error instead of the maximum and vary the value of T between 1 and 6). The left (resp. right) panel in Figure 12 reports the average total spending (resp. total number of

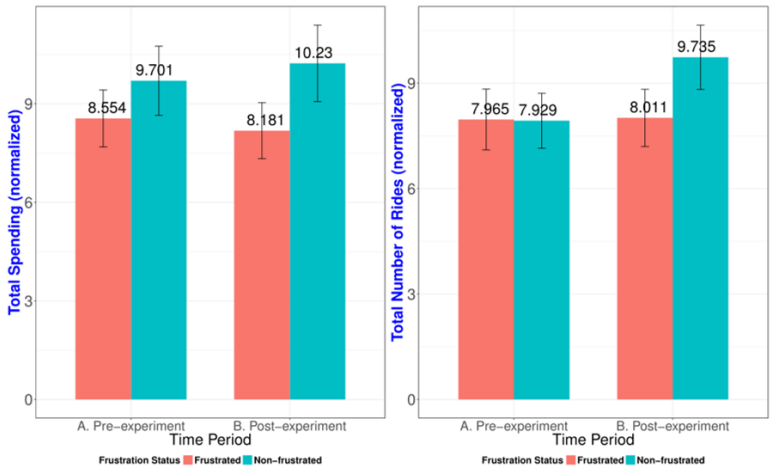

Figure 12 Average total number of rides by frustration status for Experiment 2.

rides) for both frustrated and non-frustrated riders during the pre- and the post-experiment periods. As expected, in the pre-experiment period, the frustrated group is not statistically different from the non-frustrated group. Frustrated riders spend on average 9.6% less relative to non-frustrated riders ($t(592) = 1.16,\ p = 0.25$, i.e., not statistically significant).

However, after they experience a frustration, frustrated riders spend 25.04% less relative to the non-frustrated group ($t(592) = -2.32,\ p<.05$, i.e., significant at the 95% level). The same pattern holds for the average total number of rides.[34] This allows us to conclude that experiencing a frustration negatively affects the engagement behavior.

4.3. Experiment 3: Viaversary

As mentioned in Section 2.2, this experiment includes a total of 605 riders in NYC. After applying our filter (see Section 3.3), we are left with 599 riders divided as follows: Control (175) and Credit (424). For simplicity, we next report the results for the total spending and the total number of rides during $T = 4$ weeks after being exposed to the experiment.

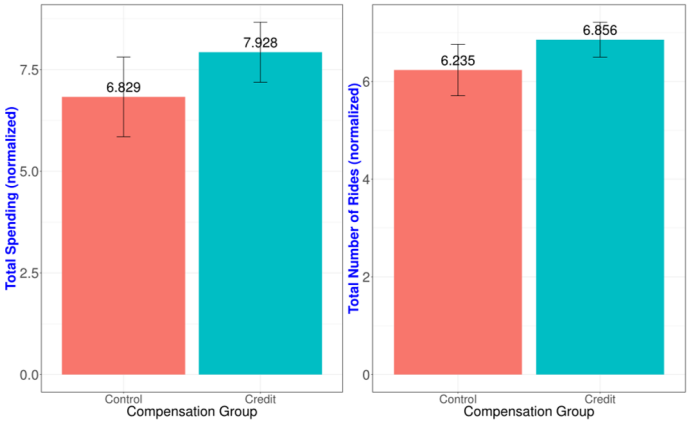

Figure 13 Results for Experiment 3.

As we can see from Figure 13, riders in the Credit condition take on average 9.96% more rides relative to riders in the Control condition during the first 4 weeks after being exposed to the experiment. However, this difference is not statistically significant ($t(597) = -.95,\ p = .34$). The same result applies to the average total spending. This implies that offering a compensation to non-frustrated riders (in this case, to riders who celebrate their joining date anniversary) does not have a significant impact on their engagement behavior. For robustness purposes, we vary the time

[34] Regarding the pre-experiment period, the frustrated riders complete on average 0.45% more rides relative to the non-frustrated riders; $t(592) = 1.34,\ p = 0.18$ (not statistically significant). Regarding the post-experiment period, the frustrated riders complete on average 21.52% less rides relative to the non-frustrated riders; $t(592) = -1.84,\ p<0.1$ (significant at the 90 % level).

window from 4 weeks to 1 week by increments of 1 week and observed consistent results across the different time windows.

Next, we conduct the same analysis by splitting the riders into high and low groups based on their pre-experiment usage. For example, we divide the riders into two groups by using the median split (we also vary this number up to top 25% by increment of 5%). As before, we observed that the Control and Credit conditions are not statistically different neither for the high group nor for the low group (the results of this test are very similar to Figure 13 and are not reported for conciseness).

We conclude this section by presenting one final comparison in terms of the impact of compensating frustrated riders versus Viaversary riders. More precisely, we compute the difference in the total number of rides between the Credit and Control conditions for Experiment 3 (i.e., Viaversary riders) as well for Experiment 1 (i.e., frustrated riders). The mean and standard deviation of the difference in the total number of rides between the Credit and Control conditions in Experiment 3 are equal to: Mean = 0.621, Standard Deviation = 9.868 (the total number of subjects is equal to 599). For Experiment 1, we have: Mean = 1.249, Standard Deviation= 12.712 (the total number of subjects is equal to 2,307). Based on this analysis,[35] the impact of the Credit condition (i.e., proactively sending a $5 credit) on the frustrated riders yields on average 1.249 - 0.621 = 0.628 extra ride relative to the gain for Viaversary riders. The practical implication from this finding is clear: Given a limited budget of promotions, it is more effective to allocate the promotions to riders who have experienced a frustration. In particular, the promotion will have a deeper impact in terms of boosting their engagement behavior.

5. Conclusion

In this paper, we investigated whether a service provider should proactively compensate users who experienced a frustration (i.e., low level of service). When a user experiences low service quality, the future engagement of this particular user is at risk. A possible strategy for the service provider is to proactively send a 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 one of the leading ride-sharing platforms, Via. We designed and ran three field experiments to study the impact of compensating riders who had experienced a frustration. Motivated by historical data and by the ride-sharing context, we

[35] By conducting a t-test, we found that the result is statistically significant at the 95% confidence level ($t(2902) = 1.754$).

considered two types of frustration: long waiting times and long travel times. We measured rider engagement by computing total spending, total number of rides, and the average time interval between rides (all during T weeks after being exposed to the experiment). By conducting ANOVA tests, regression analyses, and a difference-in-differences approach, we find that sending a compensation to frustrated riders (i) is profitable and boosts their engagement behavior (relative to not sending a compensation), (ii) works well for long waiting times but not for long travel times, (iii) is more effective than sending the same offer to non-frustrated riders, and (iv) has an impact that is moderated by past usage frequency. We also observed that the best strategy is to send credit for future usage (as opposed to waiving the charge or sending an apologetic message).

This paper is among the first to rigorously investigate the impact of proactively sending compensations to frustrated customers in the ride-sharing market. The results presented in this paper allow us to draw practical insights on best practices for proactive campaigns related to service quality. Besides boosting the engagement behavior, this type of compensation leads to additional benefits in terms of customer satisfaction. When receiving such a compensation, users are often pleasantly surprised and feel that the service provider is looking after them. Below are few of the text messages users sent to Via after receiving the compensation:

"That'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."

As one can see, users appreciate the reward, and this practice can make the difference in a competitive industry. It is worth mentioning that such users are more likely to recommend the service to their friends and hence help in regard to gaining market share. 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 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.

Acknowledgments

We would like to thank several people at Via's NYC office for insightful discussions and feedback. In particular, we are thankful to Saar Golde, Ori Klein, Alex Lavoie, and Gabrielle McCaig.

References

Andrews M, Luo X, Fang Z, Ghose A (2015) Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science* 35(2):218–233.

Angrist JD, Pischke JS (2008) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Arora N, Dreze X, Ghose A, Hess JD, Iyengar R, Jing B, Joshi Y, Kumar V, Lurie N, Neslin S, et al. (2008) Putting one-to-one marketing to work: Personalization, customization, and choice. *Marketing Letters* 19(3-4):305.

Banerjee S, Riquelme C, Johari R (2015) Pricing in ride-share platforms: A queueing-theoretic approach, working paper.

Benjaafar S, Bernhard H, Courcoubetis C (2017) Drivers, riders and service providers: The impact of the sharing economy on mobility, working paper.

Benjaafar S, Ding JY, Kong G, Taylor T (2018) Labor welfare in on-demand service platforms, working paper.

Berry LL, Parasuraman A (2004) *Marketing services: Competing through quality* (Simon and Schuster).

Bimpikis K, Candogan O, Daniela S (2016) Spatial pricing in ride-sharing networks, working paper.

Bolton RN (1998) A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing science* 17(1):45–65.

Cachon GP, Daniels KM, Lobel R (2017) The role of surge pricing on a service platform with self-scheduling capacity, forthcoming in Manufacturing & Service Operations Management.

Chen MK, Chevalier JA, Rossi PE, Oehlsen E (2017) The value of flexible work: Evidence from uber drivers, working paper.

Chen MK, Sheldon M (2016) Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. *EC*, 455.

Cohen MC, Zhang RP (2017) Coopetition and profit sharing for ride-sharing platforms, working paper.

Fisher M, Gallino S, Li J (2017) Competition-based dynamic pricing in online retailing: A methodology validated with field experiments, forthcoming in Management Science.

Fong NM, Fang Z, Luo X (2015) Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research* 52(5):726–735.

Furuhata M, Dessouky M, Ordóñez F, Brunet ME, Wang X, Koenig S (2013) Ridesharing: The state-of-the-art and future directions. *Transportation Research Part B: Methodological* 57:28–46.

Gallino S, Moreno A (2015) The value of fit information in online retail: Evidence from a randomized field experiment, working paper.

Hall JV, Horton JJ, Knoepfle DT (2017) Labor market equilibration: Evidence from uber, working paper.

Hu M, Zhou Y (2017) Price, wage and fixed commission in on-demand matching, working paper.

Kohavi R, Deng A, Frasca B, Walker T, Xu Y, Pohlmann N (2013) Online controlled experiments at large scale. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1168–1176 (ACM).

Kohavi R, Thomke S (2017) The surprising power of online experiments. *Harvard Business Review* 95(5):74–81.

Maxwell SE, Delaney HD (2004) *Designing experiments and analyzing data: A model comparison perspective*, volume 1 (Psychology Press).

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

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

Singh J, Teng N, Netessine S (2017) Philanthropic campaigns and customer behavior: Field experiments on an online taxi booking platform, forthcoming in Management Science.

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

Tang CS, Bai J, So KC, Chen XM, Wang H (2016) Coordinating supply and demand on an on-demand platform: Price, wage, and payout ratio, working paper.

Taylor T (2016) On-demand service platforms, working paper.

Zahorik AJ, Rust RT (1992) Modeling the impact of service quality on profitability: a review. *Advances in services marketing and management* 1(1):247–76.

Zeithaml VA, Berry LL, Parasuraman A (1996) The behavioral consequences of service quality. *the Journal of Marketing* 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.