NBER WORKING PAPER SERIES

SERVICE QUALITY ON ONLINE PLATFORMS: EMPIRICAL EVIDENCE ABOUT DRIVING QUALITY AT UBER

Susan Athey
Juan Camilo Castillo
Bharat Chandar

Working Paper 33087
http://www.nber.org/papers/w33087

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2024

We are grateful for funding from the Sloan Foundation and the Stanford Cyber Initiative. The paper was initiated while Bharat Chandar was an employee of Uber. He no longer retains equity in the company. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w33087

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2024 by Susan Athey, Juan Camilo Castillo, and Bharat Chandar. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Service Quality on Online Platforms: Empirical Evidence about Driving Quality at Uber
Susan Athey, Juan Camilo Castillo, and Bharat Chandar
NBER Working Paper No. 33087
October 2024
JEL No. J28, J48, L50, L91, R41

ABSTRACT

Online marketplaces have adopted new quality control mechanisms that can accommodate a flexible pool of providers. In the context of ride-hailing, we measure the effectiveness of these mechanisms, which include ratings, incentives, and behavioral nudges. Using telemetry data as an objective measure of quality, we find that drivers not only respond to user preferences but also improve their behavior after receiving warnings about their low ratings. Furthermore, we use data from a randomized experiment to show that informing drivers about their past behavior improves quality, especially for low-performing drivers. Lastly, we find that UberX drivers exhibit behavior comparable to that of UberTaxi drivers, suggesting that Uber's new quality control mechanisms successfully maintain a high level of service quality.

Susan Athey
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
athey@stanford.edu

Juan Camilo Castillo
Department of Economics
University of Pennsylvania
133 South 36th Street
Philadelphia, PA 19104
and NBER
jccast@upenn.edu

Bharat Chandar
Stanford University
579 Jane Stanford Way
Stanford, CA
chandarbharatk@gmail.com

1 Introduction

Online marketplaces have transformed many industries. Platforms like Uber, TaskRabbit, and Airbnb now allow small, independent providers to sell or rent underutilized goods, work during their free time, and smooth their income during unemployment spans (Katz and Krueger, 2016; Chen et al., 2019). When the provider base expands and diversifies and with sufficient competition, potential benefits to consumers include lower prices, more variety, and more flexibility.

Offering goods and services from independent, flexible providers also poses challenges for platforms. They need to convince consumers that strangers can provide high quality. The traditional approach of relying on ex ante screening has disadvantages for a model that aims to make use of a flexible pool of independent service providers, many of whom are new to the service or are working part time. New platforms thus follow a streamlined screening process, and to ensure a high quality they instead employ a variety of new technology-enabled methods, harnessing them to shape incentives and provide nudges to providers. For example, they use rating systems extensively, allowing customers to monitor quality through simple interfaces. They combine ratings with incentive systems that remove individuals who violate quality standards and reward those who do well. Some companies also collect objective, real-time measures of quality they can share with their providers to nudge them to perform well.

In this paper, we investigate how well the ex-post quality control methods employed by online marketplaces effectively bring about a high service quality, focusing on the context of rides intermediated by Uber. We use a combination of natural experiments and designed randomized experiments to assess the extent to which service providers—drivers—respond to different types of information, incentives, and nudges, shedding light on the mechanisms that induce a higher quality.

The ride-hailing setting that we study is especially well-suited for our purposes. Unlike many other markets, where only subjective measures of quality like ratings or reviews are available, in this setting we observe objective measures of quality in a large sample: Uber collects telemetry metrics such as speed, acceleration, braking, phone handling, and routing. In contrast to subjective measures, our telemetry metrics do not vary endogenously with equilibrium reporting behavior or incentives. This allows us to exe-

cute a detailed analysis of different methods of ex-post quality control and their impact on these objective telemetry metrics.

As a building block towards our main results, we first analyze riders' preferences for driving behavior. This allows us to determine what type of behavior should be considered high quality, and, thus, what driving behavior Uber would like to induce. To that end, we use trip rating as a measure of satisfaction. At the end of each trip, riders get the chance to rate the driver on a scale between one and five stars. After controlling for driver, origin, destination, and time of the week, we find that riders give higher ratings to trips with fewer strong brakes and accelerations. They also prefer trips where the driver does not handle the cell phone. These results are consistent with riders preferring safer trips. Riders also give higher ratings when drivers drive at a steady, intermediate speed, suggesting that there is some tension between safety and arriving quickly at the destination. Finally, riders prefer shorter trips, both in duration and distance, and they prefer pickups and drop-offs closer to their requested locations. Based on these findings, we build a score that aggregates driving metrics into a one-dimensional measure of driving quality based on a predictive model for the trip rating.

We then move on to the core of our paper, where we quantify the impact of Uber's quality control mechanisms. We first analyze the role of incentives and information provided through ratings. When drivers' ratings fall below certain thresholds, Uber sends them notifications with the aim of improving their behavior, and if ratings keep decreasing, drivers are eventually taken off the platform. We find, first of all, that the drivers that Uber removes—those with persistently low ratings—behave significantly worse than the average driver based on driving metrics and scores. Thus, ratings provide Uber with a simple criterion to remove poorly performing drivers from the platform.

We also find that drivers' quality improves substantially—in terms of both scores and metrics—after receiving notifications about their low ratings: they handle their phones less, accelerate less, drive at steadier speeds, take shorter routes, and pick up and drop off passengers closer to their intended locations. These effects are strong even for the first notification, which occurs far away from the threshold at which they are removed from the platform. Thus, we conclude that, besides direct economic incentives, ratings and notifications work as behavioral nudges that induce better driving behavior. Importantly, the effects of notifications are long-lasting: they persist even after drivers receive

notifications from Uber informing them that they are no longer at risk of being taken off the platform once their ratings improve.

Giving drivers feedback about their past behavior also has a positive impact on driving behavior. At the time our data was collected, Uber sent a weekly report to drivers summarizing how they performed according to telemetry data. Uber conducted a large-scale randomized experiment that introduced a significant upgrade to these reports. Treated drivers gained access to a dashboard within the Uber app where they could analyze their driving behavior for individual trips. We find an improvement in the behavior of treated drivers. Furthermore, the effect is stronger for drivers who performed in the bottom 10th percentile before the experiment started.

All these results indicate that Uber's quality control mechanisms positively impact driving behavior. However, it remains possible that, despite these mechanisms, Uber provides poor overall quality due to the lack of ex-ante screening. The rest of our paper examines whether that is the case. We take advantage of the fact that, apart from UberX (Uber's main ride-hailing service), Uber also offers a product called UberTaxi that allows riders to request standard taxis licensed by the local government. Whereas the quality provided by UberX is largely controlled through ratings, incentives, and nudges, UberTaxi mostly relies on traditional screening through the licensing process. Differences in driving behavior between both groups cannot be directly interpreted as the effect of different quality control methods because there are several other differences between both products. Drivers belong to different populations with different demographic characteristics. Monetary incentives also differ: UberX drivers are much more likely to own their car, and fare structures are different. Nevertheless, a direct comparison between the quality provided by both products allows us to determine the viability of a ride-hailing service that relies almost entirely on ex-post quality control mechanisms with limited ex-ante screening.

We find that the overall quality of UberX and UberTaxi trips, as measured by our scores, is roughly equivalent. However, important differences underlie this result. UberX drivers brake and accelerate less, tend to speed less, and pick up and drop off drivers closer to where they want, all of which are desirable behaviors. On the other hand, they handle their phone more and tend to take longer routes. Our scores suggest that riders prefer the way UberX drivers control their vehicle but favor the routing behavior

of UberTaxi drivers.[1] On balance, riders are roughly indifferent between the behavior of drivers of both products. These results confirm that Uber's quality control systems are sufficient to ensure a level of quality comparable to that of taxis, despite not relying on ex-ante screening.

We also measure heterogeneous differences between UberX and UberTaxi drivers to provide additional evidence that Uber's quality control systems operate as intended. We find that the gap between UberX and UberTaxi is smaller during rush hour, in particular for speed and hard brakes, when riders are more likely to be in a hurry. This is consistent with UberX providing stronger incentives to cater to riders' preferences, and speeding up at that time.

Our findings have important managerial implications. Our main results show that there are several mechanisms platforms can use to ensure a high quality. They can use rating systems coupled with messages to low-rated providers and a deactivation policy, as well as real-time monitoring and feedback to nudge providers to improve their quality. Our results also deliver key messages for regulators concerned about quality. We find that markets that mostly use ex-post quality control systems can provide similar quality as traditional markets that rely on ex-ante screening and occupational licensing. Regulators should thus consider allowing marketplaces to use well-designed incentives, nudges, and information systems as effective substitutes for rigorous ex-ante screening.

Related Work In a paper that is closely related to ours, Liu et al. (2021) find that Uber drivers are less likely than taxi drivers to take detours from airport trips. The main implication is that Uber's incentive systems are effective in eliminating this kind of moral hazard. In contrast to their work, we analyze driving behavior, a different type of incentive problem, and we focus on the mechanisms through which incentive systems affect behavior.

Our work is part of a broader literature that analyzes rating and review systems in the digital economy (Dellarocas, 2006; Tadelis, 2016). Many of these papers focus on the behavior of consumers when they rank providers (Resnick et al., 2006; Resnick and Zeckhauser, 2002; Filippas et al., 2022; Nosko and Tadelis, 2015; Proserpio and Zervas,

[1]Since our scores are derived from UberX trips (which account for 98% of our data), the quality should be interpreted as reflecting the preferences of UberX riders; unfortunately our UberTaxi data are insufficient to reliably compare preferences across categories of riders.

2017; Zervas et al., 2021) and on how consumers respond to ratings and rankings, such as in the context of online bookstores (Chevalier and Mayzlin, 2006) and restaurants (Luca, 2011). In contrast, we focus on provider behavior and how it is influenced by ratings and reviews.[2]

A broad literature analyzes how nudges and information affect agents' behavior beyond what can be explained by incentives (Thaler and Sunstein, 2009; Allcott and Kessler, 2019). Two papers focus on driving behavior, with findings that are consistent with some of our results. Jin and Yu (2021) find that text messages can nudge ride-hailing drivers to reduce cell-phone handling, and Choudhary et al. (2022) find that private car drivers respond positively to information on past behavior, especially if they were not good drivers previously. Besides these empirical findings, a key contribution of our paper is to highlight their significance in the context of quality provision in the gig economy.

Several other papers provide evidence that agents react to nudges and information in different contexts, such as surgical performance (Kolstad, 2013), energy consumption (Allcott and Rogers, 2014), health insurance (Handel, 2013), taking medicine (Macharia et al., 1992), voting (Gerber et al., 2003), and charitable donations (Shang and Croson, 2009; Frey and Meier, 2004; Edwards and List, 2014). Our work shows evidence of one setting where nudges, coupled with other mechanisms, are used by companies to achieve an outcome that is desirable to consumers.

A growing literature analyzes several aspects of Uber and ride-hailing markets. Our work most closely relates to those that focus on the labor supply side (Hall and Krueger, 2016; Chen and Sheldon, 2015; Hall et al., 2023; Cook et al., 2020). A recurring theme is the value of labor flexibility introduced by ride hailing platforms (Angrist et al., 2021; Chen et al., 2019). This kind of flexibility could in principle be accompanied by a reduction in quality; the evidence in our paper establishes that this potential cost is not borne out in practice and documents several mechanisms that contribute to Uber's ability to maintain quality.

[2]Jin and Leslie (2003) analyze providers' response to the regulation that existed before digital markets. Mayzlin et al. (2014) and Luca and Zervas (2016) analyze a different kind of provider behavior: they find evidence that hotel and restaurant owners post positive reviews for themselves and negative reviews for competitors.

Roadmap Section 2 describes the Uber market and the data we use. In preparation for our main analysis, in Section 3 we analyze riders' preferences over driving metrics and construct the driving quality scores that we use as our main outcome variables. Section 4, the core of our analysis, measures how online marketplaces' ex-post quality control methods—ratings, incentives, and behavioral nudges—affect driving behavior. In Section 5, we compare the performance of UberX and UberTaxi to provide further evidence on the success of these new quality control mechanisms. Finally, we conclude in Section 6.

2 Uber, telemetry, and ratings

We analyze the Uber market in Chicago during the first half of 2017. We limit our analysis to a region around downtown Chicago that is large enough to include Midway and O'Hare, the main airports in the region, but which excludes some far-away suburbs (Figure 1).

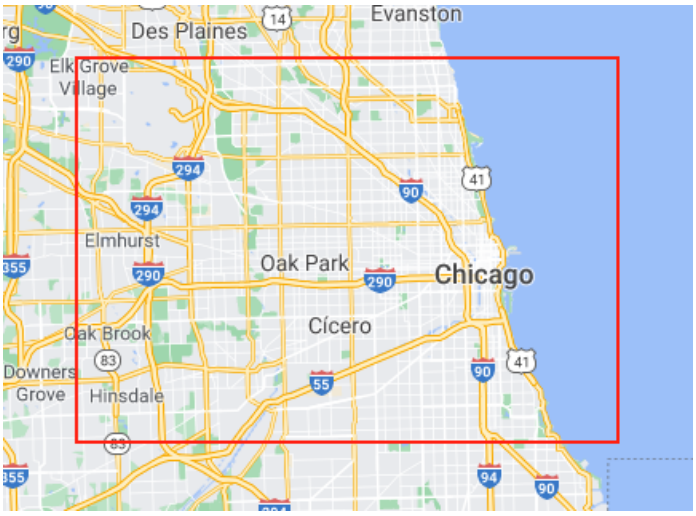


Figure 1: Region of analysis

Note: This figure highlights in red the region in Chicago we focus on.

Our data involve two Uber products. Our primary analysis focuses on the first product, UberX, which is Uber's main ride-hailing option and its largest product by number of rides. Drivers typically own their cars, and entry requirements are fairly simple: being at least 21 years old; having a driver's license, vehicle registration and insurance; and passing an online screening that reviews driving record and criminal history. The second

product is UberTaxi, which matches riders to taxis licensed by the City of Chicago. Taxi drivers must be Chicago public chauffeurs, which involves a lengthy licensing process.[3] There are two common ownership models for taxis in Chicago. Some drivers are independent operators and own both the taxi and the medallion. Most other drivers lease both the taxi and the medallion for 12 hour shifts at a flat fee and retain all earnings from working during the shift.

We focus on trips over a several month period in early 2017. The number of UberTaxi trips is much smaller than the number of UberX trips, but this is true for all cities; we chose Chicago as our region of analysis because it is the market with the largest number of UberTaxi trips. Our main dataset includes 6,901,200 UberX trips and 139,716 UberTaxi trips after filtering out trips in which any driving metrics, which we describe below, are missing.

Quality metrics Uber collects telemetry data from drivers' smartphones every two seconds while the Uber app is open. The raw data includes location, speed, and acceleration.[4] Our analysis focuses on ten trip-level metrics, whose distributions are shown in Figure 2. The first two metrics are the *accelerations* metric, the fraction of acceleration events during the trip where acceleration went above 2 m/s^2, and the *brakes* metric, which is defined similarly.[5]

Uber developed two classifiers based on accelerometer data that predict whether a driver's cell phone was mounted and whether the driver was handling the cell phone (i.e., moving it while holding it with his hands). We define the *handling* and *mounted* metrics as the average of the predictions from these classifiers over the two middle time quartiles of a trip. We do not take into account what happens at the beginning and at the end of a trip because drivers are especially likely to use their phone at the beginning and end of a trip, but often in a way that does not necessarily interfere with safety or contribute to passenger dissatisfaction.

[3]The driver must at minimum be 21 years of age, possess an active, permanent driver's license in good standing, pass a national background check, pass a two-week public chauffeur course and licensing exam, have an authorized debt clearance or payment plan, and be in good standing with court-order child support payments.

[4]Telemetry data may be less accurate when the phone is not mounted, though this problem is mitigated by using telemetry from the GPS instead of accelerometers.

[5]An acceleration event takes place when speed increases during two consecutive 2 second intervals. A braking event is defined similarly.

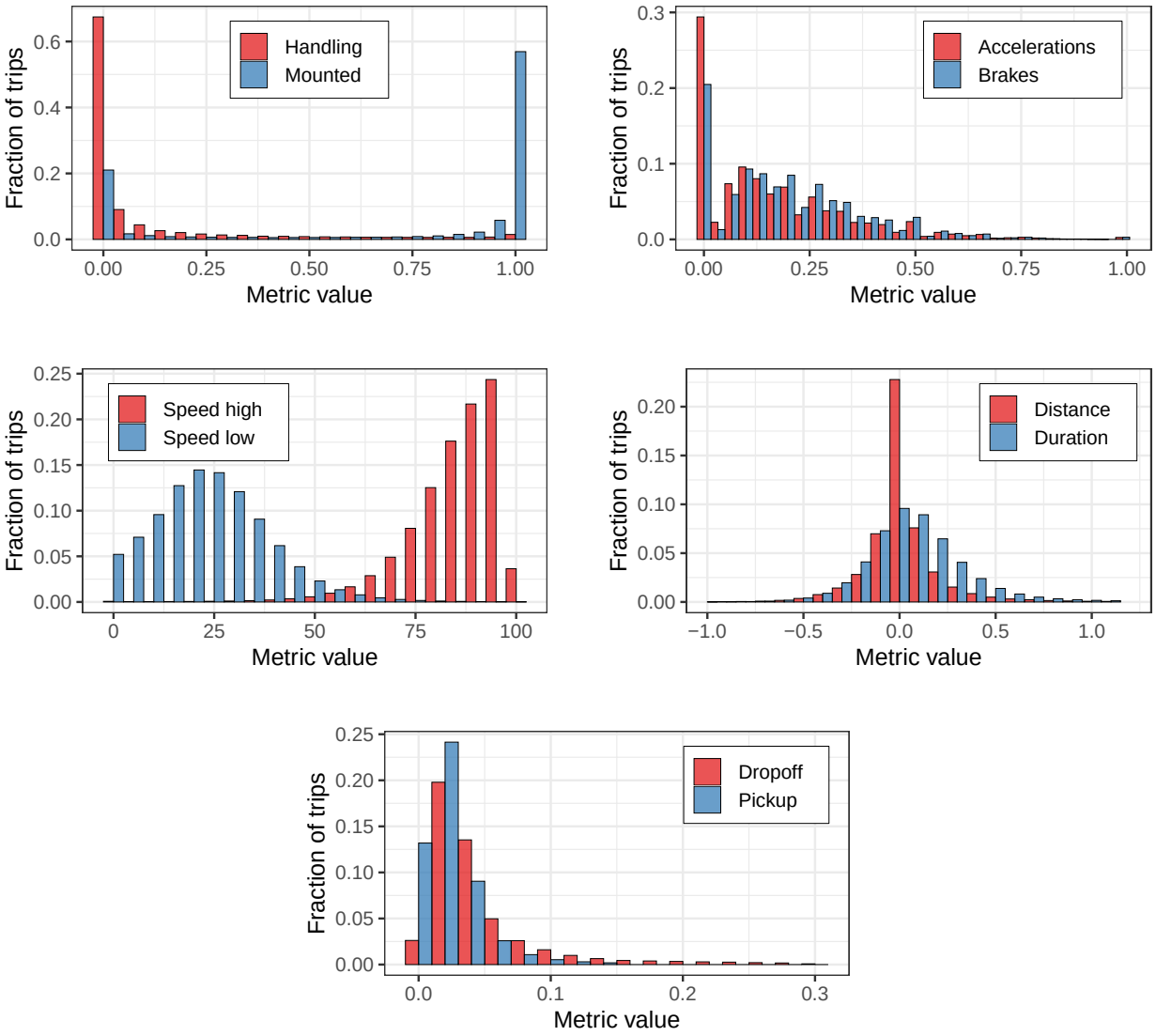


Figure 2: Distribution of the driving metrics for UberX trips

Our metrics for speed are based on a model developed by Uber to construct speed metrics for each segment (which roughly corresponds to a block) relative to other cars that went through the same segment. The *contextualized speed* for a segment is the percentile within the distribution of speeds for all cars that went through that segment. We focus on two metrics at the trip level: *Speed low* and *speed high*, the 10th and 80th percentiles, respectively, of all contextualized speeds in a trip. They measure how fast the driver was going during the slowest and fastest segments of the trip, relative to other traffic.

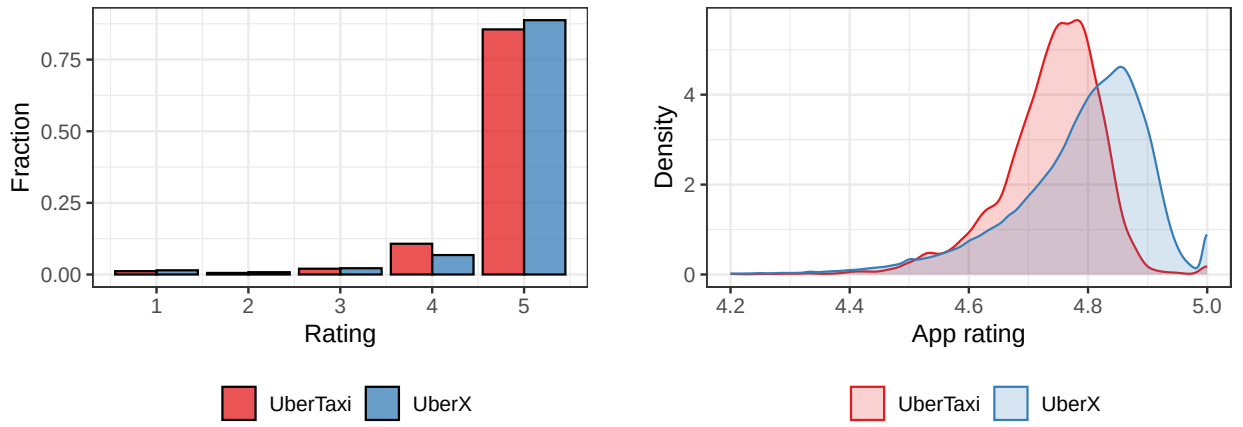
We also define four metrics that are related to the route taken by the driver. The first two assess whether the driver chose a good route: our *distance* metric, the log of the ratio of the actual trip distance to the distance Uber estimated before the trip started, and our *duration* metric, the log of the ratio of the actual duration to the duration Uber estimated before the trip. Uber does not estimate distances and durations for UberTaxi trips, so in analyses that include UberTaxi trips, we use a modified version of these metrics that is based on estimated distances and durations that we impute based on a random forest prediction (see Appendix B.3). The other two metrics are the *pickup* and *dropoff* metrics, which measure the distance between the requested pickup point and the place where the pickup actually took place, and the distance between the requested dropoff point and the place where the rider was actually dropped off.[6]

There is a large number of alternative metrics we could have used to measure brakes, accelerations, and speed. For instance, we could have used thresholds other than 2 m/s^2 to measure brakes, or we could have used different contextualized speed percentiles. We selected our main variables based on a regularized regression model (Lasso) to predict a trip's rating based on all these variables and their squares. We selected the variables that dropped out last as the penalty increased. The details of this procedure are in Appendix B.1. Our main results, however, do not change when we choose alternative metrics.

Ratings After every trip, the passenger can give a one to five star rating to the driver. 28.8% of the trips in our sample were rated. Figure 3a shows the distribution of ratings. Most trips receive five stars, consistent with the behavior studied by Filippas et al. (2022). UberTaxi trips tends to get a larger fraction of 4 star ratings.

Uber uses the *app rating*, the average of the last 500 trips, as its main measure of driver quality; drivers can see it in the app, and it is shown to passengers upon being matched to a driver before pickup. Figure 3b shows its distribution. Uber stops assigning trips to UberX drivers ("deactivation" of a driver) when ratings drop below certain thresholds, but only after a process that involves giving notifications to drivers when their ratings approach the deactivation threshold. We describe the process in detail in Section 4.1.

[6]We exclude from our sample trips in which the pickup metric is greater than 150 meters or the dropoff metric is greater than 300 meters, since these are likely trips in which the rider decided to change the origin or destination. We also exclude trips in which the absolute value of the distance metric is greater than 0.7 or in which the absolute value of the duration metric is greater than 0.8.



(a) Trip rating

(b) App rating

Figure 3: Distribution of ratings

Note: Subfigure (a) shows histograms of the ratings given to individual UberX and UberTaxi trips. Subfigure (b) shows kernel density plots of drivers' app rating (the average of the last 500 trips).

3 Driving behavior and rider preferences

In this section, we explore what driving behavior riders prefer. This is a key step toward our main results because it allows us to determine what should be considered high-quality behavior. Based on that analysis, we build one-dimensional scores that summarize driving behavior, and we use them as outcome variables in the rest of the paper.

3.1 Rider preferences

In order to determine rider preferences for driving behavior, we use different methodologies to predict the rating a rider will give to a trip, based on the observed driving metrics. This procedure faces a variety of problems. One challenge is that different kinds of trips (at different times of the day, with different origin and destination) lead to different driving behavior, such as slower trips downtown or during rush hour, some of which might be intrinsically more satisfying for riders. A second challenge arises due to unobserved characteristics that have an effect on satisfaction, such as the driver's personality, which might be correlated with driving behavior.

We address both issues by controlling for the driver and the type of trip to difference out fixed characteristics of the driver that lead to higher satisfaction (such as the comfort

of the car) and fixed characteristics of the trip (such as traffic or time of the day).[7] We thus estimate how *changes* in driving metrics lead to changes in consumer satisfaction, fixing the driver and the type of trip.

To control for the type of trip, we divide our sample into 128 rectangles by origin coordinates and 128 rectangles by destination coordinates. The rectangles are sized so that they have similar numbers of trips. We also divide our sample into 15 hour-of-the-week intervals. This results in 245,760 buckets as the cartesian product of all origin, destination, and hour of the week divisions. Most of our regressions include fixed effects based on these groups, which we call *trip characteristics* fixed effects. Appendix A gives further details about these trip characteristic groups, and it shows that, despite the large number of groups, most trips in our sample are in groups with more than five trips.

Individual metrics We start with simple linear regressions to understand how ratings relate to individual metrics. We estimate regressions of the form

$$y_i = \beta m_i + \gamma X_i + \epsilon_i. \tag{1}$$

Trips are indexed by i. The left hand side variable y_i is a measure of trip rating, and m_i is a vector of driving metrics. X_i is a set of fixed effect dummies. Table 1 shows estimates of this equation for our sample of UberX trips. All driving metrics are normalized, so coefficients measure how ratings change if metrics change by one standard deviation.

Columns (1) and (2) show results for regressions in which the dependent variable is the five star rating, and in columns (3) and (4) the dependent variable is an indicator that takes the value of 1 when the rating is equal to 5. Riders dislike phone handling and hard brakes. Some specifications show a weak preference for cell phones being mounted. When we focus on regressions with driver fixed effects, we see that riders dislike hard accelerations. The four metrics we have mentioned reflect preferences for safer trips; these trips may also provide a more comfortable ride. Preferences for speed metrics, on the other hand, are more nuanced. Riders prefer higher low speeds and lower high speeds, i.e, trips in which drivers stay at an intermediate speed throughout the trip. This

[7] In some specifications we also include rider fixed effects (see Appendix D.1). Although our numbers change somewhat, the interpretation of our results does not change. Another potential concern is that the quality of the car might alter the perception that a trip is safe. Appendix D.5 shows that car quality does not interact in meaningful ways with our safety variables.

reflects a compromise between a safe, smooth ride and getting quickly to the destination. Riders have strong preferences for shorter trips, both in terms of distance and duration. They also prefer to be picked up and dropped off close to the requested locations.

The dependent variable in columns (5) and (6) is a dummy for whether the trip was rated. Coefficients tend to have opposite signs relative to previous columns (except for speed high and drop-off). This suggests that riders are more likely to rate trips when they are unsatisfied. This type of bias is one motivation for our approach described below; by creating a score for the quality of the ride that can be evaluated whether or not the ride was actually rated, we avoid the challenge of dealing with non-random missing ratings. Thus, when interpreting magnitudes, we caution that our predictions of ratings only apply to the type of ride that was actually rated in practice, rather than overall satisfaction if all rides were rated.

One potential concern with Table 1 is that traffic may be correlated with certain forms of driving, and riders blame drivers for bad traffic conditions. That would result in a correlation between driving metrics and ratings that is unrelated to rider preferences for those metrics. In Appendix D.2, we construct a measure of traffic conditions and show that our results are robust to controlling for traffic. A second concern is that the trip characteristic fixed effects we define may be too coarse spatially. Appendix D.3 shows that our results are robust to defining trip characteristic fixed effects in different ways.

Flexible functional form We now follow a more flexible approach to capture the dependence of rating as a function of driving metrics. We estimate

$$y_i = \mu_{d(i)} + \nu_{c(i)} + s(m_i; \theta) + \epsilon_i, \tag{2}$$

where $d(i)$ indexes the driver and $c(i)$ indexes the trip characteristics group. Thus, $\mu_{d(i)}$ and $\nu_{c(i)}$ represent driver and trip characteristics fixed effects.

The term $s(m_i; \theta)$ is a flexible function of driving metrics. It is the sum of three high order polynomials. The first one is the interaction of a quadratic function of handling and a quadratic function of mounting. The second one is the interaction of a quadratic of brakes, a quadratic of accelerations, a quartic of high speed, and a quartic of low speed. The third term is the interaction of a cubic of distance, a cubic of duration, a

Table 1: Rating response to driving metrics

	Dependent variable:					
	Rating		Rating is 5		Rated	
	(1)	(2)	(3)	(4)	(5)	(6)
Mounted	0.0096***	0.0015	0.0045***	0.0010	0.00002	−0.0011**
	(0.0009)	(0.0013)	(0.0005)	(0.0006)	(0.0002)	(0.0005)
Handling	−0.0018**	−0.0050***	−0.0005	−0.0016***	0.0006**	−0.0002
	(0.0008)	(0.0009)	(0.0004)	(0.0004)	(0.0002)	(0.0003)
Brakes	−0.0089***	−0.0032***	−0.0041***	−0.0014***	0.0009***	0.0014***
	(0.0007)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Accelerations	0.0014**	−0.0035***	0.0009***	−0.0015***	0.0015***	0.0013***
	(0.0007)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Speed low	0.0069***	0.0031***	0.0033***	0.0015***	−0.0026***	−0.0028***
	(0.0007)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Speed high	−0.0019***	−0.0065***	−0.0014***	−0.0036***	−0.0020***	−0.0026***
	(0.0007)	(0.0006)	(0.0004)	(0.0003)	(0.0002)	(0.0002)
Distance	−0.0156***	−0.0155***	−0.0046***	−0.0047***	0.0006***	0.0005**
	(0.0008)	(0.0007)	(0.0004)	(0.0003)	(0.0002)	(0.0002)
Duration	−0.0265***	−0.0232***	−0.0103***	−0.0089***	0.0066***	0.0059***
	(0.0008)	(0.0007)	(0.0004)	(0.0003)	(0.0002)	(0.0002)
Pickup	−0.0133***	−0.0115***	−0.0065***	−0.0058***	0.0082***	0.0080***
	(0.0007)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Dropoff	−0.0158***	−0.0139***	−0.0066***	−0.0057***	−0.0008***	−0.0008***
	(0.0008)	(0.0006)	(0.0004)	(0.0003)	(0.0002)	(0.0002)
Trip characteristics FE	✓	✓	✓	✓	✓	✓
Driver FE		✓		✓		✓
Observations	1,991,742	1,991,742	1,991,742	1,991,742	6,901,200	6,901,200

Note: This table shows results of regressions of rating variables—five-star rating, a dummy for the rating being five, and a dummy for the trips being rated—on driving metrics. All metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

quadratic of pick-up, and a quadratic of drop-off.[8] To avoid overfitting (this specification for $s(m_i;\theta)$ has 378 parameters), we regularize our model with a lasso penalty, where higher order terms have higher penalties. We do not penalize fixed effects. We choose penalties by cross validation (see Appendix B.2). Our final specification, which has lower out-of-sample MSE than simple lasso, is a post-lasso linear regression that only keeps those terms with a nonzero coefficient from the original lasso regression.

[8]We include higher order terms of speed variables because we expect them to have more important nonlinearities than the other four variables, which are defined as fractions. Fully interacting both terms would result in a regression with 291,600 terms, and it would not be feasible to estimate it given our sample size. We thus separate the function additively into three terms, one for phone usage and the second one for driving.

Figure 4 shows the functional form of our estimated $s(m_i;\hat{\theta})$. In each subfigure we vary two of the metrics. At each point in these plots we compute the average value of $s(m_i;\hat{\theta})$ over the distribution of the metrics we are not varying. The contours, as well as their color, represent the average value of $s(m_i;\hat{\theta})$. The blue scatterplot represents the distribution of the two metrics we are varying.

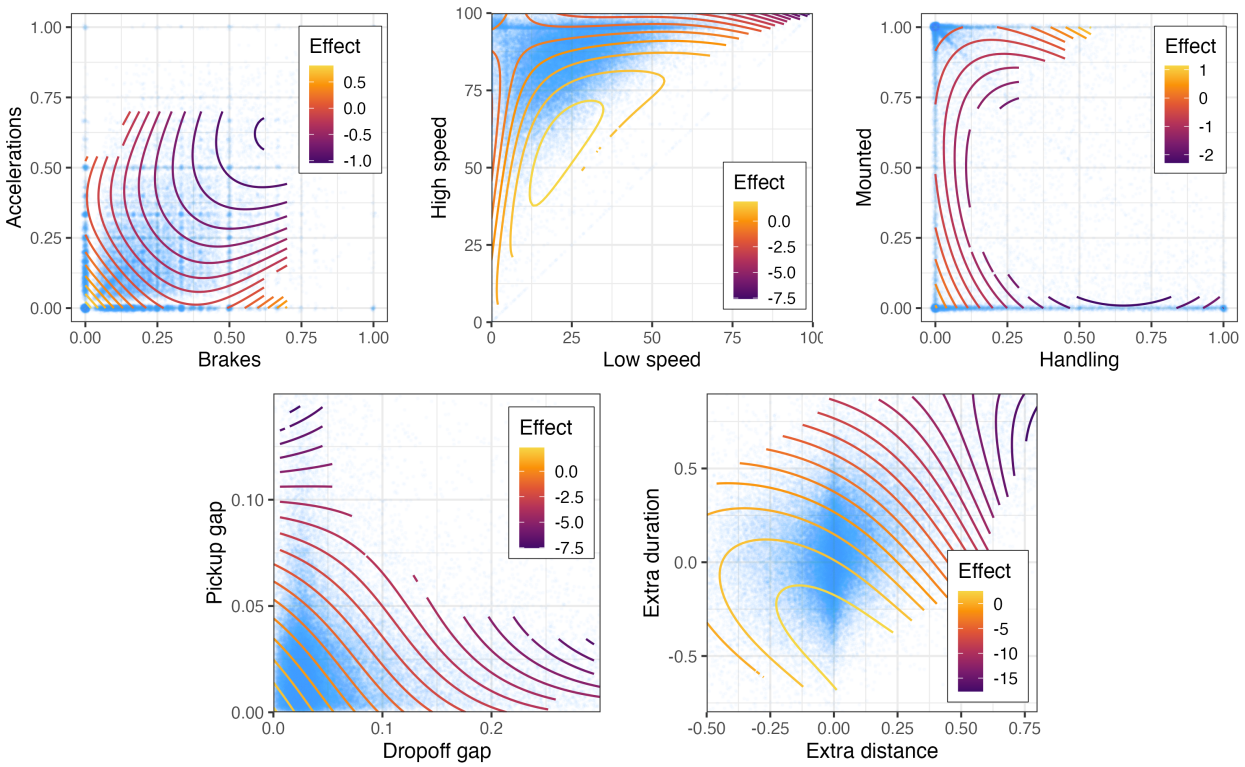


Figure 4: Effect of driving metrics on ratings

Note: Estimated effect of quality metrics on ratings. Each subfigure shows how the effect changes as two of the metrics change. Contour colors represent the magnitude of the effect. Effects are the average of the effect over the distribution of the rest of the metrics. The blue dots are a scatterplot to show the distribution of the two variables of interest. Contours are only shown in areas with a significant density.

The top left panel confirms that riders prefer fewer hard accelerations and hard brakes. The top middle panel shows preferences for speed. Riders have a bliss point near (20,55). This confirms that riders prefer trips that are neither too fast nor too slow.[9] In the top right panel, most of the mass is on the top left corner. Around that corner, riders prefer drivers to mount their phone and not to handle it. The bottom left panel shows that riders prefer to be picked up and dropped off close to the requested locations.

[9]The region with highest density is above and to the left of the bliss point. The gradient at that point is consistent with the coefficients for speed high and speed low in Table 1.

Finally, the bottom right panel shows that, for the most part, riders prefer shorter and quicker trips, although their preferences flatten out when trips become too short or too quick.

3.2 Driving scores

In the rest of this paper we compare the driving behavior across different settings. Although we can use individual driving metrics as outcomes, our previous analysis shows that riders' preferences over these metrics are nonlinear, and, in some cases, nonmonotonic. In order to summarize differences in quality across all metrics, we define a driving score $s_i^F = s(m_i; \hat{\theta})$ based on Equation (2) that we call our *full score* or *score F* (since it is computed from our full model). We measure the score in hundredths of a star.

Using our score as our main outcome variable instead of using ratings directly allows us to narrow down on a specific dimension of quality, driving behavior. This allows us to focus on objective behavior that we can observe through telemetry metrics, rather than on subjective preferences captured by ratings—which, as we show in Section 3.3, includes a large amount of rider-specific information that has nothing to do with driver behavior. It also has the advantage that it allows us to compare drivers' behavior when riders do not rate trips, which is the case for more than two thirds of the trips in our sample.

In some specifications we want to distinguish preferences related to different types of metrics. We thus define a vehicle *control score* or *score C*, denoted by s_i^C. We compute it from a model of the form of equation (2), but where $s(m_i; \theta)$ only includes terms related to handling, mounting, acceleration, brakes, and speed, which measure how the driver controls the vehicle. We also define a *routing score* or *score R*, denoted by s_i^R, where $s(m_i; \theta)$ only includes terms related to distance, duration, pick-up, and drop-off. Finally, we define a *speed score* or *score S*, denoted by s_i^S, where $s(m_i; \theta)$ only includes terms related to speed low and speed high. The speed score thus captures a subset of the information captured by the control score.

Figure 5 shows the density of our four scores. The full score has the largest variance, since it captures variation arising from all metrics. The control and routing scores each account for a substantial fraction of the variance. All four scores have a small standard deviation—4.39 hundredths of a star for the full score, 2.01 for the control score, 4.29 for

the routing score, and 1.35 for the speed score. Two factors contribute to this finding. First, the standard deviation of the rating is pretty low since most trips have a rating of 5. Second, as we will show in Section 3.3, of all the variation in ratings, only a small fraction is under the driver's control. The magnitude of our scores is on the order of magnitude of the score changes that drivers can achieve.

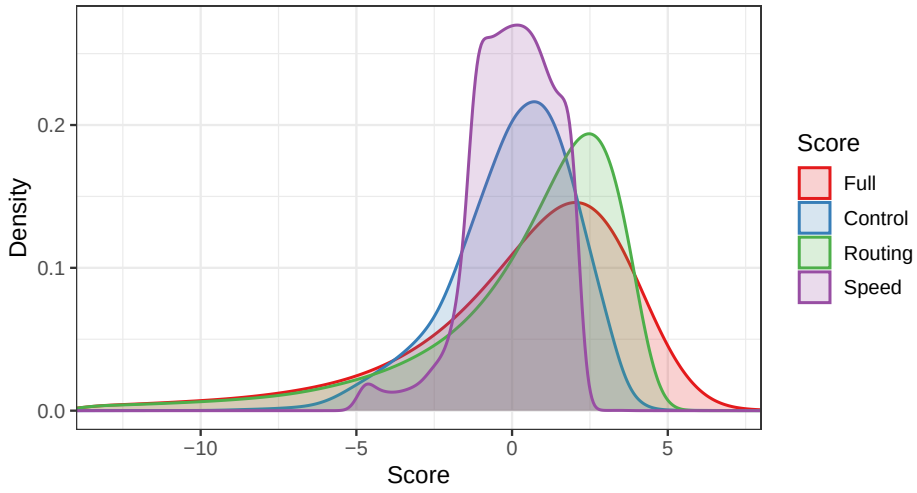


Figure 5: Distribution of scores

Note: Kernel density plots of the full score, the control score, the routing score and the speed score. The scores are measured in hundredths of a star, and they are centered so that they have mean zero.

We construct our scores based on the sample of UberX trips only. Thus, they capture the preferences of UberX riders, who may have chosen UberX precisely because their preferences tilt towards the driving behavior of UberX drivers. This may be a concern for our analysis from Section 5, where we compare the behavior of UberX and UberTaxi drivers. Appendix B.4 shows that those results look very similar when we construct scores based on the sample of all UberX and UberTaxi trips. We also run similar exercises using a score that was constructed from UberTaxi trips only—although we have less confidence in these scores because of the limited size of our sample of UberTaxi trips. See Appendix B.4 for a fuller analysis.

Scores as surrogates In using scores as an outcome variable, we follow Athey et al. (2019). They propose constructing a *surrogate index*, an outcome variable in settings where the true outcome of interest is missing—because it is a long term outcome, for instance, or because it is not systematically available. The surrogate index is defined as the predicted

value of the outcome conditional on a set of intermediate outcomes, the *surrogates*. It can be estimated in a dataset that differs from the one used to evaluate the impact of a treatment. The surrogates in our case are the driving metrics, the outcome that is sometimes missing is the trip rating, and the surrogate index are the scores. Thus, we use the score as a way to aggregate telemetry measures in a single dimension that relates to quality as experienced by riders.

We follow Athey et al. (2019) most closely in Section 4.3, where we use the scores to evaluate the impact of a randomized treatment that provides information to UberX drivers about past driving behavior. Athey et al. (2019) consider two assumptions: (i) the surrogates capture the effect of the treatment on the final outcome, and (ii) the relationship between the surrogates and the final outcome does not depend on the treatment. Given that we are focusing our attention on the impact of differences in telemetry metrics on rider satisfaction, it is natural to focus on differences in trip ratings that are captured by the driving metrics, so condition (i) is satisfied by assumption. Condition (ii) is plausible since our experiment affects the information available to drivers but does not directly impact riders.

Athey et al. (2019) show that two results hold under assumptions (i) and (ii). First, the estimated impact of the treatment on the score gives a consistent estimate of the average treatment effect on the final outcome. Second, it can be more efficient to analyze the impact of the treatment on the surrogate index rather than directly on the final outcome (even if that outcome were observed).[10]

The rest of our analysis in Section 4 uses scores to evaluate the impact of other treatments (such as warnings and notifications) that are not assigned in a random experiment. The above results hold under the additional assumption of unconfoundedness conditional on controls—the usual assumption that is required in quasi-experimental settings. Finally, in Section 5 we use the score to compare UberX to UberTaxi. Our goal is to quantify the changes in quality that would occur if riders were to request an UberX trip instead of an UberTaxi trip. We do not argue that we capture the well-being of UberTaxi riders, since they may have different preferences. Instead, our scores should be

[10]In an experiment where both surrogates and the final outcome are observed for each unit, Athey et al. (2019) show that efficiency can be gained by pooling data from the treatment group and the control group when estimating the relationship between the surrogates and the final outcome. Our analysis of a randomized experiment uses a smaller dataset; in that case, efficiency is gained by using a larger dataset to estimate the relationship between surrogates and the final outcome.

interpreted as differences in how an UberX rider would evaluate the driving behavior.

3.3 What determines driving behavior?

Our results from Section 3.1 give us a clear picture of what kind of behavior riders like, and Section 3.2 give us a way of measuring how well behavior complies with what riders like. However, it is still unclear whether that behavior is determined by the driver, by the rider, or by the type of trip. In this section we explore these questions.

Variance decomposition We start by decomposing the variance of ratings, driving metrics, and scores into different components. For outcome y_i, we run

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \gamma_{r(i)} + \epsilon_i. \tag{3}$$

where $r(i)$, $c(i)$, and $d(i)$ represent, respectively, the rider, trip characteristics, and driver of trip i. We thus decompose y_i into a driver effect, a trip characteristics effect, a rider effect, and a residual.[11]

Table 2 shows the variance of each one of the four terms in equation (3), where the outcomes y_i are driving metrics, scores, and trip rating. For control metrics and scores, the driver accounts for a significant share of the variation. This is especially true for mounted and handling. The rider is not responsible for much of the variation. We also see that trip characteristics are especially important for speed and for the pickup and dropoff metrics.[12]

Driver effects account for a small fraction of the variation in ratings. This explains, in part, why our scores have such a small standard deviation. Rider effects, on the other hand, account for a substantial fraction of the variation in ratings. This highlights the limitations of the rating itself as opposed to the scores based on telemetry data we use to evaluate ride quality; our scores are applied systematically to the telemetry measures from each ride, while the ratings have rider-specific noise that is unrelated to the driver's performance.

[11] One challenge is that $\mu_{d(i)}$ captures most of the variation for drivers with few trips, and $\gamma_{r(i)}$ captures most of the variation for riders with few trips. We thus limit our sample to drivers and riders that have more than 20 trips in our sample.

[12] This finding underscores the importance of checking that our results about the comparison between UberX and UberTaxi are present even when considering metrics other than speed, since results based on

Table 2: Variance decomposition

	Driver	Rider	Trip characteristics	Residual
Mounted	0.830	0.005	0.003	0.164
Handling	0.591	0.012	0.008	0.386
Brakes	0.220	0.024	0.038	0.725
Accelerations	0.333	0.021	0.026	0.628
Speed low	0.098	0.045	0.083	0.787
Speed high	0.137	0.047	0.116	0.651
Distance	0.018	0.058	0.040	0.847
Duration	0.039	0.059	0.042	0.831
Pickup	0.016	0.065	0.075	0.711
Dropoff	0.014	0.076	0.138	0.601
Score F	0.078	0.097	0.090	1.205
Score C	0.284	0.030	0.034	0.653
Score R	0.042	0.100	0.103	1.319
Score S	0.088	0.039	0.051	0.839
Rating	0.046	0.301	0.036	0.627

Note: The table shows the fraction of the variation of the variable in each row that can be explained by driver fixed effects, rider fixed effects, and trip characteristics fixed effects. The numbers are the result of regressions on three way fixed effects. We limit our sample to drivers and riders that have more than 20 trips in our sample.

4 Ratings, incentives, and nudges

This section presents evidence that the mechanisms Uber uses to encourage high quality positively impact driving behavior and trip quality. We first show that ratings and related notifications nudge drivers into better driving behavior, and that when Uber removes drivers with low ratings, it tends to remove those with low-quality driving. We then show that providing feedback to drivers about their driving behavior leads to improvements in driving quality.

4.1 Response to ratings and notifications

We now show evidence that drivers respond directly to ratings and to the notifications they receive when ratings become low.

Uber's rating system The main element of Uber's incentive system are rules under which UberX drivers with low ratings stop being matched to riders. These rules are coupled with notifications that are sent to drivers when their ratings reach certain thresholds.

speed may rely more heavily on carefully controlling for trip characteristics.

Table 3 summarizes the deactivation process, which moves sequentially through the steps in each row. In order to move to the next step, a driver has to satisfy the conditions on average ratings for the last 50 and 500 trips and the condition for the number of rated trips (they cannot enter the process before they complete 500 trips in total). At each one of the steps, the driver gets a notification by email, by text messaging, and through the Uber app. The notification explains that they are getting closer to deactivation and provides links to resources with help to improve ratings. Drivers can also progress backwards through the deactivation process if their ratings improve sufficiently. Drivers go back to the state right before the last notification if the app rating goes above certain thresholds.[13]

Table 3: Deactivation process for UberX

Event	Last 500 rating	Last 50 rating	Rated trips
Notification 1	< 4.6	< 4.6	25 since first trip
Notification 2	< 4.5	< 4.5	25 since notification 1
Temporary deactivation	< 4.4	< 4.4	25 since notification 2
Reactivation	*Passed quality improvement course*		
Notification 3	< 4.4	< 4.4	25 since reactivation
Permanent deactivation	< 4.4	< 4.4	25 since notification 3

An important fraction of drivers satisfy the conditions to be in the deactivation process. 64% of trips are completed by drivers who have completed more than 500 trips. 7.4% of trips are completed by drivers with a rating below 4.6, and 1.3% by drivers with a rating below 4.4. Even more importantly, a larger fraction of drivers are close to these thresholds and should therefore be worried about their behavior if they want to avoid entering the process or being deactivated. For 11.8% of drivers, receiving a 3-star rating for the next 15 trips would result in an app rating below 4.6; for 1.8% of drivers, this sequence would result an app rating below 4.4. Appendix E shows further statistics related to the deactivation process.

Response to rating and notifications We first explore how drivers' current app ratings affects the way they drive. Additionally, we want to see if the notification system influences their behavior. For outcome y_i, we estimate the model

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \alpha r_i^{\text{app}} + \beta w_i + P_2(n_i; \gamma) + \epsilon_i, \tag{4}$$

[13]The thresholds to go back are 4.6, 4.5, and 4.45 after notification 1, 2, and 3, respectively.

where r_i^{app} is the rating the driver observes in the app at the time of the trip, and w_i is a vector of dummies that characterizes the stage along the deactivation process in which the driver is in. We include fixed effects at the driver and trip characteristics level.

Since we are including driver fixed effects, we focus on variation within drivers, a large fraction of which might be driven by the experience of the driver. This might be problematic, especially when trying to measure α, the response to the app rating. For that reason, we control for $P_2(n_i;\gamma)$, a quadratic of the number of Uber trips the driver has completed, which is a proxy for experience. In order to identify β we exploit variation due to drivers that crossed some threshold and got a notification. Appendix E shows that there is a large number of such crossings, both from above and below.

Table 4 shows the results of this exercise. In Panel A, w_i is simply a dummy for being in any notification state. In other words, whenever a driver gets his first notification, w_i switches from zero to one and stays like that forever. We see that notifications have a positive effect on ratings and on all scores. Table 5, which presents the same exercise for individual metrics, shows that notifications also affect every driving metric in the direction that riders prefer according to Table 1, although not all coefficients are significant. One potential concern with these results is that they may be driven by reversion to the mean. Appendix D.7 shows that results are similar if we exclude the last 10 or 20 trips before the driver rating crosses the threshold to receive a notification, suggesting that this form of mean reversion is not the main driver of these findings.

The estimates of the app rating coefficient α suggest that drivers respond to low ratings by changing their behavior in a way that increases future ratings. This can be by improving how they drive, but it can also be through some channels we are not able to measure. For instance, drivers can be more friendly with their riders, or they can start cleaning their car. Columns (2)-(5) suggest that this is mostly unrelated to driving behavior, since we find small, positive coefficients that are marginally significant. A potential concern with giving a causal interpretation to this coefficient is that there may be factors that lead to serial correlation in driver behavior. In Appendix D.6, we consider a related exercise in which we measure how different outcomes respond to shocks in the rating from the last trip that the driver completed, using an instrumental variables strategy to address these potential endogeneity concerns. Our findings are consistent with these results.

Table 4: Response to ratings and notifications

	Dependent variable:				
	Rating	Score F	Score C	Score R	Score S
	(1)	(2)	(3)	(4)	(5)
	Panel A: General effect of warnings				
Has received notif.	0.108***	0.101***	0.051***	0.078***	0.020***
	(0.007)	(0.020)	(0.017)	(0.017)	(0.007)
App rating	−0.296***	0.067*	0.043*	0.065**	0.025**
	(0.013)	(0.036)	(0.025)	(0.033)	(0.012)
Observations	1,983,507	6,874,960	6,874,960	6,874,960	6,874,960
	Panel B: Decomposition of effect of warnings				
1st notification	0.100***	0.102***	0.049***	0.071***	0.017**
	(0.007)	(0.021)	(0.017)	(0.018)	(0.007)
2nd notification	0.125***	0.062*	0.033	0.057*	0.003
	(0.013)	(0.035)	(0.028)	(0.031)	(0.012)
3rd notification	0.239***	0.171*	0.232***	0.028	0.081**
	(0.036)	(0.094)	(0.081)	(0.081)	(0.032)
1st notif. expired	0.104***	0.104***	0.039**	0.108***	0.019**
	(0.008)	(0.024)	(0.020)	(0.021)	(0.009)
2nd notif. expired	0.089***	0.021	−0.028	0.063	0.034
	(0.024)	(0.074)	(0.051)	(0.064)	(0.023)
3rd notif. expired	0.277***	0.291*	0.405**	0.043	0.196***
	(0.053)	(0.156)	(0.158)	(0.138)	(0.051)
App rating	−0.298***	0.063*	0.040	0.060*	0.022*
	(0.013)	(0.036)	(0.025)	(0.033)	(0.012)
Observations	1,983,507	6,874,960	6,874,960	6,874,960	6,874,960

Note: This table shows regressions of ratings and scores on app ratings and dummies for having received notifications. Panel A uses a single dummy for having received notifications. Panel B uses different dummies for different stages in the notification process. All regressions include driver and trip characteristics fixed effects, and control for a quadratic function of the number of Uber trips the driver has completed. Standard errors are clustered by driver. All driving metrics are normalized to mean zero and variance one. A few observations are missing due to drivers that have never been rated. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Panel B of Tables 4 and 5 separates the effect of notifications across different notification states. The state is defined by the last event that took place. All effects are measured relative to the level before getting any notification. As we can see, effects tend to have the same signs as the main effect, although not all coefficients are significant. Notification 2 does not seem to have any effect on top of the original effect of notification 1. Notification 3, on the other hand, seems to have the strongest effect, consistent with the imminent threat of deactivation.

The main takeaway from these results is that, once drivers enter the deactivation pro-

Table 5: Response to ratings and notifications

	Dependent variable:									
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Distance	Duration	Pickup	Dropoff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: General effect of warnings									
Has received notif.	0.051***	−0.042***	−0.003	−0.016**	0.012**	−0.002	−0.013***	−0.014***	−0.015***	−0.002
	(0.012)	(0.012)	(0.007)	(0.007)	(0.005)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)
App rating	0.041**	−0.025	0.010	0.002	0.043***	0.017*	0.010	−0.023***	−0.007	−0.011*
	(0.017)	(0.019)	(0.011)	(0.012)	(0.009)	(0.009)	(0.007)	(0.007)	(0.007)	(0.006)
Observations	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960
	Panel B: Decomposition of effect of warnings									
1st notification	0.049***	−0.039***	−0.006	−0.017**	0.007	−0.012**	−0.012***	−0.011***	−0.010**	−0.004
	(0.012)	(0.012)	(0.007)	(0.008)	(0.006)	(0.006)	(0.004)	(0.004)	(0.005)	(0.004)
2nd notification	0.024	−0.039**	−0.012	−0.027**	0.005	0.016*	−0.013*	−0.012*	−0.022***	0.009
	(0.020)	(0.019)	(0.012)	(0.012)	(0.009)	(0.010)	(0.007)	(0.007)	(0.008)	(0.006)
3rd notification	0.078*	−0.117*	−0.064**	−0.029	0.020	−0.042	−0.025	−0.005	−0.036**	0.006
	(0.047)	(0.065)	(0.029)	(0.029)	(0.024)	(0.026)	(0.018)	(0.020)	(0.018)	(0.018)
1st notif. expired	0.065***	−0.042***	0.016*	−0.004	0.022***	0.017**	−0.013***	−0.020***	−0.020***	−0.003
	(0.015)	(0.014)	(0.009)	(0.010)	(0.007)	(0.007)	(0.005)	(0.005)	(0.005)	(0.004)
2nd notif. expired	0.009	−0.010	0.016	0.012	0.030*	0.033*	−0.019	−0.015	−0.013	−0.003
	(0.032)	(0.031)	(0.020)	(0.019)	(0.016)	(0.020)	(0.013)	(0.015)	(0.014)	(0.012)
3rd notif. expired	0.109	−0.176	−0.098**	−0.132**	0.078**	−0.050	−0.033	−0.009	−0.048*	0.005
	(0.077)	(0.123)	(0.042)	(0.059)	(0.034)	(0.032)	(0.022)	(0.025)	(0.027)	(0.032)
App rating	0.037**	−0.023	0.007	0.001	0.040***	0.015	0.010	−0.022***	−0.005	−0.011*
	(0.017)	(0.019)	(0.011)	(0.011)	(0.009)	(0.009)	(0.007)	(0.007)	(0.007)	(0.006)
Observations	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960	6,874,960

Note: This table shows regressions of metrics on app ratings and dummies for having received notifications. Panel A uses a single dummy for having received notifications. Panel B uses different dummies for different stages in the notification process. All regressions include driver and trip characteristics fixed effects, and control for a quadratic function of the number of Uber trips the driver has completed. Standard errors are clustered by driver. All driving metrics are normalized to mean zero and variance one. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

cess and start receiving notifications, they make substantial improvements to their behavior. This is evidence that the notifications system is working as intended. Importantly, these effects are persistent, since they remain after notifications expire—after drivers start behaving better and receive enough high ratings to get out of the deactivation process.

4.2 Driver deactivation

We now show evidence that the deactivation process described in the previous section is able to weed out drivers that perform badly. Figure 6 compares the distribution of ratings and scores completed by drivers who were deactivated with the distribution for the general population. For deactivated drivers, we want to avoid looking at their behavior after responding to incentives, so we only consider trips that took place before they got

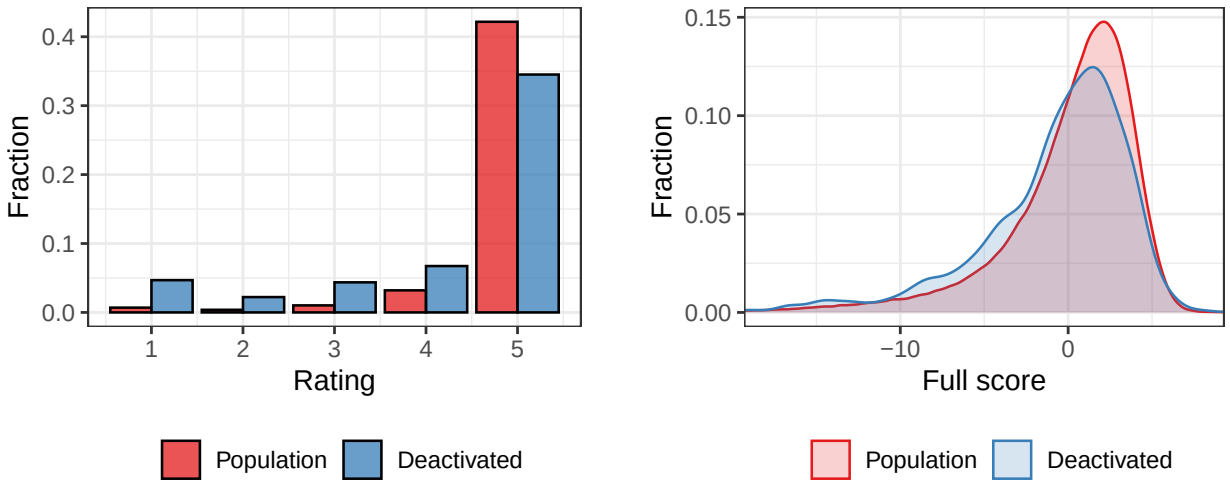


Figure 6: Distribution of ratings and scores for drivers who were deactivated

Note: These figures show the distribution of the rating and of the full score for two different samples. The whole population of trips is shown in red. The trips by drivers who were eventually deactivated *that took place before getting the first notification* are shown in blue.

their first warning.

Deactivated drivers have a much larger share of trips with low ratings. This relation is mechanical—drivers are deactivated when they have low ratings. But we see that low scores are also more common among drivers that are eventually deactivated. In particular, we see that the distribution of scores has a longer left tail: the deactivation process disproportionately kicks out drivers that complete trips with very bad scores.

We now follow a more nuanced regressions analysis. We start, again, with the sample of trips of drivers who were eventually deactivated, but which took place before the driver got the first notification. To avoid confounders related to the time when trips took place or the places where trips took place, we match every such trip to the nearest ten trips by drivers who were not deactivated. We use a Euclidean metric in which a half-hour difference is equivalent to a difference in origin or destination of one kilometer. We then estimate the effect of the trip being done by a driver who was eventually deactivated as

$$\hat{\tau} = \frac{1}{N} \sum_{i \in I_{\mathrm{D}}} \left(y_i - \frac{1}{10} \sum_{j \in I_{ND,i}} y_j \right), \tag{5}$$

where y_i is the outcome variable for trip i, I_{D} is the set of trips that were completed by a driver who was eventually deactivated, and $I_{ND,i}$ denotes the set of i's nearest neighbors

Table 6: Characteristics of trips by drivers who were later deactivated

	Dependent variable:						
	Score F	Score D	Score R	Score S	Mounted	Handling	Brakes
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Deactivated	-0.914^{***}	-0.466^{***}	-0.925^{***}	-0.123^{**}	-0.412^{***}	0.441^{***}	0.037
	(0.137)	(0.164)	(0.114)	(0.062)	(0.143)	(0.162)	(0.063)
Observations	3,024	3,024	3,024	3,024	3,024	3,024	3,024
	Dependent variable:						
	Accels.	Speed low	Speed high	Distance	Duration	Dropoff gap	Pickup gap
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Deactivated	-0.040	-0.284^{***}	-0.212^{***}	0.081^{***}	0.156^{***}	0.047^{**}	0.210^{***}
	(0.087)	(0.040)	(0.060)	(0.022)	(0.037)	(0.019)	(0.026)
Observations	3,024	3,024	3,024	3,024	3,024	3,024	3,024

Note: This table shows regressions of driving metrics as well as scores on dummies for whether the driver was eventually deactivated. The sample includes all trips by riders who were not deactivated, and trips by drivers who were deactivated *before they got their first notification*. All metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

among trips by drivers who were never deactivated.[14] We compute standard errors as in Abadie and Imbens (2005) with an adjustment for clustering by driver.

Table 6 presents the results from this exercise. Drivers who are deactivated perform badly in terms of all four scores. They also tend to display worse behavior in terms of individual metrics. They are less likely to mount their phone and more likely to handle it. There is no significant difference in terms of brakes and accelerations, but they are slow drivers, their trips are longer in terms of distance and duration, and they pick up and drop off riders farther away from the desired locations. All of these results are consistent with the hypothesis that the deactivation process is able to eliminate drivers who had bad driving behavior, offsetting to some extent the fact that they could have been able to enter the platform because of simplified screening.

[14]This estimator is an average treatment effect on the treated (ATT) since it weights observations according to the distribution of characteristics of trips by drivers that were eventually deactivated.

4.3 Information on past behavior

We now quantify the extent to which drivers' behavior responds to feedback provided by Uber on their past behavior.

Uber informed drivers of their past performance in terms of driving metrics.[15] Every week, drivers received a simple report on their smartphone screen summarizing their driving metrics in the past week and comparing them with other drivers'. Receiving this information could motivate drivers to improve their behavior, but it could also lead to worse behavior if well-behaving drivers decrease their efforts.

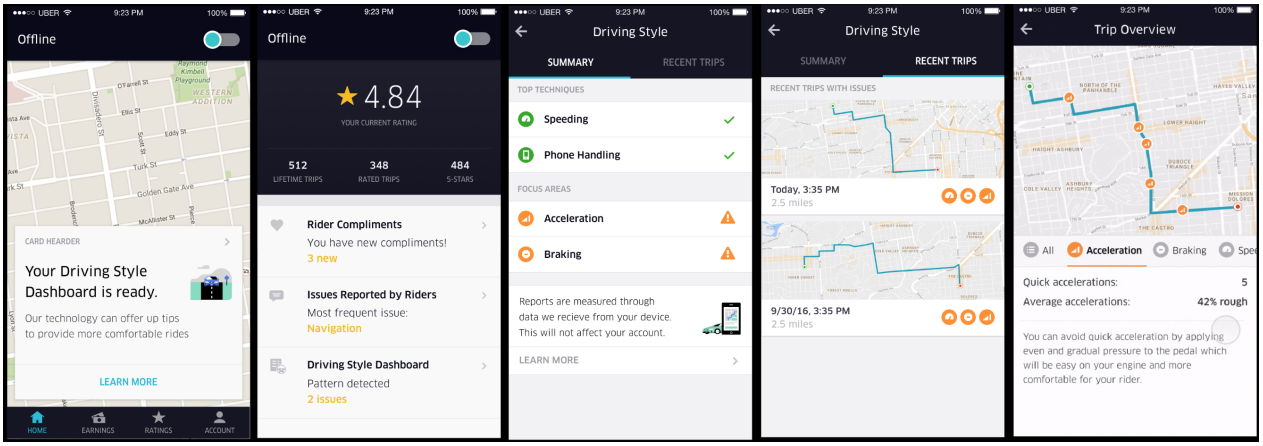


Figure 7: Images of the dashboard with detailed information about past driving behavior

Experimental setup Every driver in our sample received the simple report, so there is no direct way to tell to what extent it affected their behavior. However, Uber ran a closely related experiment. Uber was considering a major upgrade of the simple report. The new version would be a complete dashboard in the Uber app with detailed information about past behavior. The dashboard included information on individual trips and on specific segments within each trip. Figure 7 shows screenshots of the dashboard.

Uber initially treated a random set of drivers with the upgraded dashboard, and kept on sending the original report to a control group. We use this experimental setting to measure how additional information on past behavior influenced trip quality. In Appendix F.1 we report results of a balance test of pre-experiment averages by driver for all metrics and scores. We include only drivers who took at least ten trips in the

[15]This feature of the product was discontinued in late 2018

month before the experiment. We do not find statistically significant differences for most metrics and scores, but we find significant differences for the speed metrics and for the distance metric. Since we do not have perfect balance in the pre-treatment period for these variables, we control for pre-treatment averages in analyzing the experiment.

We start by analyzing the results of the experiments with regressions of the form

$$y_i = \alpha + \tau T_{d(i)} + \gamma X_i + \epsilon_i, \tag{6}$$

where y_i is some outcome and $T_{d(i)}$ is a dummy for whether the driver was in the treatment group. Our estimate for τ is thus an estimate of the intent to treat (ITT) effect. We limit our sample to those trips that took place after the experiment started. X_i includes the driver's pre-experiment mean for the outcome and trip characteristics fixed effects.

We are also interested in measuring the average treatment effect (ATE) of interacting with the new dashboard. We construct an indicator variable I_i that is equal to one if driver $d(i)$ interacted with the dashboard in the week prior to trip i. We are then interested in a model of the form

$$y_i = \alpha + \theta I_i + \gamma X_i + \epsilon_i. \tag{7}$$

We estimate this regression by 2SLS, instrumenting I_i with the treatment dummy and its interaction with the driver's pre-treatment mean for each of the telemetry measures and scores. Treatment status and the pre-treatment outcomes are demeaned for interpretability.[16] Our estimate of θ is thus an estimate of the ATE.

Columns (1)-(4) in Table 7 show results for regressions of the form of equations (6) and (7) where the dependent variables are our four scores. In Panel A we measure a positive effect of the ITT, although it is only significant for the full and routing scores. Panel B shows that interacting with the dashboard leads to a significant improvement in all four scores. The effect is especially pronounced for the full and control scores, where we would have expected the strongest impact. Interestingly, we also find an effect on the routing score which may be somewhat surprising given that the feedback provided is unrelated to routing. This might occur because feedback about their behavior may increase overall awareness of their behavior, even in dimensions not directly addressed by the feedback. We see the smallest effect on the speed score, possibly because the feedback (intended to encourage slower driving) is misaligned with rider preferences.

[16] We obtain similar results if we use a less rich set of instrumental variables.

Appendix F.2 presents a similar table for individual metrics. We do not measure significant effects.[17]

We are also interested in seeing how the effects of the dashboard differ by drivers' pre-treatment behavior. To see how the experiment affected poorly-performing drivers, we first compute the average of each outcome for each driver in the pre-treatment period (excluding drivers with fewer than ten trips in the month before the experiment launched). We then code a pre-treatment dummy variable, "Bottom 10th Perc. Before," which indicates whether the driver was among the worst-performing 10% of drivers in terms of the outcome variable during the pre-period.[18]

Columns (5)-(8) in Table 7 present results from regressions similar to Equations 6 and 7, where access and interaction with the dashboard are interacted with the pre-treatment dummy. For the full and control scores, the effects of the dashboard are primarily driven by drivers who performed worst during the pre-treatment period. These findings align with those of Choudhary et al. (2022), who observed similar results for general private drivers in India. This suggests that informing poorly-performing drivers about their performance encourages increased efforts and leads to performance improvement. Additionally, there appears to be a small improvement for drivers who were not performing poorly, indicating that more detailed information does not deteriorate the performance of previously well-performing drivers. The patterns for the routing and speed scores are less clear and noiser. This is consistent with our claim that these scores are less likely to be influenced by feedback, suggesting that one should not expect a larger improvement from the lowest-performing drivers.

4.4 Response to individual rider effects

Our previous results in this section show how drivers improve their driving behavior in response to incentives and nudges from the platform. We now analyze whether individual riders have an impact on driving behavior. One takeaway from Table 2 is that rider effects account for a substantial fraction of the variation, especially for speed and routing

[17] We pulled data from other cities to try to increase the power of our regressions, but we found inconsistent results on metrics. For scores, we found a stronger effect in San Francisco, somewhat weaker results in LA, DC, and Boston, and no evidence of an effect in New York.

[18] We obtain similar results if we create similar dummies with different cutoffs, or if we use a continuous measure instead of a dummy variable.

Table 7: Effect of the driving dashboard

	Dependent variable:							
	Score F	Score C	Score R	Score S	Score F	Score C	Score R	Score S
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Intent to treat estimator							
Pre-Period Mean	0.593***	0.820***	0.569***	0.767***				
	(0.007)	(0.006)	(0.006)	(0.011)				
Treatment	0.037***	0.018*	0.028***	0.004				
	(0.011)	(0.009)	(0.008)	(0.004)				
Bottom 10th Perc. Before					−1.777***	−2.591***	−1.179***	−0.752***
					(0.042)	(0.043)	(0.039)	(0.033)
Treatment x Not Bottom 10th Perc.					0.040***	0.010	0.031***	0.007
					(0.015)	(0.014)	(0.011)	(0.005)
Treatment x Bottom 10th Perc.					0.093	0.172***	−0.008	0.044
					(0.058)	(0.058)	(0.050)	(0.040)
Observations	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965
	Panel B: 2SLS estimator							
Interaction	0.062***	0.051***	0.032**	0.016**				
	(0.019)	(0.017)	(0.015)	(0.008)				
Bottom 10th Perc. Before					0.050	−0.066	−0.020	−0.016
					(0.041)	(0.048)	(0.035)	(0.019)
App Interaction x Not Bottom 10th Perc.					0.056***	0.026	0.036**	0.015**
					(0.019)	(0.017)	(0.015)	(0.007)
App Interaction x Bottom 10th Perc.					0.159	0.313***	−0.027	0.030
					(0.100)	(0.096)	(0.082)	(0.040)
Observations	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965

Note: Panel A shows regressions of scores on a dummy for being in the treatment group of the dashboard experiment and on the pre-period mean score. Panel B shows 2SLS regressions of scores on a dummy for observing the dashboard, using the treatment dummy and interactions with pre-treatment metrics and scores as instruments. Columns (4)-(6) measures heterogeneous effects, depending on whether the driver was among the 10% of worst performing drivers in terms of the pre-treatment mean of the outcome variable. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

metrics. Here we explore if drivers tailor their behavior to riders' heterogeneous preferences. For instance, it could be the case that some individual riders tend to be in a hurry and put pressure on their driver to get to their destination quickly. Such findings would suggest that Uber's rating systems encourage drivers to follow the form of driving riders want.

For some outcome y_i, let $\bar{y}_i^{LO}$ be the average value of the outcome for all trips taken

by $r(i)$, the rider who took trip i, leaving out the current trip. We run

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \beta \bar{y}_i^{LO} + \epsilon_i, \tag{8}$$

where the first two terms denote driver and trip characteristics fixed effects. Our main object of interest is β, which measures to what extent y_i is correlated within riders. The driver and trip fixed effects account for the fact that riders will tend to have similar trips (e.g. trips from their home), and that drivers may focus their driving around certain areas. Many riders only have a small number of trips in our database, so $\bar{y}_i^{LO}$ could be a noisy measure. Thus, we restrict our sample to riders with 20 or more trips.[19]

Panel A in Table 8 shows results for this exercise, where the outcome variables are metrics. We see a negative but negligible effect on mounted and handling. On the other hand, we see a strong and positive effect on brakes and accelerations, and especially on speed and routing metrics. This means that there is a strong correlation of metrics within riders. We interpret this finding as establishing that riders influence driver behavior. Appendix C.1 presents similar results for scores: the strong correlations in metrics are reflected as correlations in scores.

We also analyze whether the effect individual riders have on driving intensifies during rush hour, when riders are presumably in a hurry.[20] Appendix D.4 confirms that riders who take trips during the morning rush hour tend to prefer faster trips, based on regressions similar to those in Table 1 but that also allow for heterogeneous preferences. One should then expect stronger within-rider correlations during the morning rush hour if riders who use UberX at that time do so consistently.[21] We estimate regressions of the form of equation (8), but where we also interact $\bar{y}_i^{LO}$ with dummies for different times of the week: morning rush hour, afternoon rush hour, and off-peak times. Panel B in Tables 8 and 15 show the result of this exercise. We see especially large coefficients for trips during the morning rush hour, as expected.

[19]Restricting to riders with 20 or more trips reduces the sample size in the regression. As a result, a higher share of trip characteristics groups have few or no trips. Accordingly we use coarser fixed effects than in previous regressions. We divide the sample into 64 groups by origin, 64 groups by destination, and 15 groups by hour of the week. See Appendix A for more details.

[20]We define morning rush hour trips as those starting during weekdays between 6 am and 10 am, and afternoon rush hour trips as those starting between 5 pm and 9 pm on weekdays.

[21]We note that that this coefficient might also pick up unobserved trip characteristics; one set of heavy UberX riders might be riders who regularly use UberX during the morning commute, and the route may be identical, inducing correlation among scores within a rider.

Table 8: Response to rider preferences

	Dependent variable:									
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Distance	Duration	Pickup	Dropoff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Main effect									
Mean by rider	-0.006^{***}	-0.004	0.140^{***}	0.113^{***}	0.338^{***}	0.352^{***}	0.518^{***}	0.516^{***}	0.545^{***}	0.601^{***}
	(0.001)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
Observations	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440
	Panel B: Heterogeneous effects by time of the week									
Mean by rider	-0.002	0.001	0.158^{***}	0.125^{***}	0.302^{***}	0.326^{***}	0.432^{***}	0.457^{***}	0.502^{***}	0.525^{***}
× Off-peak	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
Mean by rider	-0.025^{***}	-0.028^{***}	0.110^{***}	0.101^{***}	0.512^{***}	0.507^{***}	0.861^{***}	0.752^{***}	0.704^{***}	0.974^{***}
× AM rush	(0.004)	(0.006)	(0.007)	(0.007)	(0.006)	(0.008)	(0.007)	(0.006)	(0.005)	(0.005)
Mean by rider	-0.006^{*}	-0.002	0.104^{***}	0.083^{***}	0.310^{***}	0.311^{***}	0.428^{***}	0.463^{***}	0.531^{***}	0.403^{***}
× PM rush	(0.003)	(0.005)	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.004)
Observations	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440	3,231,440

Note: This table shows regressions of quality metrics on leave-out means by rider. Only riders with more than 20 trips are included. All regressions control for driver and trip characteristics fixed effects. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

5 UberX versus UberTaxi

Having studied various dimensions in which UberX drivers respond to Uber's nudges and incentives, we now compare the driving behavior of UberX and UberTaxi trips. The differences we measure cannot be attributed solely to the different quality control mechanisms—ratings, information, and nudges for UberX and ex-ante screening for UberTaxi—since many other factors vary between the two products. The demographics of both pools of drivers differ, for instance, and the drivers face different monetary incentives. However, our estimates measure the causal effect of requesting an UberX instead of an UberTaxi for the same trip. Beyond its direct relevance for a rider deciding which product to choose, this causal effect tells us whether a ride-hailing platform that relies almost entirely on ex-post quality control can achieve the same level of quality of one that relies on ex-ante screening.

The main question we address is: Given the characteristics (origin, destination, and time) of an UberTaxi trip, how would the driving behavior experienced by the rider have changed if she had instead requested an UberX trip? In other words, we aim to estimate the average treatment effect of getting an UberX trip, for the distribution of UberTaxi trips. This estimate is, hence, the average treatment effect on the untreated (ATU). We do not estimate the more conventional average treatment effect on the treated (ATT) because,

while we can find similar UberX trips for most UberTaxi trips, the reverse is not true due to the much smaller number of UberTaxi trips in our data. The difference between an ATU and and ATT is not particularly relevant in this setting because what is a treatment and what is a control is arbitrary: we could have just as easily defined the treatment to be requesting an UberTaxi trip, in which case the estimate we obtain is the negative of the ATT. The key assumption underlying our estimation is unconfoundedness when controlling for trip characteristics: conditional on the trip's origin, destination, and time of day, there are no further unobserved characteristics of trips that would impact the telemetry metrics.

We use two different methodologies with very similar results. First, similar to the matching estimator we describe in equation (5), we match UberX and UberTaxi trips according to their origin and destination coordinates and the hour of the week at which they started. Concretely, we match every UberTaxi trip to its 10 nearest UberX neighbors, using a Euclidean metric in which a half-hour difference is equivalent to a difference in origin or destination of one kilometer. We then estimate the UberX effect as

$$\hat{\tau}^{match} = \frac{1}{N} \sum_{i \in I_{\text{taxi}}} \left(y_i - \frac{1}{10} \sum_{j \in C_i} y_j \right), \tag{9}$$

where I_{taxi} is the set of UberTaxi trips, and C_i denotes the set of i's nearest neighbors among UberX trips. We compute standard errors as in Abadie and Imbens (2005) with an adjustment for clustering by driver.

The second methodology is a simple fixed effects regression using the same trip characteristics as in Section 3. The specification we run is

$$y_i = \tau x_i + \beta X_i + \epsilon_i, \tag{10}$$

where x_i is a dummy that equals one if driver $d(i)$ is an UberX driver, and βX_i is a set of fixed effects. Our coefficient of interest is τ. In general, this estimator is not consistent for the ATU as it is usually defined, but it converges to a weighted average of treatment effects. As we see below, both methodologies result in almost identical results.

Table 9 shows the estimates from these specifications for metrics as the dependent variable. The matching estimator (Panel A) has almost identical results to the OLS estimator with trip characteristics fixed effects (Panel B). Our estimates change somewhat if

Table 9: Comparison of driving behavior between UberX and UberTaxi trips

	Dependent variable:									
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Distance	Duration	Pickup	Dropoff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Matching estimator									
UberX	0.9539***	0.0743***	−0.0480***	−0.2524***	−0.0171**	−0.2170***	0.2036***	0.1672***	−0.0825***	−0.2190***
	(0.0247)	(0.0129)	(0.0119)	(0.0144)	(0.0085)	(0.0071)	(0.0032)	(0.0058)	(0.0039)	(0.0066)
Observations	139,716	139,716	139,716	139,716	139,716	139,716	139,716	139,716	139,716	139,716
	Panel B: Trip characteristics fixed effects									
UberX	0.9518***	0.0757***	−0.0359***	−0.2449***	−0.0304***	−0.2194***	0.1974***	0.1529***	−0.0806***	−0.2226***
	(0.0250)	(0.0138)	(0.0121)	(0.0147)	(0.0087)	(0.0073)	(0.0035)	(0.0061)	(0.0040)	(0.0066)
Observations	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729
	Panel C: Trip characteristics and rider fixed effects									
UberX	0.9467***	0.0634***	−0.0573***	−0.2559***	0.0024	−0.1993***	0.1517***	0.1182***	−0.0782***	−0.2017***
	(0.0262)	(0.0148)	(0.0134)	(0.0159)	(0.0101)	(0.0089)	(0.0058)	(0.0083)	(0.0064)	(0.0081)
Observations	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729

Note: This table compares the driving behavior of UberX and UberTaxi trips according to driving metrics. Panel A uses a matching estimator, where every UberTaxi trip is compared to the ten nearest UberX trips based on origin and destination coordinates and time of the week. Panel B presents results from a linear regression that includes trip characteristics fixed effects. Panel C presents results from a linear regression that includes trip characteristics as well as rider fixed effects. All driving metrics are normalized to mean zero and variance one. Standard errors are adjusted as in Abadie and Imbens (2005) and clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

we also include rider fixed effects (Panel C), but not enough for the interpretation of the coefficients to change.

UberX drivers are more likely to mount their cell phones. The difference is almost one standard deviation. This is not surprising: Uber has led campaigns to ensure drivers mount their phones, sometimes giving away phone mounts. UberX drivers are, however, also more likely to handle their phones. This is also not too surprising, since they rely more on their cell phone to find the next trip, and they are probably more tech-savvy and thus more likely to navigate with their phones.

UberX drivers have fewer hard brakes and accelerations. The difference is especially pronounced for accelerations, with a difference of roughly one quarter of a standard deviation. They also drive more slowly in terms of both speed measures. However, the UberX effect is much stronger for *speed high* than for *speed low*, which means that UberX drivers tend to drive at a steadier speed than UberTaxi drivers, as one would expect if they pay more attention to riders' preference for steady speeds.

UberX drivers perform worse in terms of both the distance and duration of trips.[22]

[22]For these results, we use the distance and duration metrics based on imputed expected duration and distance, as explained in Appendix B.3.

Table 10: Comparison of driving behavior between UberX and UberTaxi trips

	Dependent variable:			
	Score F	Score C	Score R	Score S
	(1)	(2)	(3)	(4)
	Panel A: Matching estimator			
UberX	−0.0257	0.1188***	−0.2767***	0.2974***
	(0.0250)	(0.0257)	(0.0190)	(0.0092)
Observations	139,716	139,716	139,716	139,716
	Panel B: Trip characteristics fixed effects			
UberX	0.0026	0.1013***	−0.2466***	0.2874***
	(0.0260)	(0.0265)	(0.0201)	(0.0095)
Observations	7,147,729	7,147,729	7,147,729	7,147,729
	Panel C: Trip characteristics and rider fixed effects			
UberX	0.1059***	0.1431***	−0.1454***	0.2935***
	(0.0346)	(0.0289)	(0.0294)	(0.0121)
Observations	7,147,729	7,147,729	7,147,729	7,147,729

Note: This table compares the driving behavior of UberX and UberTaxi trips according to driving scores. Panel A uses a matching estimator, where every UberTaxi trip is compared to the ten nearest UberX trips based on origin and destination coordinates and time of the week. Panel B presents results from a linear regression that includes trip characteristics fixed effects. Panel C presents results from a linear regression that includes trip characteristics as well as rider fixed effects. All driving metrics are normalized to mean zero and variance one. Standard errors are adjusted as in Abadie and Imbens (2005) and clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

This finding aligns with Liu et al. (2021), who find that UberX drivers in New York generally take longer routes than taxi drivers. Appendix C.3 shows that UberX drivers have more trips that are longer than Uber's expected route, which may indicate moral hazard—driving longer to increase earnings. However, Appendix C.3 also shows that UberX drivers have fewer trips that are shorter than Uber's expected route, suggesting that they are less likely to deviate from the suggested route to take shortcuts, possibly due to less experience with the city's streets. Finally, UberX drivers tend to pick up and drop off riders closer to their desired locations, which could also contribute to the longer trips that we observe.

Table 10 reports estimates of the ATE on driving scores. The results reveal a mixed picture. UberX drivers perform better in terms of the control score, particularly in dimensions related to speed. However, they perform worse in terms of the routing score. The full score, which aggregates all dimensions, suggest that UberX and UberTaxi drivers

perform roughly equally, with perhaps a slight advantage for UberX drivers.

The main takeaway from this analysis is that there is no clear answer as to whether UberX or UberTaxi drivers provide a better quality in terms of the metrics we analyze. On one hand, UberX drivers perform better in terms of braking, accelerating, and driving at appropriate speeds. On the other hand, they perform worse in terms of handling their phones and routing. Overall, our findings suggest that passengers are roughly indifferent between the driving behavior of both types of drivers.

In the Appendix, we consider a variety of extensions and robustness checks; for example, in Appendix D.5, we make use of estimates of the market value of the cars, and show that our results are robust to comparing UberX and UberTaxi rides with similar market prices. One caveat is that, as we show in Appendix B.4, some of our results suggest that UberTaxi riders may prefer the driving behavior of UberTaxi drivers. Those results, however, only account for the preferences of a sample of riders who constitute 2% of the trips in our data. An overwhelming majority of Uber riders in fact prefer the driving behavior of UberX drivers.

UberX vs. UberTaxi during rush hour We now perform an exercise that gives suggestive evidence for whether UberX or UberTaxi drivers are more responsive to riders' preferences. As one might expect, we show in Appendix D.4 that riders prefer faster trips during rush hours. We thus analyze whether UberX trips are especially fast during those hours. To that end, we split our dataset into morning rush hour trips, afternoon rush hour trips, and off-peak trips. We match every UberTaxi trip to its 10 nearest UberX trips within its subsample, and compute the average treatment effect for each subsample, using the matching estimator from equation (9).

Table 11 shows the result from this exercise. UberX drivers handle their phones more relative to UberTaxi drivers during rush hour than during off-peak hours. The gap in hard accelerations, hard brakes, and speed high shrinks during rush hour. Whereas speed low is lower for UberX drivers during off-peak hours, it is higher during rush hours. Furthermore, UberX drivers tend to take longer routes during rush hour but their trips are faster, and they tend to pick up and drop off riders closer to the desired locations during the morning rush hour. This behavior is consistent with UberX drivers being more responsive to passenger preferences, paying more attention to riders who want to get to

their destination on time by driving faster (therefore braking and accelerating more), by handling the phone more to find better routes to avoid traffic, and by rerouting to ensure the rider gets to the destination more quickly.[23] It is perhaps surprising that we also find similar (but smaller) effects during the afternoon rush hour, when riders are not in as much of a hurry as in the morning.

Table 11: Heterogeneity in effect of UberX using a matching estimator

	Dependent variable:									
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Distance	Duration	Pickup	Dropoff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
UberX (Off-peak)	0.948***	0.061***	−0.066***	−0.282***	−0.054***	−0.265***	0.200***	0.178***	−0.077***	−0.210***
	(0.025)	(0.014)	(0.013)	(0.015)	(0.009)	(0.008)	(0.004)	(0.006)	(0.004)	(0.007)
UberX (AM rush)	0.989***	0.095***	−0.017	−0.221***	0.025**	−0.161***	0.229***	0.163***	−0.103***	−0.238***
	(0.032)	(0.015)	(0.016)	(0.018)	(0.011)	(0.009)	(0.005)	(0.009)	(0.007)	(0.009)
UberX (PM rush)	0.920***	0.081***	−0.040***	−0.209***	0.028**	−0.158***	0.176***	0.138***	−0.069***	−0.217***
	(0.032)	(0.016)	(0.014)	(0.018)	(0.013)	(0.009)	(0.007)	(0.010)	(0.008)	(0.011)
Observations	139,716	139,716	139,716	139,716	139,716	139,716	139,716	139,716	139,716	139,716

Note: This table show the effect of UberX varies during rush hours. In order to do so, we split the sample into three subsamples, off-peak trips, morning rush hour trips, and afternoon rush hour trips. Each column presents results for a different response variable. We use a matching estimator, where every UberTaxi trip is compared to the ten nearest UberX trips within the same subsample, based on origin and destination coordinates and time of the week. All driving metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Appendix C.2 presents similar results for scores. The full and control scores are lower for UberX during the morning rush hour, whereas the routing score is higher. However, these results do not necessarily reflect whether these adjustments are good or bad for drivers; as shown in Appendix D.4, riders' preferences are somewhat different during the morning rush hour.

6 Conclusions

Observers have expressed concern that ride-hailing platforms might reduce quality by allowing inexperienced drivers on their platform, a consequence of their streamlined screening process that is designed to allow for a flexible workforce. Using objective measures of driving quality, we provide empirical evidence that the ratings, incentives, nudges, and information systems set up by Uber constitute important offsetting forces

[23]Note, however, that this result could be driven by alternative explanations. For instance, it could be that UberX drivers are more likely to be in a hurry during rush hours, and thus drive faster.

that significantly improve the quality of driving. Ultimately, we find that UberX drivers provide quality that is roughly comparable to that provided by UberTaxi drivers. While we cannot fully attribute differences in quality to the differences in quality control mechanismsm, our findings suggest that concerns about low quality on ride-hailing platforms are misguided. Thus, our findings indicate that policymakers concerned about quality should consider allowing marketplaces to rely on well-designed ex-post quality control systems rather than on ex-ante screening and occupational licensing.

Our paper highlights the need to understand the specific channels that contribute to the effectiveness of these quality control mechanisms as well as to the differences between UberX and UberTaxi driving behavior. This question may be best answered by randomized experiments that can be conducted in the future. For instance, future research could investigate whether UberX drivers are primarily motivated by an intrinsic desire to create a good experience for passengers, or whether perceived or real economic incentives play a more important role in motivating behavior.

References

Abadie, Alberto and Guido W. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 2005, *74* (1), 235–267.

Allcott, Hunt and Judd B. Kessler, "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons," *American Economic Journal: Applied Economics*, January 2019, *11* (1), 236–76.

_ **and Todd Rogers**, "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *American Economic Review*, October 2014, *104* (10), 3003–37.

Angrist, Joshua D., Sydnee Caldwell, and Jonathan V. Hall, "Uber versus Taxi: A Driver's Eye View," *American Economic Journal: Applied Economics*, July 2021, *13* (3), 272–308.

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang, "Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index," *Working paper*, 2019.

Chen, M. Keith and Michael Sheldon, "Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Platform," *Working paper*, 2015.

_ , Judith A. Chevalier, Peter E. Rossi, and Emily Oehlsen, "The Value of Flexible Work: Evidence from Uber Drivers," *Journal of Political Economy*, 2019, *127* (6), 2735–2794.

Chevalier, Judith A. and Dina Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 2006, *43* (3), 345–354.

Choudhary, Vivek, Masha Shunko, Serguei Netessine, and Seongjoon Koo, "Nudging Drivers to Safety: Evidence from a Field Experiment," *Management Science*, 2022, *68* (6), 4196–4214.

Cook, Cody, Rebecca Diamond, Jonathan V Hall, John A List, and Paul Oyer, "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers," *The Review of Economic Studies*, 11 2020, *88* (5), 2210–2238.

Dellarocas, Chrysanthos, "Reputation mechanisms," *Handbook on Economics and Information Systems*, 2006, pp. 629–660.

Edwards, James T. and John A. List, "Toward an understanding of why suggestions work in charitable fundraising: Theory and evidence from a natural field experiment," *Journal of Public Economics*, 2014, *114*, 1 – 13.

Filippas, Apostolos, John J. Horton, and Joseph M. Golden, "Reputation Inflation," *Marketing Science*, 2022, *41* (4), 733–745.

Frey, Bruno S. and Stephan Meier, "Social Comparisons and Pro-social Behavior: Testing "Conditional Cooperation" in a Field Experiment," *American Economic Review*, December 2004, *94* (5), 1717–1722.

Gerber, Alan S, Donald P Green, and Ron Shachar, "Voting may be habit-forming: evidence from a randomized field experiment," *American Journal of Political Science*, 2003, *47* (3), 540–550.

Hall, Jonathan, John Horton, and Dan Knoepfle, "Ride-Sharing Markets Re-Equilibrate," *Working paper*, 2023.

Hall, Jonathan V and Alan B Krueger, "An analysis of the labor market for Uber's driver-partners in the United States," Technical Report, National Bureau of Economic Research 2016.

Handel, Benjamin R., "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, December 2013, *103* (7), 2643–82.

Jin, Ginger Zhe and Phillip Leslie, "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards," *The Quarterly Journal of Economics*, 2003, *118* (2), 409–451.

Jin, Yizhou and Thomas Yu, "How to Prevent Traffic Accidents: Moral Hazard, Inattention, and Behavioral Data," *Working paper*, 2021.

Katz, Lawrence F and Alan B Krueger, "The rise and nature of alternative work arrangements in the United States, 1995-2015," Technical Report, National Bureau of Economic Research 2016.

Kolstad, Jonathan T., "Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards," *American Economic Review*, December 2013, *103* (7), 2875–2910.

Liu, Meng, Erik Brynjolfsson, and Jason Dowlatabadi, "Do Digital Platforms Reduce Moral Hazard? The Case of Uber and Taxis," *Management Science*, 2021, *67* (8), 4665–4685.

Luca, Michael, "Reviews, reputation, and revenue: The case of Yelp.com," *Working Paper*, 2011.

__ **and Georgios Zervas**, "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 2016, *62* (12), 3412–3427.

Macharia, WM, G Leon, BH Rowe, BJ Stephenson, and R Haynes, "An overview of interventions to improve compliance with appointment keeping for medical services," *JAMA*, 1992, *267* (13), 1813–1817.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier, "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, August 2014, *104* (8), 2421–55.

Nosko, Chris and Steven Tadelis, "The limits of reputation in platform markets: An empirical analysis and field experiment," *Working paper*, 2015.

Proserpio, Davide and Georgios Zervas, "Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews," *Marketing Science*, 2017, *36* (5), 645–665.

Resnick, Paul and Richard Zeckhauser, "Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system," *The Economics of the Internet and E-commerce*, 2002, *11* (2), 23–25.

_ , _ , **John Swanson, and Kate Lockwood**, "The value of reputation on eBay: A controlled experiment," *Experimental economics*, 2006, *9* (2), 79–101.

Shang, Jen and Rachel Croson, "A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods," *The Economic Journal*, 2009, *119* (540), 1422–1439.

Tadelis, Steven, "Reputation and Feedback Systems in Online Platform Markets," *Annual Review of Economics*, 2016, *8* (1), 321–340.

Thaler, Richard H. and Cass R. Sunstein, *Nudge: Improving decisions about health, wealth, and happiness*, Penguin, 2009.

Zervas, Georgios, Davide Proserpio, and John W Byers, "A first look at online reputation on Airbnb, where every stay is above average," *Marketing Letters*, 2021, *32* (1), 1–16.

Online Appendix

Appendix A Construction of trip characteristics groups

In order to size the origin rectangles to have similar numbers of trips, we first divide the sample into two equally sized groups by origin latitude. Each group is then subdivided into two equally sized groups by origin longitude. Then we divide each group again by latitude, and repeat this process 6 times.[24] We follow an analogous process to divide the sample into 128 groups by destination coordinates.

The 15 hour-of-the-week intervals are: 7:00-9:00 am Mon-Fri, 9:00-11:00 am Mon-Fri, 11:00 am-1:00 pm Mon-Fri, 1:00-4:00 pm Mon-Fri, 4:00-6:00 pm Mon-Thu, 6:00-8:00 pm Mon-Thu, 8:00-10:00 pm Mon-Thu, 10:00 pm-1:00 am Mon-Thu, 4:00-8:00 pm Fri, 8:00 pm-midnight Fri-Sat, midnight-4:00 am Sat-Sun, 9:00 am-2:00 pm Sat-Sun, 2:00 pm-8:00 pm Sat, 2:00 pm-8:00 pm Sun, and all remaining times.

Figure 8 shows a histogram of the number of trips in the group each trip in our sample belongs to. We can see that the majority of trips are in groups with more than five trips.

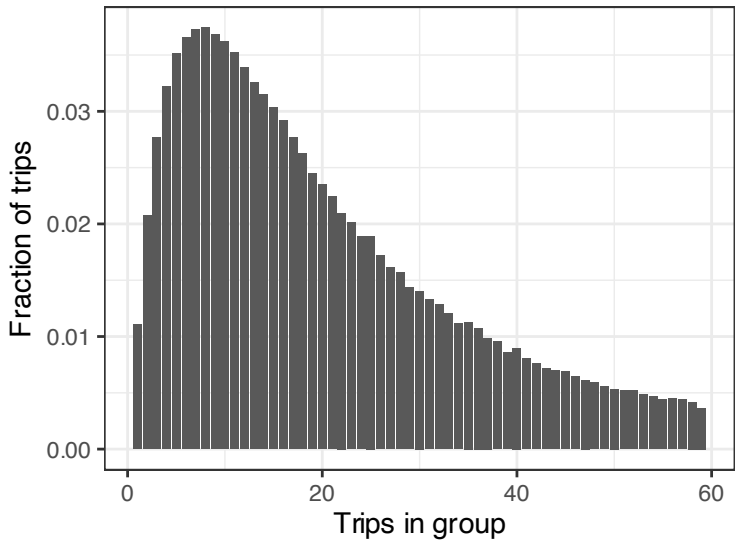


Figure 8: Distribution of trip group size

Note: We split the sample into groups by the origin and destination coordinates and by hour of the week. This figure shows the distribution of the number of trips in each group, weighted by number of trips.

[24] We are grateful to John Horton for suggesting this procedure.

Table 12: Selection of driving metrics

Metric	Penalty
Cont. speed 70	0.00001
Cont. speed mean	0.00003
Cont. speed 40	0.00003
Cont. speed 50	0.00004
Cont. speed 20	0.00004
Accelerations 3 m/s^2	0.0001
Brakes 3 m/s^2	0.0002
Cont. speed 100	0.0002
Cont. speed 30	0.0002
Accelerations 2.5 m/s^2	0.0003
Mounted	0.0008
Cont. speed 0	0.0012
Cont. speed 90	0.0012
Cont. speed 60	0.0013
Cont. speed 10	0.0014
Brakes 2.5 m/s^2	0.0023
Handling	0.0030
Brakes 2 m/s^2	0.0033
Accelerations 2 m/s^2	0.0033
Avg. speed when moving	0.0037
Cont. speed 80	0.0048
Dropoff gap	0.0093
Pickup gap	0.0093
Excess distance	0.0112
Excess duration	0.0178

Note: We fit a lasso regression of trip rating including all candidate metrics. The table shows the level of the penalty at which each variable is dropped.

Appendix B Variable selection and score construction

B.1 Selecting driving metrics

In order to select our main variables, we run a lasso regression of trip rating that includes all candidate metrics as well as their squares. We also include driver and trip characteristics fixed effects without penalization. The candidate metrics include metrics for brakes and accelerations using thresholds of 2, 2.5 and 3.06 m/s^2 (the industry standard is 3.06 m/s^2), and metrics for 12 different moments of the distribution of contextualized speeds within each trip (percentiles 0, 10, 20, ..., 100, as well as the mean).

Table 12 shows the order in which variables are dropped as we start increasing the penalty, and the value for the penalty at which they are dropped. Distance and duration are the most predictive variables. We choose the most predictive accelerations and brakes variables, those that use a threshold of 2 m/s^2. The most predictive speed variables are percentiles 80 and 10.

B.2 Score model

We tried a variety of ways of regularizing our model. In order to test them, we split our sample into three sets. The first is a train set with 47.5% of observations that we use to choose penalty parameters. We also have an estimation set with 47.5% of the data to estimate the model parameters. We set apart the remaining 5% of our observations as a test set. Our selection criterion is test-set mean square error (MSE).

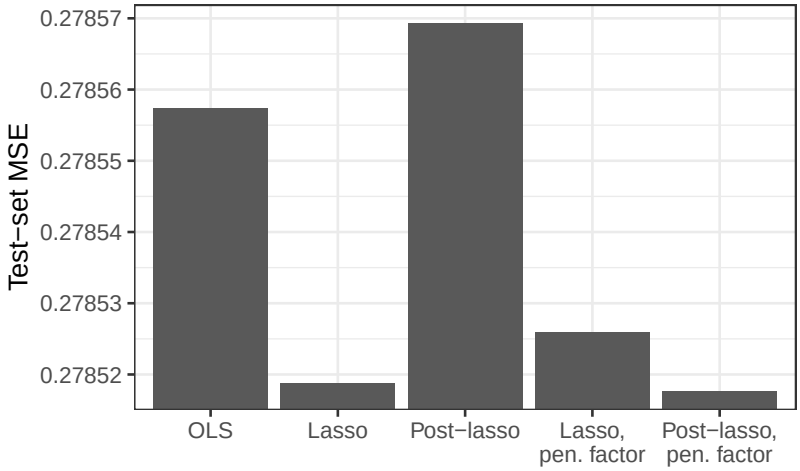


Figure 9: Performance of different models for score.

Note: In this figure we show the test-set mean-squared error for a variety of candidate models for trip rating as a function of a high-dimensional polynomial of driving metrics.

Our baseline model is an OLS regression with no penalization. We also run a lasso model with no penalties on fixed effects, as well as a post-lasso model that keeps all terms with nonzero coefficients in the lasso regression. We choose the penalty by 10-fold cross validation within the train set. We also run a lasso model with an increasing penalty factor. In other words, the penalty factor for an n-th order term is $\lambda\mu^n$, where λ is the base penalty for the model and μ is the penalty factor. We choose μ by 20-fold cross validation within the train set, and we choose λ by 10-fold cross validation within the remaining data in each fold.

Figure 9 compares the performance of all these models. The one that performs best is the post-lasso model with a penalty factor. The lasso model without a penalty factor performs almost as well. We prefer the post-lasso model with a penalty factor since the final model has no penalty, which means our coefficients have no asymptotic bias.

The final score we create uses the same procedure as the post-lasso model with a

penalty factor, but we use all the data to estimate it. In other words, we split the sample into two equally sized training and estimation sets (without leaving out any data in a test set).

While we use this procedure to choose our main methodology, if we use a different methodology the coefficients we measure throughout our paper change very little and the interpretation of all our results stays the same.

B.3 Imputing duration and distance metrics for UberTaxi trips

Since UberTaxi trips do not have estimated trip distance and duration, we cannot compute our duration and distance metrics. For that reason, we impute their estimated trip distance and duration based on random forests that we train on our sample of UberX trips. The covariates we use are the timestamp, hour of the day, day of the week, origin coordinates, and the destination coordinates. We also include several transformations of these variables to improve the fit of the random forest.[25]

Both random forests result in good predictions. The out-of-sample R^2 for the distance random forest is 0.958, and for the duration random forest it is 0.936. We use these random forests to impute the distance and duration of UberTaxi trips. We then use these imputed values to compute our driving metrics and scores for UberTaxi trips. In all results that use UberTaxi trips, we use these imputed metrics for UberX trips as well to ensure we use the same metrics.

B.4 Alternative samples for score construction

In order to measure how our results are affected by estimating preferences based on UberX trips only, we estimate regressions similar to those in Table 1, based on the sample of all trips (UberX and UberTaxi). Columns (1)-(2) in Table 13 present those results. We also reestimate our main results from Table 10 using scores that are constructed based on the full sample of trips (see Columns (1)-(2) in Table 14). All results look very similar to our main results in the paper.

[25] We include the sine and cosine of the day of the week times $2/7\pi$ (so that they lie on a unit circle) as well as a 45° rotation of those two variables. Finally, we include rotations of the origin and destination by multiples of 22.5°. We also include the difference between the origin and destination coordinates and rotations thereof, the straight-line distance between them, and the angle of the direction of the trip.

Table 13: Rating response to driving metrics for the full sample of trips and for UberTaxi trips

	Dependent variable: Rating					
	All trips		UberTaxi			
	(1)	(2)	(3)	(4)	(5)	(6)
Mounted	0.0099^{***}	0.0019	-0.0038	-0.0111	0.0044	0.0148^{**}
	(0.0009)	(0.0012)	(0.0093)	(0.0111)	(0.0040)	(0.0067)
Handling	-0.0018^{**}	-0.0056^{***}	-0.0165	-0.0020	-0.0119^{**}	-0.0059
	(0.0008)	(0.0008)	(0.0146)	(0.0105)	(0.0047)	(0.0057)
Brakes	-0.0084^{***}	-0.0030^{***}	0.0074	0.0091	-0.0037	-0.0025
	(0.0007)	(0.0005)	(0.0100)	(0.0058)	(0.0039)	(0.0038)
Accelerations	0.0010	-0.0035^{***}	0.0001	0.0045	0.00002	-0.0014
	(0.0007)	(0.0006)	(0.0090)	(0.0054)	(0.0038)	(0.0037)
Speed low	0.0041^{***}	0.0006	-0.0037	-0.0073	-0.0014	-0.0033
	(0.0006)	(0.0005)	(0.0113)	(0.0061)	(0.0040)	(0.0038)
Speed high	-0.0029^{***}	-0.0071^{***}	-0.0073	-0.0123	0.0052	0.0092^{*}
	(0.0007)	(0.0006)	(0.0140)	(0.0083)	(0.0056)	(0.0052)
Distance	-0.0148^{***}	-0.0148^{***}	-0.0172	-0.0125	-0.0014	0.0013
	(0.0007)	(0.0006)	(0.0175)	(0.0094)	(0.0048)	(0.0044)
Duration	-0.0284^{***}	-0.0252^{***}	-0.0293^{**}	-0.0223^{***}	-0.0220^{***}	-0.0190^{***}
	(0.0007)	(0.0006)	(0.0124)	(0.0068)	(0.0046)	(0.0041)
Pickup	-0.0127^{***}	-0.0110^{***}	-0.0073	-0.0033	-0.0135^{***}	-0.0122^{***}
	(0.0006)	(0.0005)	(0.0114)	(0.0062)	(0.0039)	(0.0036)
Dropoff	-0.0139^{***}	-0.0120^{***}	-0.0030	-0.0066	-0.0084^{**}	-0.0087^{**}
	(0.0007)	(0.0006)	(0.0121)	(0.0067)	(0.0039)	(0.0035)
Trip characteristics FE fine	✓	✓	✓	✓		
Trip characteristics FE coarse					✓	✓
Driver FE		✓		✓		✓
Observations	2,126,993	2,126,993	36,115	36,115	36,115	36,115

Note: This table shows results of regressions of five-star rating on driving metrics. Columns 1 and 2 show results for the full sample that combines UberX and UberTaxi trips. Columns 3 and 4 show results for UberTaxi trips using granular fixed effects. Columns 5 and 6 show results for UberTaxi trips using coarser fixed effects. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table 14: Comparison of UberX and UberTaxi trips for alternative scores

	Dependent variable:							
	Pooled F	Pooled C	Pooled R	Pooled S	Taxi F	Taxi C	Taxi R	Taxi S
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Matching estimator							
UberX	0.0753^{***}	0.2958^{***}	-0.2811^{***}	0.2790^{***}	-2.0838^{***}	-0.4644^{***}	-0.1468^{***}	-0.2803^{***}
	(0.0232)	(0.0232)	(0.0188)	(0.0091)	(0.0573)	(0.0357)	(0.0178)	(0.0148)
Observations	139,716	139,716	139,716	139,716	139,716	139,716	139,716	139,716
	Panel B: Trip characteristics fixed effects							
UberX	0.0990^{***}	0.2790^{***}	-0.2503^{***}	0.2690^{***}	-2.0055^{***}	-0.4213^{***}	-0.1317^{***}	-0.2651^{***}
	(0.0242)	(0.0241)	(0.0200)	(0.0094)	(0.0601)	(0.0373)	(0.0193)	(0.0148)
Observations	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729
	Panel C: Trip characteristics and rider fixed effects							
UberX	0.2005^{***}	0.3144^{***}	-0.1520^{***}	0.2782^{***}	-1.8475^{***}	-0.4859^{***}	-0.0131	-0.2869^{***}
	(0.0329)	(0.0265)	(0.0293)	(0.0121)	(0.0738)	(0.0490)	(0.0338)	(0.0257)
Observations	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729	7,147,729

Note: This table compares the driving behavior of UberX and UberTaxi trips according to scores constructed using alternative samples. Columns 1 through 4 present results for scores constructed on the full sample of UberX and UberTaxi trips. Columns 5-8 present results for scores constructed using only UberTaxi trips (to construct these scores, we fix the set of variables that remain after the variable selection for the main scores from section 3.2). As in Table 10, Panel A uses a matching estimator, while Panel B presents results from a linear regression that includes trip characteristics fixed effects. Panel C presents results from a linear regression that includes trip characteristics as well as rider fixed effects. All driving metrics are normalized to mean zero and variance one. Standard errors are adjusted as in Abadie and Imbens (2005) and clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

We also estimate similar results based on the sample of UberTaxi trips, although the very small number of UberTaxi trips limits us in ways that make us less confident about our results. In particular, we face a tradeoff between bias (failing to control granularly for trip characteristics) and variance (noisy estimates resulting from a small sample available to estimate effects when granular trip characteristics are included). Columns 3 and 4 in Table 13 show that if we run the exact same specification as in Table 1 on that sample, the coefficients we estimate on driving metrics have wide confidence intervals. This is not surprising since 87% of the trip characteristics groups constructed in Appendix A have no UberTaxi trips, and conditional on having at least one trip on average there are only 1.8 UberTaxi trips per group. In Columns 5 and 6 we present results based on coarser fixed effects.[26] The precision of the coefficients is higher, but the only coefficients that are significantly different from zero are those for handling, speed high, duration, pickup, and dropoff, suggesting that riders may prefer faster trips and better routing. Overall, these results suggest, with the caveats about omitted variable bias we have mentioned, that riders who take UberTaxi have preferences for higher speed.

Columns 4-6 in Table 14 present results based on scores that are estimated on the sample of UberTaxi trips. The variable selection is unstable if we use the same procedure as when creating our baseline scores, so we fix the set of covariates that remain after the variable selection for the baseline scores. Our findings suggest that UberTaxi riders may in fact prefer UberTaxi driving behavior. Note, however, that UberTaxi trips account for only 2% of our sample. Thus, even though this specific sample of riders may prefer the driving behavior of UberTaxi, the overwhelming majority of Uber riders prefer the behavior of UberX drivers or at least view them comparably.

[26] We divide the sample into 8 longitude groups by 8 latitude groups, in comparison to 128 x 128 groups for our baseline groups in Appendix A. We continue to use 15 hour of the week intervals as in Appendix A. For these coarser groups, 87% of groups have at least one UberTaxi trip, and conditional on having at least one there are on average 16 taxi trips. These frequencies are comparable to those for UberX using our baseline trip characteristics.

Appendix C Additional results

C.1 Response to riders preferences on scores

Table 15 presents results similar to those in Table 8, but where the outcome variables are scores. The results are almost identical if we use scores that are constructed using the sample of morning rush hour trips.

Table 15: Response to rider preferences

	Dependent variable:			
	Score F	Score C	Score R	Score S
	(1)	(2)	(3)	(4)
	Panel A: Main effect			
Mean by rider	0.151***	0.170***	0.066***	0.279***
	(0.006)	(0.003)	(0.006)	(0.003)
Observations	3,231,440	3,231,440	3,231,440	3,231,440
	Panel B: Heterogeneous effects by time of the week			
Mean by rider	0.129***	0.139***	0.053***	0.229***
× Off-peak	(0.008)	(0.003)	(0.008)	(0.003)
Mean by rider	0.275***	0.308***	0.141***	0.488***
× AM rush	(0.015)	(0.007)	(0.014)	(0.007)
Mean by rider	0.118***	0.150***	0.043***	0.243***
× PM rush	(0.014)	(0.006)	(0.013)	(0.006)
Observations	3,231,440	3,231,440	3,231,440	3,231,440

Note: This table shows regressions of quality metrics on leave-out means by rider. Only riders with more than 20 trips are included. All regressions control for driver and trip characteristics fixed effects. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

C.2 Heterogeneity in UberX effect on scores

Table 16 shows the heterogeneity in effect of UberX by morning and evening rush hours.

C.3 Distribution of distance metric for UberX and UberTaxi trips

Figure 10 shows the distribution of distance metrics for UberX and UberTaxi trips.

Table 16: Heterogeneity in effect of UberX using a matching estimator

	Dependent variable:			
	Score F	Score C	Score R	Score S
	(1)	(2)	(3)	(4)
UberX (Off-peak)	0.022	0.167***	−0.314***	0.305***
	(0.027)	(0.027)	(0.021)	(0.010)
UberX (AM rush)	−0.162***	0.014	−0.263***	0.281***
	(0.034)	(0.033)	(0.028)	(0.013)
UberX (PM rush)	0.041	0.127***	−0.177***	0.298***
	(0.038)	(0.034)	(0.033)	(0.014)
Observations	139,716	139,716	139,716	139,716

Note: This table how the effect of UberX on scores varies during rush hours. In order to do so, we split the sample into three subsamples, off-peak trips, morning rush hour trips, and afternoon rush hour trips. Each column presents results for a different response variable. We use a matching estimator, where every UberTaxi trip is compared to the ten nearest UberX trips within the same subsample, based on origin and destination coordinates and time of the week. All driving metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

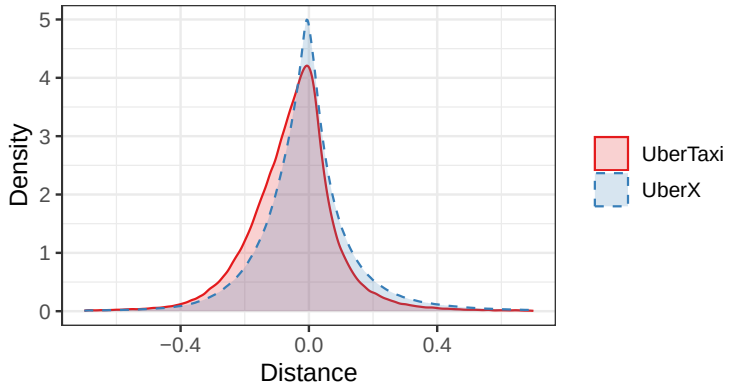


Figure 10: Distribution of distance metric for UberX and UberTaxi trips

Note: This figure presents kernel densities of the distribution of the distance metric for UberX and UberTaxi trips.

Appendix D Robustness checks

D.1 Rider fixed effects

In Table 17 we run the regressions from Table 1, but we also include rider fixed effects. The numbers change somewhat, but the interpretation of the coefficients does not change.

D.2 Traffic conditions

To control for traffic conditions, we exploit our detailed trip data to construct a metric for traffic conditions. For every trip, we residualize the average trip speed on granular origin

Table 17: Rating response to driving metrics

	Dependent variable:		
	Rating	Rating is 5	Rated
	(1)	(2)	(3)
Mounted	0.0005	0.0012^{***}	-0.0009^{**}
	(0.0010)	(0.0005)	(0.0004)
Handling	-0.0039^{***}	-0.0015^{***}	0.0002
	(0.0006)	(0.0003)	(0.0002)
Brakes	-0.0022^{***}	-0.0009^{***}	0.0016^{***}
	(0.0004)	(0.0002)	(0.0002)
Accelerations	-0.0030^{***}	-0.0016^{***}	0.0019^{***}
	(0.0005)	(0.0002)	(0.0002)
Speed low	0.0016^{***}	0.0008^{***}	-0.0022^{***}
	(0.0004)	(0.0002)	(0.0002)
Speed high	-0.0048^{***}	-0.0024^{***}	-0.0004^{**}
	(0.0005)	(0.0002)	(0.0002)
Distance	-0.0124^{***}	-0.0041^{***}	0.0007^{***}
	(0.0005)	(0.0002)	(0.0002)
Duration	-0.0197^{***}	-0.0077^{***}	0.0050^{***}
	(0.0005)	(0.0002)	(0.0002)
Pickup	-0.0109^{***}	-0.0052^{***}	0.0018^{***}
	(0.0004)	(0.0002)	(0.0002)
Dropoff	-0.0123^{***}	-0.0051^{***}	0.0002
	(0.0005)	(0.0002)	(0.0002)
Observations	1,991,742	1,991,742	6,901,200

Note: This table shows results of regressions of rating variables—five-star rating, a dummy for the rating being five, and a dummy for the trips being rated—on driving metrics. All regressions include three-way fixed effects by rider, driver, and trip characteristics. All metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

by destination fixed effects, which we create by following the same procedure described in Appendix A to split the our sample into 64 origin and 64 destination rectangles. We call this residual of trip speed the *speed deviation*, which tells us whether the trip speed was atypically high or low given the origin and destination. We then use this variable to construct a traffic variable for every trip: the leave-out mean of the speed deviation of other trips with similar origins and destinations that took place *during the exact same hour*. To define similar trips, we split Chicago into 8 origin groups and 8 destination groups, also following the procedure described in Section 3.1; similar trips are those that start and end in the same groups.[27]

[27]To prevent issues arising from groups of similar trips with too few trips, we also define coarser origin and destination groups, with 4 groups each. We use this second definition whenever the finer group has

Table 18 shows results similar to those in Table 1, but we also include this traffic variable as a control. In columns 1, 3, and 5, we simply control for traffic. In columns 2, 4, and 6, we control for a cubic function of traffic. In both cases we observe that the coefficients on our driving metrics are virtually unchanged after controlling for traffic.

Table 18: Rating response to driving metrics

	Dependent variable:					
	Rating		Rating is 5		Rated	
	(1)	(2)	(3)	(4)	(5)	(6)
Mounted	0.0015	0.0015	0.0010	0.0010	-0.0011^{**}	-0.0011^{**}
	(0.0013)	(0.0013)	(0.0006)	(0.0006)	(0.0005)	(0.0005)
Handling	-0.0050^{***}	-0.0051^{***}	-0.0016^{***}	-0.0017^{***}	-0.0003	-0.0003
	(0.0009)	(0.0009)	(0.0004)	(0.0004)	(0.0003)	(0.0003)
Brakes	-0.0035^{***}	-0.0035^{***}	-0.0016^{***}	-0.0016^{***}	0.0012^{***}	0.0012^{***}
	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Accelerations	-0.0038^{***}	-0.0038^{***}	-0.0017^{***}	-0.0017^{***}	0.0012^{***}	0.0012^{***}
	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Speed low	0.0030^{***}	0.0030^{***}	0.0014^{***}	0.0014^{***}	-0.0029^{***}	-0.0029^{***}
	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Speed high	-0.0067^{***}	-0.0067^{***}	-0.0037^{***}	-0.0037^{***}	-0.0027^{***}	-0.0027^{***}
	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Distance	-0.0159^{***}	-0.0159^{***}	-0.0049^{***}	-0.0049^{***}	0.0004^{*}	0.0003
	(0.0007)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Duration	-0.0224^{***}	-0.0222^{***}	-0.0085^{***}	-0.0084^{***}	0.0063^{***}	0.0064^{***}
	(0.0007)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Pickup	-0.0115^{***}	-0.0115^{***}	-0.0058^{***}	-0.0058^{***}	0.0080^{***}	0.0080^{***}
	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Dropoff	-0.0138^{***}	-0.0138^{***}	-0.0056^{***}	-0.0056^{***}	-0.0008^{***}	-0.0008^{***}
	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Traffic	-0.0019^{***}	-0.0023^{***}	-0.0010^{***}	-0.0012^{***}	-0.0008^{***}	-0.0011^{***}
	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Traffic^2		-0.00004^{***}		-0.00002^{**}		-0.00003^{***}
		(0.00001)		(0.00001)		(0.00001)
Traffic^3		0.000003^{***}		0.000001^{***}		0.000002^{***}
		(0.000001)		(0.000000)		(0.000000)
Trip characteristics FE	✓	✓	✓	✓	✓	✓
Driver FE	✓	✓	✓	✓	✓	✓
Observations	1,991,657	1,991,657	1,991,657	1,991,657	6,900,944	6,900,944

Note: This table shows results of regressions of rating variables—five-star rating, a dummy for the rating being five, and a dummy for the trips being rated—on driving metrics and traffic conditions. All metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

fewer than 100 trips.

D.3 Alternative trip characteristics fixed effects

In Table 19 we run the regressions from Table 1, but we use different fixed effects for trip characteristics. In columns 1, 3, and 5, we split space into 64 origin groups and 64 destination groups, and we split time into $24 \times 7 = 168$ hours of the week. In columns 2, 4, and 6, we split space into 32 origin groups and 832 destination groups, and we split time into hours. Columns (1)-(5) look almost identical to those in Table 1. We observe some changes in column (6), although the interpretation of the coefficients does not change.

D.4 Heterogeneity by time of the week

We run regressions similar to Equation (1), but we interact our metrics with dummies for whether the trip took place during the morning rush hour, the afternoon rush hour, or during off-peak hours. Table 20 shows the results when we use rating as our dependent variable. Every row corresponds to one metric, and every column represents one dummy for time of the week. The main difference across columns is that riders have stronger preferences for faster trips during the morning rush hour, consistent with people having to arrive to work on time. In some alternative specifications we also included dummies for trips that start and end at airports, but we did not find any noticeable difference with non-airport trips.

D.5 Response to car prices

One potential concern is the effect that car quality might have on riders' preferences. In order to address that issue, we construct a car price variable based on car make, model, year, and mileage. We do not observe mileage, so we assume that cars are driven twice the average mileage of 13,476 mi per year since the car was produced, given Uber cars are used more intensely than average cars. We use Kelley Blue Book data for prices that were collected manually by Uber. This is a time consuming task, so we only have prices for the most common car models, which account for roughly 60% of our trips.

Table 21 shows regressions of rating variables on driving metrics, as well as on prices. Columns (2), (4), and (6) also include the interaction of prices and driving metrics. Neither the car price nor its interactions seem to have any noticeable effect on ratings. Fur-

Table 19: Rating response to driving metrics

	Dependent variable:					
	Rating		Rating is 5		Rated	
	(1)	(2)	(3)	(4)	(5)	(6)
Mounted	0.0023*	0.0013	0.0012*	0.0011	−0.0006	−0.0005
	(0.0013)	(0.0014)	(0.0006)	(0.0007)	(0.0005)	(0.0005)
Handling	−0.0057***	−0.0065***	−0.0022***	−0.0021***	−0.0003	−0.0003
	(0.0008)	(0.0010)	(0.0004)	(0.0005)	(0.0003)	(0.0003)
Brakes	−0.0036***	−0.0045***	−0.0015***	−0.0017***	0.0016***	0.0003
	(0.0006)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Accelerations	−0.0042***	−0.0044***	−0.0017***	−0.0016***	0.0013***	−0.0007***
	(0.0006)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Speed low	0.0038***	0.0038***	0.0016***	0.0014***	−0.0027***	−0.0005**
	(0.0006)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Speed high	−0.0056***	−0.0059***	−0.0029***	−0.0028***	−0.0033***	−0.0051***
	(0.0006)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Distance	−0.0158***	−0.0131***	−0.0048***	−0.0038***	0.0008***	0.0016***
	(0.0007)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Duration	−0.0233***	−0.0257***	−0.0089***	−0.0096***	0.0058***	0.0046***
	(0.0007)	(0.0008)	(0.0003)	(0.0004)	(0.0002)	(0.0003)
Pickup	−0.0115***	−0.0124***	−0.0058***	−0.0063***	0.0088***	0.0096***
	(0.0005)	(0.0006)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Dropoff	−0.0138***	−0.0141***	−0.0055***	−0.0057***	−0.0006***	−0.0004*
	(0.0006)	(0.0007)	(0.0003)	(0.0003)	(0.0002)	(0.0002)
Trip char. FE 1	✓		✓		✓	
Trip char. FE 2		✓		✓		✓
Driver FE	✓	✓	✓	✓	✓	✓
Observations	1,991,742	1,991,742	1,991,742	1,991,742	6,901,200	6,901,200

Note: This table shows results of regressions of rating variables—five-star rating, a dummy for the rating being five, and a dummy for the trips being rated—on driving metrics. All metrics are normalized to mean zero and variance one. In odd columns, trip characteristics fixed effects are based on 64 origin groups and 64 destination groups as well as on hours of the week. In odd columns, trip characteristics fixed effects are based on 32 origin groups and 32 destination groups as well as on the exact hour. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

thermore, we do not observe any major changes from the main coefficients in Table 1.

D.6 Response to last rating

We conduct an exercise in which, instead of exploring how drivers respond to their app rating, we look at how they respond to the last rating they received. Let r_i represent the rating given to the driver after trip i. Let $l(i)$ represent the index of the last trip by driver $d(i)$ that received a rating before trip i takes place. Then $r_{l(i)}$ represents the last rating

Table 20: Heterogeneity in the response of rating to driving metrics by time of the week

	Dependent variable: Rating		
	Interaction of covariate with:		
	Off-peak	AM rush	PM rush
	(1)	(2)	(3)
Mounted	0.0008	0.0014	0.0036**
	(0.0013)	(0.0019)	(0.0018)
Handling	−0.0049***	−0.0073***	−0.0038***
	(0.0009)	(0.0018)	(0.0015)
Brakes	−0.0032***	−0.0031**	−0.0041***
	(0.0007)	(0.0015)	(0.0013)
Accelerations	−0.0036***	−0.0039***	−0.0033**
	(0.0007)	(0.0015)	(0.0013)
Speed low	0.0030***	0.0031**	0.0036***
	(0.0007)	(0.0015)	(0.0013)
Speed high	−0.0065***	−0.0044***	−0.0090***
	(0.0007)	(0.0016)	(0.0015)
Distance	−0.0147***	−0.0217***	−0.0164***
	(0.0008)	(0.0019)	(0.0015)
Duration	−0.0210***	−0.0351***	−0.0200***
	(0.0009)	(0.0017)	(0.0014)
Pickup	−0.0103***	−0.0156***	−0.0134***
	(0.0007)	(0.0016)	(0.0012)
Dropoff	−0.0126***	−0.0141***	−0.0191***
	(0.0008)	(0.0015)	(0.0015)
Obervations: 1,991,742			

Note: This table shows the result of one regression of rating variables on quality metrics interacted with dummies for morning and afternoon rush hour trips. Rows represent quality metrics, and columns represent rush hour dummies. All ratings are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

that the driver received before the trip started. We run regressions of the form

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \alpha r_{l(i)} + P_2(n_i;\beta) + \epsilon_{ijkt} \tag{11}$$

where $P_2(n_i;\beta)$ is quadratic function of the number of trips completed by the driver before trip i.

A potential concern with giving this finding a causal interpretation is that there may be factors that lead to serial correlation in driver behavior. In order to isolate the effect of the rating, we use two different instrumental variables strategies. We wish to focus on variation in the previous rating that is not explained by the driver's own behavior or

Table 21: Response to ratings and notifications, including car prices

	Dependent variable:					
	Rating		Rating is 5		Rated	
	(1)	(2)	(3)	(4)	(5)	(6)
Mounted	0.0040**	0.0030	0.0018**	0.0015	−0.0005	−0.0006
	(0.0016)	(0.0032)	(0.0008)	(0.0015)	(0.0006)	(0.0012)
Handling	−0.0043***	−0.0018	−0.0017***	−0.0002	−0.0002	−0.0008
	(0.0010)	(0.0021)	(0.0005)	(0.0010)	(0.0004)	(0.0008)
Brakes	−0.0037***	−0.0045***	−0.0014***	−0.0015**	0.0014***	0.0013**
	(0.0007)	(0.0014)	(0.0003)	(0.0007)	(0.0003)	(0.0005)
Accelerations	−0.0027***	−0.0030**	−0.0012***	−0.0007	0.0014***	0.0019***
	(0.0008)	(0.0015)	(0.0004)	(0.0007)	(0.0003)	(0.0006)
Speed low	0.0036***	0.0039***	0.0016***	0.0019***	−0.0028***	−0.0026***
	(0.0007)	(0.0014)	(0.0003)	(0.0007)	(0.0003)	(0.0005)
Speed high	−0.0059***	−0.0081***	−0.0032***	−0.0041***	−0.0026***	−0.0023***
	(0.0008)	(0.0014)	(0.0004)	(0.0007)	(0.0003)	(0.0005)
Distance	−0.0152***	−0.0141***	−0.0043***	−0.0036***	0.0003	0.0003
	(0.0008)	(0.0016)	(0.0004)	(0.0007)	(0.0003)	(0.0005)
Duration	−0.0237***	−0.0260***	−0.0091***	−0.0108***	0.0057***	0.0056***
	(0.0008)	(0.0016)	(0.0004)	(0.0008)	(0.0003)	(0.0006)
Pickup	−0.0121***	−0.0112***	−0.0061***	−0.0056***	0.0087***	0.0080***
	(0.0007)	(0.0013)	(0.0003)	(0.0006)	(0.0002)	(0.0005)
Dropoff	−0.0149***	−0.0144***	−0.0061***	−0.0062***	−0.0006**	−0.0003
	(0.0008)	(0.0014)	(0.0004)	(0.0007)	(0.0003)	(0.0005)
Price	0.0015*	0.0015*	0.0006	0.0006	0.0004	0.0004
	(0.0008)	(0.0008)	(0.0004)	(0.0004)	(0.0003)	(0.0003)
Price × Mounted		0.0001		0.00003		0.00002
		(0.0004)		(0.0002)		(0.0001)
Price × Handling		−0.0003		−0.0002*		0.0001
		(0.0002)		(0.0001)		(0.0001)
Price × Brakes		0.0001		0.00001		0.00001
		(0.0002)		(0.0001)		(0.0001)
Price × Accels.		0.00005		−0.0001		−0.0001
		(0.0002)		(0.0001)		(0.0001)
Price × Speed low		−0.00004		−0.00004		−0.00003
		(0.0002)		(0.0001)		(0.0001)
Price × Speed high		0.0003*		0.0001		−0.00004
		(0.0002)		(0.0001)		(0.0001)
Price × Dist.		−0.0002		−0.0001		−0.00001
		(0.0002)		(0.0001)		(0.0001)
Price × Dur.		0.0003*		0.0002***		0.000004
		(0.0002)		(0.0001)		(0.0001)
Price × Pickup		−0.0001		−0.0001		0.0001*
		(0.0002)		(0.0001)		(0.0001)
Price × Dropoff		−0.0001		0.00001		−0.00004
		(0.0002)		(0.0001)		(0.0001)
Trip characteristics FE	✓	✓	✓	✓	✓	✓
Driver FE	✓	✓	✓	✓	✓	✓
Observations	1,226,114	1,226,114	1,226,114	1,226,114	4,287,352	4,287,352

Note: This table shows regressions of the trip rating on driving metrics, car price, and their interaction. All metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

changing characteristics (e.g. car condition). We instrument for the last rating using the average residual of other similar trips, where by similar we mean trips taken on the same calendar day and hour in the same location. Thus, if exogenous factors lead all drivers to deliver an experience that riders perceive as low quality (e.g. traffic accidents, weather), this shock to the driver's rating is unrelated to driver-specific changes over time.

To implement this, we first take the residual from the model in Equation (3). Then we group trips by 16 origin and destination areas and by calendar day and hour. The instrument for the previous rating is the average residual of all other trips that took place in the group corresponding to the previous trip. The second instrument is the leave-out average of all ratings given by the previous rider.[28]

Table 22 shows the result of these regressions. Panel A shows estimates from an OLS regression, and Panels B and C show results of 2SLS regressions. Throughout, our results are consistent with Tables 4 and 5: ratings have a negative effect on new ratings, but they do not have a large effect on our telemetry metrics or scores.

Table 22: Response to last rating

	Dependent variable:				
	Rating	Score F	Score C	Score R	Score S
	(1)	(2)	(3)	(4)	(5)
	Panel A: OLS				
Last rating	-0.0040^{***}	0.0026	0.0014	0.0008	-0.0008
	(0.0008)	(0.0025)	(0.0011)	(0.0024)	(0.0008)
Observations	1,979,500	6,861,226	6,861,226	6,861,226	6,861,226
	Panel B: IV, average rating by rider				
Last rating	-0.0047^{**}	-0.0001	0.0012	-0.0017	-0.0022
	(0.0021)	(0.0077)	(0.0035)	(0.0073)	(0.0025)
Observations	1,811,637	6,284,068	6,284,068	6,284,068	6,284,068
	Panel C: IV, both instruments				
Last rating	-0.0142^{***}	-0.0037	-0.0016	0.0003	-0.0052^{*}
	(0.0026)	(0.0086)	(0.0040)	(0.0083)	(0.0028)
Observations	771,482	2,689,979	2,689,979	2,689,979	2,689,979
	All safety metrics are normalized to mean zero and variance one.				

Note: This table shows regressions of quality metrics and scores on the rating for the last trip completed by the driver. Panel A presents an OLS estimator. Panel B presents a 2SLS estimator, where we use the rating of trips that took place nearby and during the same time as an instrument. Panel B also presents a 2SLS estimator, where we use the leave-out average rating by rider as an instrument. All driving metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

[28]Our results are very similar if we exclude trips with riders with fewer than 10 trips

Table 23: Response to last rating

	Dependent variable:									
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Distance	Duration	Pickup	Dropoff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: OLS									
Last rating	0.0012***	−0.0011*	−0.0010*	−0.0004	−0.0010*	0.0006	0.0005	−0.0004	0.0004	−0.0005
	(0.0004)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0005)	(0.0006)	(0.0005)	(0.0005)
Observations	6,861,226	6,861,226	6,861,226	6,861,226	6,861,226	6,861,226	6,861,226	6,861,226	6,861,226	6,861,226
	Panel B: IV, average rating by rider									
Last rating	0.0016	−0.0006	−0.0034*	−0.0039**	0.0015	0.0021	0.0008	−0.0024	−0.0013	0.0004
	(0.0013)	(0.0016)	(0.0018)	(0.0018)	(0.0019)	(0.0018)	(0.0017)	(0.0017)	(0.0017)	(0.0015)
Observations	6,284,068	6,284,068	6,284,068	6,284,068	6,284,068	6,284,068	6,284,068	6,284,068	6,284,068	6,284,068
	Panel C: IV, both instruments									
Last rating	0.0031**	−0.0023	−0.0008	−0.0021	−0.0005	0.0023	−0.0008	−0.0019	0.0002	0.0019
	(0.0014)	(0.0018)	(0.0021)	(0.0020)	(0.0021)	(0.0019)	(0.0019)	(0.0020)	(0.0019)	(0.0017)
Observations	2,689,979	2,689,979	2,689,979	2,689,979	2,689,979	2,689,979	2,689,979	2,689,979	2,689,979	2,689,979

Note: This table shows regressions of quality metrics and scores on the rating for the last trip completed by the driver. Panel A presents an OLS estimator. Panel B presents a 2SLS estimator, where we use the rating of trips that took place nearby and during the same time as an instrument. Panel B also presents a 2SLS estimator, where we use the leave-out average rating by rider as an instrument. All driving metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

D.7 Excluding trips before warnings

Tables 24 and 25 present similar results to those in Panel A on Tables 4 and 5, but in which the last few trips before receiving a warning are excluded from the sample. We observe very similar results.

Appendix E Further details about the deactivation process

Table 26 gives a sense of the number of drivers in each one of the ranges for the app rating (the average rating from the last 500 trips). This table also shows how many drivers already completed 500 rated trips, so that every additional trip only contributes one five hundredth to the rating after the trip.

To get a sense of how strong these incentives are, Figure 11 shows the percentage of drivers that are at risk of falling below each threshold. It shows how many 3-star trips the driver would have to complete in order to fall below the threshold. Only a very small number of drivers are likely to eventually reach the 4.4 threshold for deactivation. A somewhat more important fraction of drivers are close and even below the 4.6 threshold

Table 24: Response to ratings and notifications

	Dependent variable:				
	Rating	Score F	Score C	Score R	Score S
	(1)	(2)	(3)	(4)	(5)
	Panel A: Excluding 10 trips before warning				
Has received notif.	0.086***	0.096***	0.035*	0.083***	0.016**
	(0.007)	(0.022)	(0.019)	(0.019)	(0.008)
App rating	−0.296***	0.075**	0.043*	0.071**	0.028**
	(0.014)	(0.037)	(0.026)	(0.033)	(0.012)
Observations	1,973,515	6,843,534	6,843,534	6,843,534	6,843,534
	Panel B: Excluding 20 trips before warning				
Has received notif.	0.075***	0.087***	0.018	0.083***	0.013
	(0.008)	(0.024)	(0.021)	(0.021)	(0.008)
App rating	−0.310***	0.076**	0.037	0.075**	0.030**
	(0.014)	(0.038)	(0.026)	(0.034)	(0.013)
Observations	1,966,316	6,819,548	6,819,548	6,819,548	6,819,548

Note: This table shows regressions of ratings and scores on app ratings and dummies for having received notifications, similar to those in Panel A of table 4. Panel A excludes the last 10 trips before receiving a warning, and Panel B excludes the last 20 trips before receiving a warning. All regressions include driver and trip characteristics fixed effects, and control for a quadratic function of the number of Uber trips the driver has completed. Standard errors are clustered by driver. All driving metrics are normalized to mean zero and variance one. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

Table 25: Response to ratings and notifications

	Dependent variable:									
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Distance	Duration	Pickup	Dropoff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Excluding 10 trips before warning									
Has received notif.	0.050***	−0.038***	0.004	−0.012	0.013**	−0.0001	−0.011***	−0.017***	−0.014***	−0.004
	(0.013)	(0.013)	(0.008)	(0.008)	(0.006)	(0.006)	(0.004)	(0.005)	(0.005)	(0.004)
App rating	0.040**	−0.023	0.009	0.002	0.044***	0.017*	0.009	−0.025***	−0.007	−0.013**
	(0.018)	(0.019)	(0.011)	(0.012)	(0.009)	(0.009)	(0.007)	(0.007)	(0.007)	(0.007)
Observations	6,843,534	6,843,534	6,843,534	6,843,534	6,843,534	6,843,534	6,843,534	6,843,534	6,843,534	6,843,534
	Panel B: Excluding 20 trips before warning									
Has received notif.	0.044***	−0.031**	0.007	−0.009	0.008	−0.001	−0.009**	−0.018***	−0.011**	−0.006
	(0.015)	(0.015)	(0.009)	(0.009)	(0.007)	(0.007)	(0.004)	(0.005)	(0.006)	(0.005)
App rating	0.038**	−0.015	0.012	0.001	0.045***	0.016	0.009	−0.025***	−0.003	−0.015**
	(0.018)	(0.019)	(0.012)	(0.012)	(0.009)	(0.010)	(0.007)	(0.007)	(0.007)	(0.007)
Observations	6,819,548	6,819,548	6,819,548	6,819,548	6,819,548	6,819,548	6,819,548	6,819,548	6,819,548	6,819,548

Note: This table shows regressions of metrics on app ratings and dummies for having received notifications, similar to those in Panel A of table 4. Panel A excludes the last 10 trips before receiving a warning, and Panel B excludes the last 20 trips before receiving a warning. All regressions include driver and trip characteristics fixed effects, and control for a quadratic function of the number of Uber trips the driver has completed. Standard errors are clustered by driver. All driving metrics are normalized to mean zero and variance one. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

for the first notification. This suggests that if this deactivation process has an effect on driving behavior it is most likely through behavioral nudges instead of through actual incentives.

Table 27 shows how frequently drivers cross one of these thresholds, both from above and from below. We see that there is a large number of events, even for the threshold at

Table 26: Number of trips during which driver ratings satisfy each condition

	Number	Fraction
Total	6,901,197	
Lifetime trips >500	4,412,839	0.639
Rating <4.6	507,310	0.074
Rating <4.5	198,536	0.029
Rating <4.4	89,375	0.013

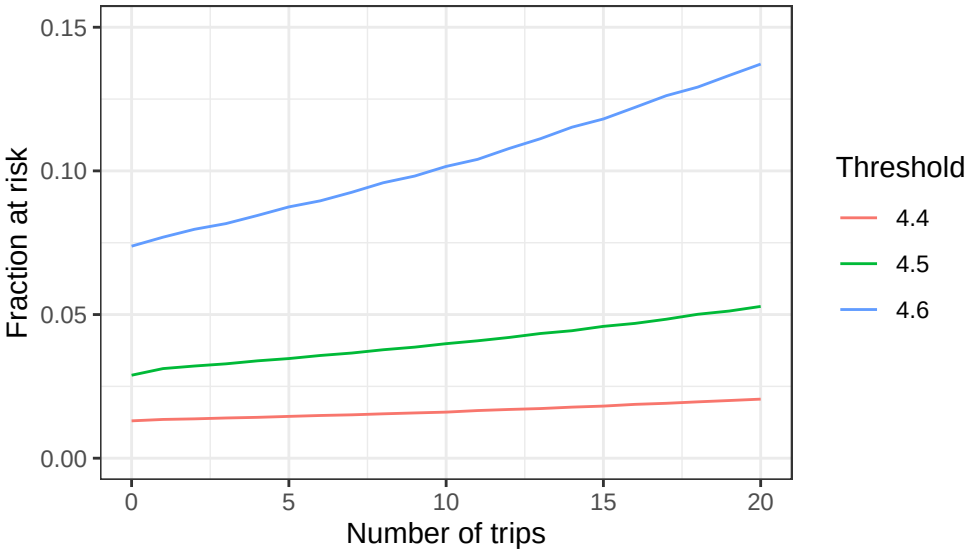


Figure 11: Fraction of drivers at risk of falling below rating thresholds

Note: This plots shows the fraction of drivers whose app rating would fall below a certain threshold if the next N consecutive trips received a 3-star rating, where N is displayed on the x-axis. The values corresponding to $N = 0$ represent the fraction whose app rating currently falls below the threshold.

4.4, close to which there are not that many drivers.

Table 27: Number of threshold crossings

	From above		From below	
Threshold	Crossings	Unique drivers	Crossings	Unique drivers
4.6	7,201	4,650	6,896	4,372
4.5	4,758	3,240	4,181	2,778
4.4	3,476	2,612	2,670	1,955

Note: Number of events in our dataset in which a driver's rating crosses one of the rating thresholds, either from above or from below.

Appendix F Additional experimental results

F.1 Balance of experimental sample

Table 28 shows results of a balance test for mean pre-experiment period outcomes for each driver. Estimates are across trips in the experimental period and are clustered by driver. While all of the non-speed metrics and scores have insignificant results, there seems to be some difference in the speed outcomes.

Table 28: Balance Test for Experiment

	Dependent variable:						
	Score F	Score C	Score R	Score S	Mounted	Handling	Brakes
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	-0.009	-0.017	-0.007	-0.006	0.701^{***}	0.102^{***}	0.219^{***}
	(0.014)	(0.014)	(0.010)	(0.005)	(0.005)	(0.002)	(0.001)
Treatment	0.016	0.029	0.013	0.011^{*}	0.0003	-0.002	0.00001
	(0.018)	(0.018)	(0.013)	(0.006)	(0.007)	(0.003)	(0.001)
Observations	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965
	Dependent variable:						
	Accels.	Speed low	Speed high	Distance	Duration	Pickup	Dropoff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	0.186^{***}	23.735^{***}	86.038^{***}	0.014^{***}	0.056^{***}	-0.00001	-0.0001
	(0.001)	(0.060)	(0.053)	(0.0004)	(0.001)	(0.00004)	(0.0001)
Treatment	-0.001	0.234^{***}	0.159^{**}	-0.002^{***}	-0.0003	0.00002	0.0002^{*}
	(0.002)	(0.079)	(0.069)	(0.001)	(0.001)	(0.0001)	(0.0001)
Observations	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965

Note: The table shows regressions of driving metrics and scores on a treatment dummy. We focus on trips that took place before the beginning of the experiment. All metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

F.2 Effect of Dashboard Experiment on Metrics

Table 29 shows results for regressions of the form in Equations 6 and 7 where the dependent variables are the quality metrics.

Table 29: Results of experiment, metrics

	Dependent variable:									
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Distance	Duration	Pickup	Dropoff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: Intent to treat estimator									
Treatment	−0.004	−0.015**	−0.0004	0.002	0.003	−0.002	−0.002	−0.006***	−0.002	−0.0005
	(0.006)	(0.006)	(0.004)	(0.005)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
Pre-Period Mean	2.081***	3.697***	4.679***	4.836***	0.051***	0.074***	1.691***	1.871***	13.136***	5.392***
	(0.009)	(0.036)	(0.031)	(0.030)	(0.0004)	(0.001)	(0.028)	(0.019)	(0.340)	(0.407)
Observations	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,682,194	3,692,289	3,785,965	3,785,965
	Panel B: 2SLS estimator									
Interaction	−0.005	−0.032***	−0.010	−0.005	0.006	−0.004	−0.001	−0.007*	−0.001	0.001
	(0.012)	(0.011)	(0.008)	(0.009)	(0.006)	(0.006)	(0.003)	(0.004)	(0.003)	(0.003)
Observations	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,785,965	3,682,194	3,692,289	3,785,965	3,785,965

Note: Panel A shows regressions of driving metrics on a dummy for being in the treatment group of the dashboard experiment. Panel B shows 2SLS regressions of driving metrics on a dummy for observing the dashboard, using the treatment dummy as an instrument. All metrics are normalized to mean zero and variance one. Standard errors are clustered by driver. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.