Service Quality in the Gig Economy: Empirical Evidence about Driving Quality at Uber*

Susan Athey† Juan Camilo Castillo‡ Bharat Chandar§

September 24, 2019

Abstract

The rise of marketplaces for goods and services has led to changes in the mechanisms used to ensure high quality. We analyze this phenomenon in the Uber market, where the system of pre-screening that prevailed in the taxi industry has been diminished in favor of (automated) quality measurement, reviews, and incentives. This shift allows greater flexibility in the workforce but its net effect on quality is unclear. Using telematics data as an objective quality outcome, we show that UberX drivers provide better quality than UberTaxi drivers, controlling for all observables of the ride. We then explore whether this difference is driven by incentives, nudges, and information. We show that riders' preferences shape driving behavior. We also find that drivers respond to both user preferences and nudges, such as notifications when ratings fall below a threshold. Finally, we show that informing drivers about their past behavior increases quality, especially for low-performing drivers.

*We are grateful for funding from the Sloan Foundation and the Stanford Cyber Initiative, and to numerous seminar audiences for useful feedback. The conclusions of this paper are those of the authors and do not represent the views of any corporation or institution.

†Stanford Graduate School of Business. E-mail: athey@stanford.edu

‡Economics Department, Stanford University. E-mail: jccast@stanford.edu

§Stanford Graduate School of Business. E-mail: chandarb@stanford.edu. This paper was completed while Bharat Chandar was an employee of Uber.

1 Introduction

Many industries have been transformed in the last few years with the arrival of the gig economy. Travelers can book a room from an apartment owner on Airbnb, commuters can get to their destination by ordering a ride from Uber, and people can get a wide variety of tasks done by hiring someone on TaskRabbit. These platforms allow small, independent providers to earn money by using their underutilized time or goods (such as cars or housing), and they potentially reduce prices and offer more variety and flexibility for consumers.

The mechanisms used by these platforms to ensure quality differ substantially from what used to be the norm. Traditional companies screen their employees beforehand to ensure they will comply with their standards. This is typically a burdensome and lengthy process that results in a fixed pool of full time workers. In contrast, a key feature of the gig economy is a streamlined screening process that enables a flexible pool of independent contractors who work during their free time and smooth their income during unemployment spans (Katz and Krueger, 2016; Chen et al., 2017; Angrist and Caldwell, 2017). New platforms thus rely on a variety of new technological possibilities and market forces to ensure that service providers are incentivized to provide high quality services.

For example, many platforms use rating systems extensively, allowing customers to monitor the quality of service and share the information with other consumers. Apps have simple interfaces that enable customers to rate service providers with little effort. Such platform companies may combine ratings and other quality tracking systems with incentive systems that remove individuals who violate quality standards and reward those who do well. Some companies, such as ride-hailing apps, also collect objective measures of quality like telematics data collected in real time. Platforms can share this individualized quality information with service providers on their platform in the form of "nudges" that remind drivers to perform well.[1]

It is unclear whether shifting from ex-ante screening towards ex-post quality control has led to higher or lower quality. Lower barriers to entry may allow some service providers who provide low quality to join and work on a platform. With ex-post quality control, low quality providers might not be detected until after a period of observation, which can result in low quality service for a few unlucky consumers. On the other hand, in more traditional industries with high upfront

[1]We define a nudge as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" Thaler and Sunstein (2008).

screening and entry barriers, there can be few incentives to provide good service after passing the initial screening.

In this paper we analyze whether the shift to ex-post quality control has increased or decreased quality in transportation services. We focus on UberX, Uber's main ride-hailing product, which provides an ideal setting for two reasons. First, we observe telemetry measures like speed, acceleration, braking, and phone handling, which means we have objective measures of quality. This is unusual; in most other markets only subjective measures like ratings or reviews are available. Second, Uber has a product called UberTaxi that allows riders to request standard taxis licensed by the local government.[2] This allows a direct comparison between UberX, a service in which quality is largely controlled through ratings feedback, incentives, and nudges, and UberTaxi, which relies on traditional screening through the licensing process and is affected to a much lesser extent by Uber's rating and incentive systems.[3]

Our main finding is that UberX drivers perform better than UberTaxi drivers. According to our main measure of driving quality—a score that summarizes riders' preferences for driving metrics—UberX trips are 0.16 standard deviations better than similar UberTaxi trips. Differences in quality control mechanisms most likely do not account for this difference in its entirety. We do find, however, that the performance of UberX drivers improves when incentives are stronger, when they receive nudges about low ratings, and when Uber shares more detailed information about past performance. This is evidence that the elements Uber has set up to control quality are responsible, at least in part, for UberX providing a better service than taxis.

We start by analyzing riders' preferences for driving behavior using 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 fixed effects, origin, destination, and time of the week, we find that riders give higher ratings on trips with fewer strong brakes and accelerations. They also prefer trips

[2]To drive a taxi in Chicago 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. To drive on UberX, a driver must be at least 21; have valid driver's license, vehicle registration, and insurance; and pass an online background check that reviews driving record and criminal history.

[3]Although some elements of Uber's rating and incentive systems are also present for UberTaxi trips, there is no process to take drivers out of the platform. Furthermore, only a small share of UberTaxi drivers' trips come from the Uber app, which means those incentives play a much smaller role for them.

where the driver does not handle their 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 soon to the destination.

We build a score that aggregates driving metrics into a one dimensional measure of driving quality according to our predictive model. We then use it to compare UberX and UberTaxi trips. UberX drivers perform significantly better: the score is on average 0.16 standard deviations higher than for UberTaxi trips that are similar in terms of origin, destination, and time of the week. Looking at individual metrics, UberX drivers have fewer hard brakes and accelerations, and they are much more likely to mount their phone. They drive somewhat slower, but they tend to drive at a steadier speed. However, they are more likely to handle their phone, which is not surprising given that they need to interact more with the Uber app.

These results are consistent with UberX drivers responding to ratings, incentives, and nudges. There are, however, several alternative explanations that might account for them. An UberX trip might be a more personal experience than an UberTaxi trip, which might result in intrinsic motivation in drivers. Monetary incentives are also somewhat different: UberX drivers typically own their car, whereas a large fraction of UberTaxi drivers do not, and fare structures are slightly different. We are not able to fully decompose the gap between UberX and Uber Taxi into all these possible channels. Instead, we focus on the role of Uber's quality control systems. The rest of our paper presents a variety of empirical evidence that sheds light on the role of Uber's ratings, incentives, and nudges.

We first examine the extent to which riders' preferences affect drivers' behavior. We find correlation in driving behavior for trips within the same rider, suggesting that drivers respond to what riders want. We also find that, relative to UberTaxi trips, UberX trips are faster and have more hard brakes and accelerations during rush hour, when riders are more likely to be in a hurry. This is consistent with UberX providing stronger incentives to cater to riders' preferences.

The incentives and information provided through ratings might also play an important role. When drivers' ratings fall below certain thresholds, drivers get notifications with the aim of improving their behavior, and if their ratings keep decreasing, drivers are eventually taken off the platform. We find that the difference between UberX and UberTaxi behavior is greater for drivers that have a low rating and are thus at a greater risk of being taken off the platform. We also see that drivers' quality improves substantially after they receive notifications. This effect is

strong even for the first notification, which occurs far away from the threshold at which they are kicked off the platform. Thus, we conclude that, besides direct economic incentives, ratings and notifications work as behavioral nudges that induce better driving behavior.

Giving drivers feedback about their past behavior also plays a role. At the time, Uber sent a weekly report to drivers summarizing how they performed according to telematics metrics. During our period of analysis, 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, which is mainly driven by drivers who performed in the bottom 10th percentile before the experiment started.

There are additional ways in which Uber's ratings and incentives can contribute to higher quality. For instance, the rating system ensures that drivers that provide low quality are weeded out of the system. Another example concerns the stronger incentives provided by a centralized platform that takes complaints more seriously than a city government. These channels only underscore our main point: Despite much simpler screening, UberX drivers provide better driving quality than UberTaxi drivers, and part of the difference can be attributed to ratings, incentives, nudges, and information. However, we are not able to precisely quantify the relative contribution of the different components.

This paper is organized as follows. Section 2 describes the Uber market and the data we use in our analysis. Section 3 analyzes riders' preferences over driving metrics in order to construct a one dimensional rating score. We then use all metrics and our rating score to compare the behavior of UberX and UberTaxi drivers in Section 4. In Section 5 we decompose driving behavior into trip, driver, and rider characteristics. In Section 6 we analyze ratings, incentives, and behavioral nudges and how they affect driving behavior, and we conclude in Section 7.

Related Work

In a paper that is closely related to ours, Liu et al. (2018) 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 explore the mechanisms through which incentive

systems affect drivers' behavior.

Our work, as well as the work by Liu et al., is part of a broader literature that analyzes rating and review systems in the digital economy (Dellarocas, 2006; Tadelis, 2016). In an early work on eBay, Resnick and Zeckhauser (2002) find that buyers review very often despite the incentive to free-ride. There is evidence that consumers respond to ratings and reviews, in the context of online bookstores (Chevalier and Mayzlin, 2006) and restaurants (Luca, 2011). Most of this work focuses on the behavior of consumers when they rank providers (Filippas et al., 2017; Nosko and Tadelis, 2015) and on how consumers respond to ratings and rankings. In contrast, we focus on provider behavior and how it is influenced by rating behavior, as in Mayzlin et al. (2014), where the authors find evidence that hotel owners post positive reviews for themselves and negative reviews for neighboring hotels. In contrast to this paper, we study the quality of the service provided and how it responds to information and incentives.

Many works analyze how nudges and information affect agents' behavior, beyond incentives they might provide (Leonard, 2008; Allcott and Kessler, 2017). They may affect behavior through intrinsic motivation and reoptimization in response to new information (Kolstad, 2013). Other papers provide evidence that nudges have consequences beyond what would have been expected from rational agents in different contexts, such as with energy consumption (Allcott and Rogers, 2014), savings (Buessing and Soto, 2006), 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). More generally, some works analyze how to incentivize people, emphasizing how forming habits results in persistent effects even after incentives cease (Charness and Gneezy, 2009; Acland and Levy, 2015).

A growing literature analyzes several aspects of Uber and ride-hailing markets. Our work most closely relates to several papers that analyze the labor supply side (Hall and Krueger, 2016; Chen and Sheldon, 2015; Hall et al., 2017; Cook et al., 2018). A recurring theme is the value of labor flexibility introduced by ride hailing platforms (Angrist and Caldwell, 2017; Chen et al., 2017). 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 explores the reasons why this is possible thanks to the shift in quality assurance mechanisms we analyze.

2 Uber, telematics, and ratings

We analyze the Uber market in Chicago during the first half of 2017. The definition used by Uber for the Chicago market includes regions more than 100 km away from downtown Chicago, so we limit our analysis to a smaller region that excludes most suburbs of Chicago but which is large enough to include Midway and O'Hare, the main airports in the region. Figure 2.1 shows a map of the region of analysis.

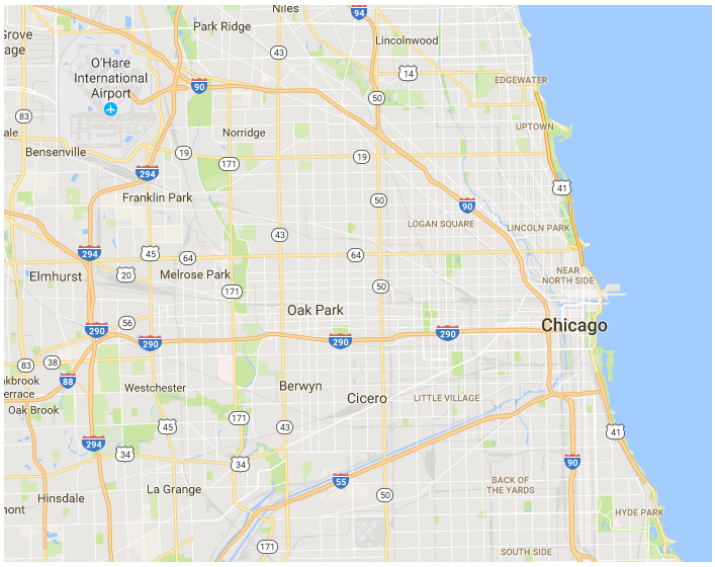


Figure 2.1: Region of analysis (from Google Maps)

We focus on two of Uber's products. The first one is UberX, Uber's main ride-hailing option and its largest product as measured by number of rides. Drivers typically own their cars, and entry requirements are far less than a traditional professional license: 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. 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 7,849,896

UberX trips and 164,288 UberTaxi trips after filtering out trips in which any driving metrics, which we describe below, are missing.

Uber collects telematics data measured through the GPS on drivers' smartphones every two seconds while the Uber app is open. The raw data includes location, speed, and acceleration.[4] Our analysis focuses on six trip-level metrics, whose distributions are shown in Figure 2.2. Our *accelerations* metric is the fraction of acceleration events where acceleration went above 2 m/s^2.[5] Our *brakes* metric is defined similarly.

Uber developed two classifiers based on accelerometer data that tell 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). Based on these classifiers, we define the *handling* and *mounted* metrics as the average of these classifiers over the two middle time quartiles of a trip (i.e., we do not take into account what happens at the beginning and at the end of a trip). We focus on the middle quartiles to make our measures easier to compare across UberX and UberTaxi trips, since Uber 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.

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

One potential concern is that speed might be highly correlated with trip time. Additionally, as pointed out by Liu et al. (2018), riders may have preferences over the route. In order to account for this, we define two additional measures, which we use in part of our analysis, for *distance* and *duration*: the log of the ratio of actual distance to estimated distance, and the log of the ratio of actual duration to estimated duration. A major limitation of these two metrics is that Uber does not compute estimated distances and durations for UberTaxi trips, so we cannot use

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

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

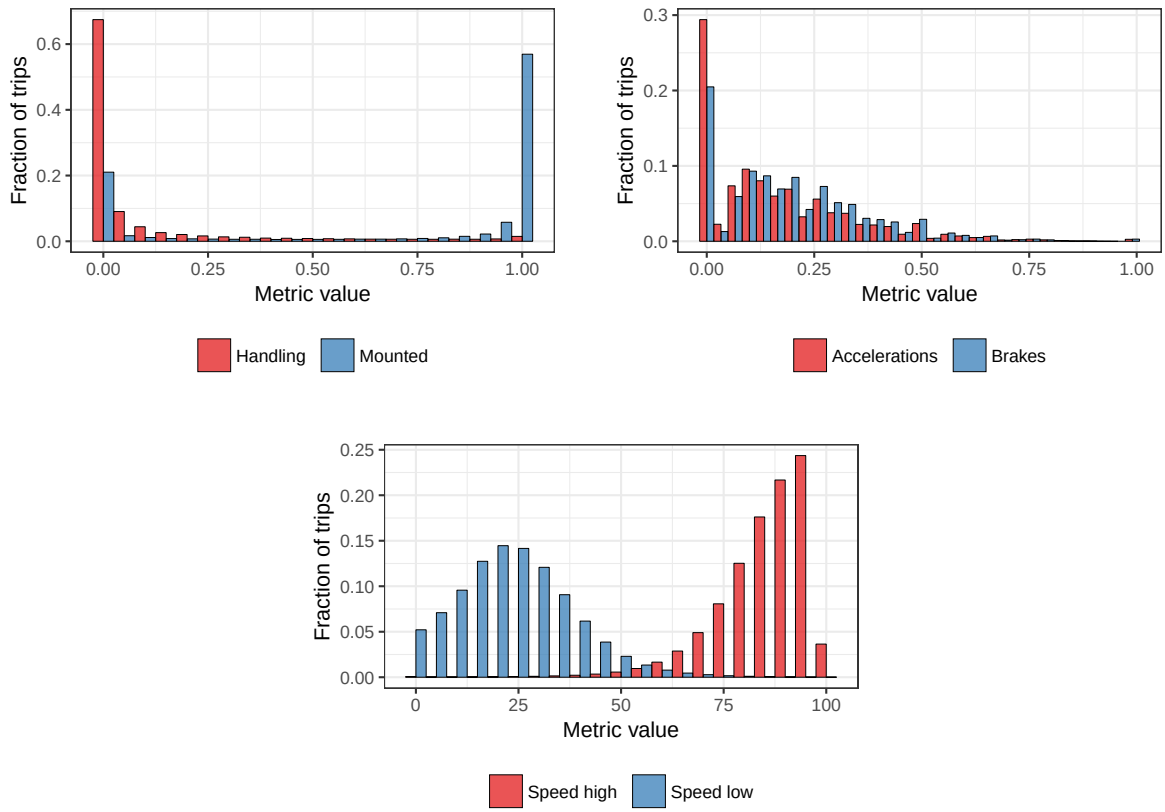


Figure 2.2: Histograms of the safety metrics for UberX trips.

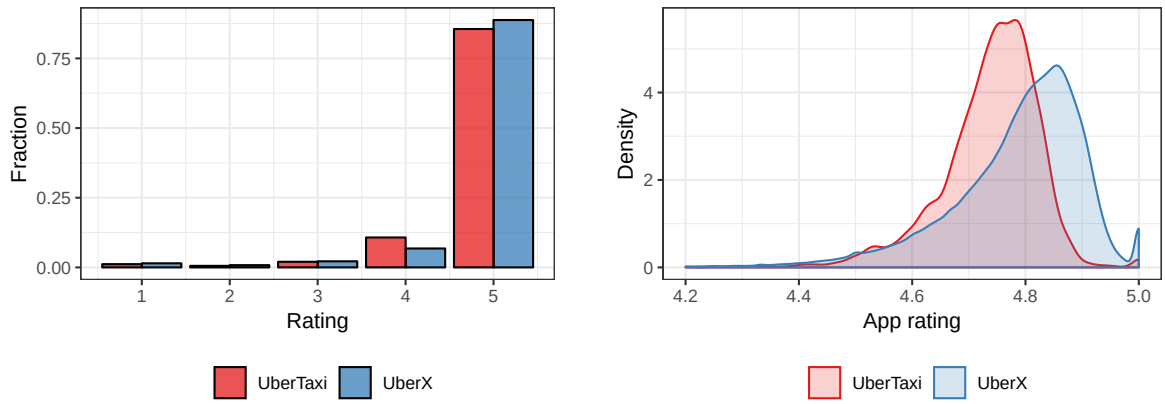
these for our main purpose of comparing UberX and UberTaxi.

There are 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 lasso model to predict a trip's rating based on all these variables and their squares. We chose the variables that dropped out last as the penalty increased. The details of this procedure are in Appendix A. Our main results, however, do not change when we choose alternative metrics.

Uber uses a rating system in which each passenger can give a one to five star rating to the driver after each trip. 29.8% of the trips in our sample were rated.[6] Figure 2.3a shows the distribution of ratings. The majority of trips receive five stars, which means that trips typically get five stars unless there was some problem. We also see that UberTaxi trips tends to get a larger fraction of 4 star ratings. Uber uses

[6]This fraction is typical for the new app interface introduced in late 2016. Previous interfaces showed a rating screen after every trip, which resulted in around 60% of trips being rated.

the *app rating*, the average of the last 500 trips, as its main measure of driver quality; drivers can see their app rating in the app, and it is shown to passengers upon being matched to a driver before pickup. Figure 2.3b shows its distribution. Uber stops giving 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. The full details of the process are described in Section 6.



(a) Trip rating

(b) App rating

Figure 2.3: Distribution of ratings for UberX and UberTaxi

3 Riders' preferences over driving behavior

In this section we explore what kind of driving behavior is preferred by riders. In order to do so we use different methodologies to predict the rating a rider will give to a trip based on our driving metrics. Simple regressions, or even nonparametric regressions, of rating on driving metrics face a variety of problems. The main issue is that different kinds of trips (at different times of the day, with different origin and destination, downtown or on a highway) lead to different kinds of driving behavior, such as slower trips downtown and in rush hour, some of which might be intrinsically more satisfying for riders. Second, there are unobserved characteristics that have an effect on satisfaction, such as the driver's personality, which might be correlated with driving behavior. We address both of these issues by controlling for type of trip and driver. Our goal is to estimate how *changes* in driving metrics lead to changes in consumer satisfaction.

To control for the type of trip, we divide our sample into 128 rectangles by origin

coordinates. In order to size the 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. We follow an analogous process to divide the sample into 128 groups by destination coordinates. Finally, we also divide our sample into 15 hour of the week intervals.[7] This results in 245,760 buckets as the cartesian product of all origin, destination, and hour of the week divisions. We inevitably end up with some groups with zero or one trips, but Figure 3.1 shows that the majority of trips in our sample are in groups with more than five trips. Most of our regressions include fixed effects based on these groups, which we call *trip characteristics* fixed effects.

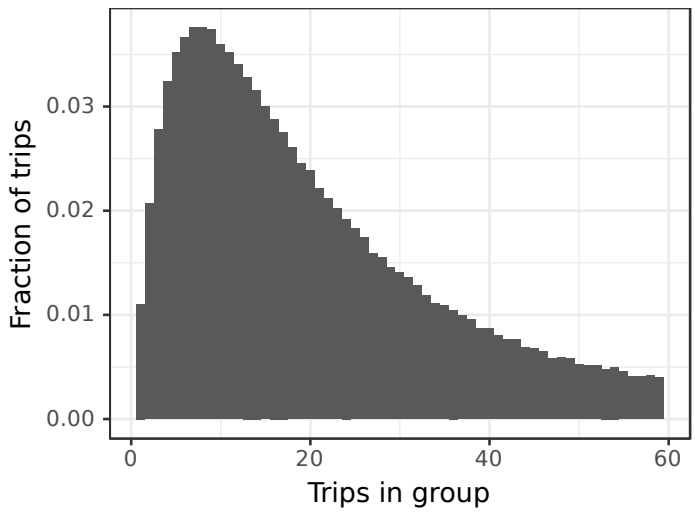


Figure 3.1: Fraction of trips in a group with each number of trips.

We start with simple linear regressions to show how metrics relate to ratings. Our main estimating equation takes 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 ratings, and m_i is a vector of driving metrics. X_i is a set of fixed effect dummies. Table 3.1 shows estimates of this equation for our sample of UberX trips, including the six metrics we focus on. In columns (3), (6), and (9) we also include routing metrics.

[7] The 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.

All driving metrics are normalized, so coefficients measure how ratings change if metrics change by one standard deviation.

Columns (1)-(6) show results for rating and for a dummy for the rating being 5, with consistent results. Riders dislike phone handling and hard brakes. Some specifications show a weak preference for cell phones being mounted. Focusing on regressions with driver fixed effects, we see that riders dislike hard accelerations. These four metrics reflect preferences for safer trips; these trips may also provide a more comfortable ride. Riders' preferences for speed metrics are more nuanced. They prefer higher low speeds and lower high speeds; in other words, riders prefer it when drivers stay at an intermediate speed throughout the trip (recalling that all speeds are expressed as a percentile of typical speeds on the route). This reflects a compromise between a safe, smooth ride and getting quickly to the destination.

In columns (3) and (6) we also include routing metrics. Riders have strong preferences for shorter trips, both in terms of distance and duration. Including these variables somewhat changes the rest of the coefficients, but the main patterns remain the same. The dependent variable in columns (7)-(9) is a dummy for whether trips were rated. We see that coefficients tend to have opposite signs relative to previous columns (except for speed high and handling, the latter of which does not have a significant coefficient). This suggests that riders are more likely to rate trips when they are unsatisfied. This type of bias is one motivation for our approach; 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.

Table 3.1: Rating response to driving metrics

	Dependent variable:								
	Rating			Rating is 5			Rated		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mounted	0.0100***	0.0028*	0.0016	0.0047***	0.0014**	0.0009	−0.0001	−0.0013***	−0.0010**
	(0.0009)	(0.0015)	(0.0015)	(0.0004)	(0.0007)	(0.0007)	(0.0002)	(0.0005)	(0.0005)
Handling	−0.0039***	−0.0082***	−0.0056***	−0.0012***	−0.0026***	−0.0019***	0.0005**	−0.0002	−0.0002
	(0.0008)	(0.0010)	(0.0010)	(0.0004)	(0.0004)	(0.0005)	(0.0002)	(0.0003)	(0.0003)
Brakes	−0.0093***	−0.0035***	−0.0040***	−0.0042***	−0.0015***	−0.0017***	0.0012***	0.0017***	0.0015***
	(0.0007)	(0.0007)	(0.0007)	(0.0003)	(0.0003)	(0.0003)	(0.0002)	(0.0002)	(0.0002)
Accelerations	0.0016**	−0.0040***	−0.0044***	0.0009***	−0.0017***	−0.0019***	0.0018***	0.0018***	0.0016***
	(0.0007)	(0.0007)	(0.0007)	(0.0003)	(0.0003)	(0.0004)	(0.0002)	(0.0002)	(0.0002)
Speed low	0.0133***	0.0087***	0.0044***	0.0058***	0.0036***	0.0021***	−0.0040***	−0.0040***	−0.0030***
	(0.0006)	(0.0006)	(0.0007)	(0.0003)	(0.0003)	(0.0003)	(0.0002)	(0.0002)	(0.0002)
Speed high	0.0026***	−0.0029***	−0.0061***	0.0002	−0.0024***	−0.0035***	−0.0026***	−0.0030***	−0.0027***
	(0.0007)	(0.0007)	(0.0007)	(0.0003)	(0.0003)	(0.0004)	(0.0002)	(0.0002)	(0.0002)
Duration			−0.0233***			−0.0088***			0.0054***
			(0.0008)			(0.0004)			(0.0002)
Distance			−0.0105***			−0.0033***			0.0005**
			(0.0008)			(0.0004)			(0.0002)
Trip characteristics FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Driver FE		✓	✓		✓	✓		✓	✓
Observations	2,296,362	2,296,362	2,132,158	2,296,362	2,296,362	2,132,158	7,685,608	7,685,608	7,387,288

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

All safety metrics are normalized to mean zero and variance one.

We now explore how preferences change with the 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.[8] Table 3.2 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 at work on time.[9]

Table 3.2: Heterogeneous ratings response to driving metrics. The three columns report coefficients for one single regression.

	Dependent variable: Rating		
	Interaction of covariate with:		
	Off-peak	AM rush	PM rush
	(1)	(2)	(3)
Mounted	0.0020	0.0035	0.0050**
	(0.0015)	(0.0022)	(0.0020)
Handling	−0.0083***	−0.0099***	−0.0069***
	(0.0011)	(0.0020)	(0.0017)
Brakes	−0.0034***	−0.0042**	−0.0040***
	(0.0008)	(0.0017)	(0.0015)
Accelerations	−0.0042***	−0.0038**	−0.0034**
	(0.0008)	(0.0017)	(0.0015)
Speed low	0.0079***	0.0123***	0.0088***
	(0.0008)	(0.0017)	(0.0014)
Speed high	−0.0031***	0.0005	−0.0048***
	(0.0008)	(0.0018)	(0.0017)
Note:			*p<0.1; **p<0.05; ***p<0.01
			Obervations: 2,296,362

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

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

where $d(i)$ indexes the driver and $c(i)$ indexes the trip characteristics group.

The term $s(m_i; \theta)$ is a flexible function of driving metrics. It is the sum of two high order polynomials. The first one is the interaction of a quadratic function of

[8]Morning rush hour trips are those starting during weekdays between 6 am and 10 am. Afternoon rush hour trips start between 5 pm and 9 pm on weekdays.

[9]In some 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.

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.[10] To avoid overfit—this specification for $s(m_i;\theta)$ has 232 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). 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.

Figure 3.2 shows the functional form of our estimated $s(m_{ijkt};\hat{\theta})$. In each subfigure we vary two of the metrics. At each point in these plots we compute the average value of $s(m_{ijkt};\hat{\theta})$ over the distribution of the metrics we are not varying.[11] The color, as well as the contours, represent the value of $s(m_{ijkt};\hat{\theta})$. More intense colors represent areas with a high density of observations, and white represents areas with no observations.

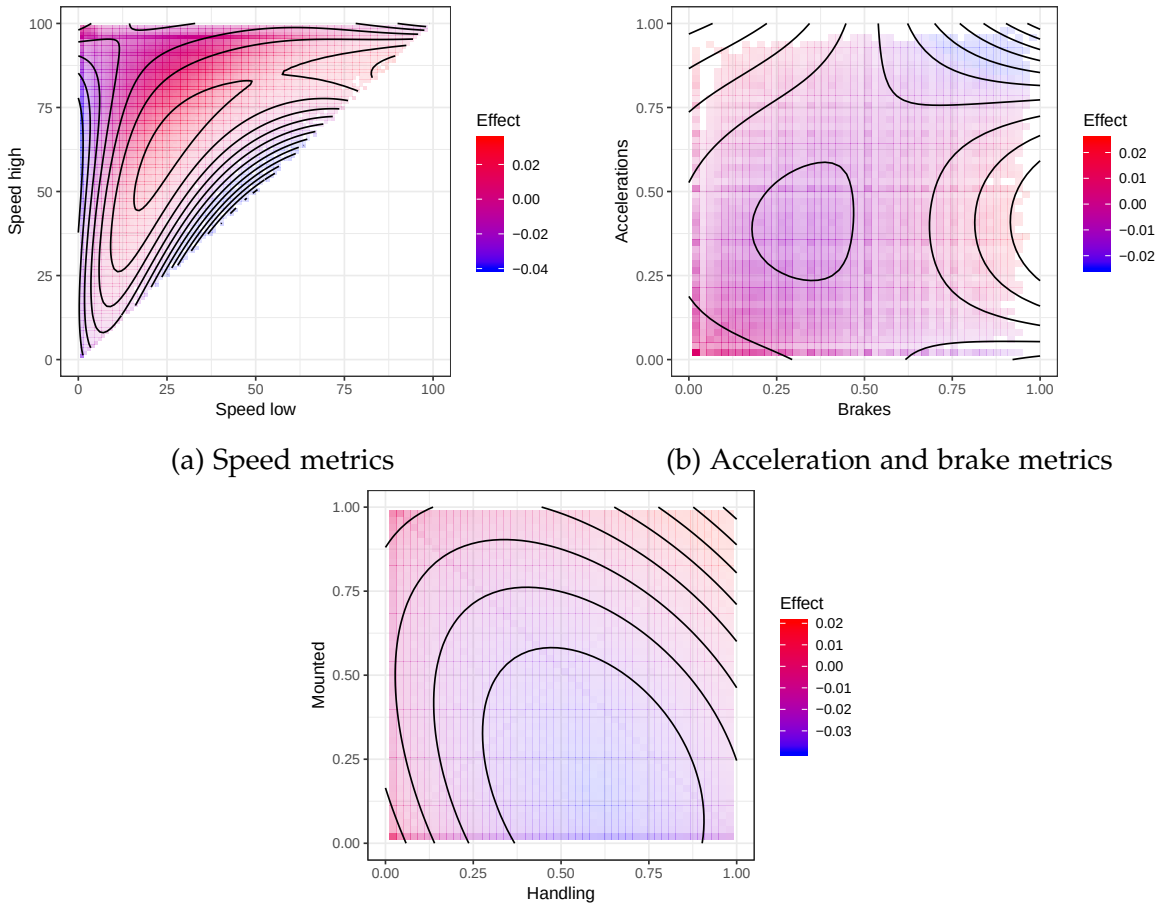
Figure 3.2a shows that riders have a bliss point for speed around (30,60). This confirms that riders prefer trips that are neither too fast nor too slow. The region with highest density is around (25,85), above and to the left of the bliss point, which explains the negative coefficient for speed high and the positive coefficient for speed low in Table 3.1.

Most observations (98% of trips) in Figure 3.2b are below (0.75, 0.75). There is a clear pattern in that region: riders prefer few hard accelerations and hard brakes. The somewhat unexpected patterns at the upper left and lower right corners are driven by a very small number of observations. From Figure 3.2c, we can see that riders prefer drivers to mount their phone and not to handle it. The behavior at the upper right corner is somewhat unexpected, but it is also driven by very few observations—only 1.5% of trips are above and to the right of (0.5,0.5)—reflecting problems in the classifiers that generate the data: a driver should not be able to handle a mounted phone.

Throughout the rest of our analysis we compare drivers' behavior under differ-

[10] 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 2025 terms, and it would not be feasible to estimate it given our sample size. We thus separate the function additively into two terms, one for phone usage and the second one for driving.

[11] The curse of dimensionality makes it hard to evaluate this average over a smooth distribution, so we bin observations into the cartesian product of the quartiles of all variables we are not varying. We then estimate $s(m_{ijkt};\hat{\theta})$ at the midpoint of all observations within each bin. Each evaluation gets a weight proportional to the number of observations in the bin.



(a) Speed metrics

(b) Acceleration and brake metrics

(c) Cell phone use metrics

Figure 3.2: Effect of driving metrics on scores. Colors, as well as contours, represent the magnitude of the effect. Stronger colors represent regions with a higher density of observations.

ent circumstances. Although we can use individual driving metrics as outcomes, our analysis shows that riders' preferences over these metrics are nonlinear. In order to summarize differences in quality across different settings, we define a driving score $s^F_{ijkt} = s(m_{ijkt}; \hat{\theta})$ which we call our *full score* or *score F*—since it is computed from our full model. This score is measured in units of stars. More precisely, if we compare two groups of rides, we will evaluate the difference by comparing the difference in the average score between the two groups, and interpret as the difference in the average quality (as perceived by the riders in our training set, who are all UberX riders).

In using this type of score as an outcome variable, we follow Athey et al. (2016). They propose constructing a *surrogate index*, an outcome variable in settings where the true outcome of interest may be missing in the relevant timeframe—because it is

a long term outcome, for instance, or because it is not systematically available in an experiment. 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, and the outcome that is sometimes missing is the star rating (taxi trips are rated substantially less often than UberX trips, and the interpretation of their ratings may be different). If the surrogates capture the effect of a "treatment" on the final outcome, and if the relationship between the surrogates and the final outcome does not depend on the treatment, then 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 available and observed in the relevant timeframe).[12]

Given that we are focusing our attention on the impact of differences in driver behavior on rider satisfaction, it is natural to focus on differences in trip ratings that are captured by the driving metrics. Thus, the conditions for the use of a surrogate index are satisfied in this application, and we will proceed to use our constructed score to measure differences between UberX and taxis. A caveat is that it is possible that the preferences of the full set of UberX riders are different than those who use taxis; if so, we should be careful to interpret our results as measuring quality as perceived by UberX riders.

In some specifications we want to distinguish the nonmonotonic preferences for speed from the monotonic preferences for the remaining metrics. We thus define a *speed score* or *score S*, denoted by s^S_{ijkt}, which we compute from a model of the form (2), but where $s(m_{ijkt};\theta)$ is the interaction of a quartic function of speed low and a quartic function of speed high—i.e., it only captures preferences for speed. Finally, we define a *no speed score* or *score NS*, denoted by s^N_{ijkt}, where $s(m_{ijkt};\theta)$ is the interaction of a quadratic for handling, a quadratic for mounting, a cubic for acceleration, and a cubic for brakes. This score only captures preferences for metrics other than speed. We center our three scores so that they have mean zero. Figure 3.3 shows the density of these scores. Our full score has the largest variance, since it captures the most variation. We can also see that speed and non-speed metrics

[12]In an experiment where both surrogates and the final outcome are observed for each unit, Athey et al. (2016) 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. One exercise in this paper analyzes a randomized experiment with a smaller dataset; in that case, efficiency is gained by using a larger dataset to estimate the relationship between surrogates and the final outcome.

are roughly equally important, since they have similar variances.

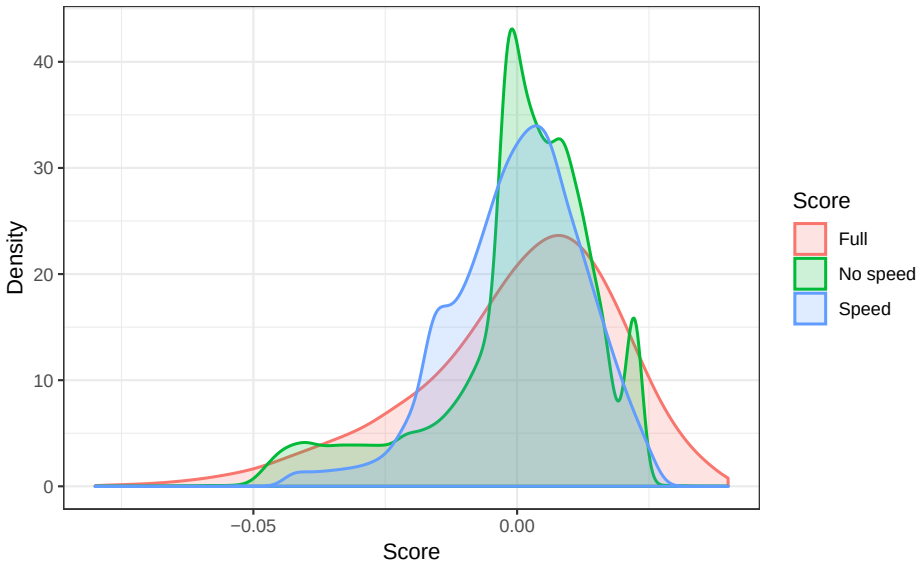


Figure 3.3: Kernel density for the distribution of scores.

One might be worried that our scores F and S are capturing preferences for getting to the destination in time, and not only preferences for driving.[13] In order to tackle this concern, we also compute additional scores based on the specification from equation (2) in which we also include a flexible function of the routing metrics. In Appendix D we show that the new scores that arise from this procedure behave very similarly to our main metrics. We are not able to use these new scores in the rest of our paper since our routing metrics are not available for UberTaxi trips.

4 UberX versus UberTaxi

In this section we compare the driving behavior of UberX and UberTaxi trips. Our main question will be the following: 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 want to estimate the average treatment effect of being an UberX trip, for the distribution of trips that were taken on UberTaxi.[14] The key assumption we rely on

[13]Ideally, we would like to create an additional score that also captures routing behavior. However, since our routing metrics are not available for UberTaxi, we cannot make the main comparison between UberX and UberTaxi we are interested in.

[14]We focus on the UberTaxi population because for most UberTaxi trips, we can find similar UberX trips, but the reverse is not true given the much larger share of UberX trips in our data.

for this estimation is unconfoundedness when controlling for trip characteristics. That is, controlling for trip origin location, destination location, and time of day, there are no further unobserved characteristics of trips that would lead the trips to have different telematics.

We use two different methodologies with very similar results. First, we match UberX and UberTaxi trips according to their origin and destination coordinates and the hour of the week at which they started. We do so by matching every UberTaxi trip to its 10 nearest UberX neighbors, using a Euclidean metric in which a half an hour difference is equivalent to a difference in origin or destination of one kilometer.[15] We then estimate the ATE as

$$\hat{\tau}^{match} = \frac{1}{N} \sum_{i \in I_{\text{taxi}}} \left(y_i - \frac{1}{10} \sum_{j \in C_i} y_j \right), \tag{3}$$

where I_{taxi} is the set of UberTaxi trips, and C_i denotes the set of nearest neighbors of trip i. 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

$$y_i = \tau x_j + \beta X_i + \epsilon_i, \tag{4}$$

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 estimate for the treatment effect is the OLS estimate for τ. In general, this estimator is not consistent for the ATE, but it converges to a weighted average of treatment effects. We will see, however, that both methodologies result in almost identical results.

Table 4.1 shows the results of running these specifications for different dependent variables. The matching estimator has almost identical results to the OLS estimator with trip characteristics fixed effects. Our estimates change somewhat if we also include rider fixed effects, but not enough for the interpretation of the coefficients to change.

[15] Our results do not change if we use similar metrics where half an hour is equivalent to more or less than one kilometer.

Table 4.1: Comparison of driving behavior between UberX and UberTaxi trips.

	Dependent variable:								
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Panel A: Matching estimator								
UberX	0.9552***	0.0697***	−0.0475***	−0.2453***	−0.0269***	−0.2225***	0.0035***	0.0031***	0.0015***
	(0.0245)	(0.0129)	(0.0116)	(0.0140)	(0.0083)	(0.0072)	(0.0003)	(0.0001)	(0.0002)
Observations	164,288	164,288	164,288	164,288	164,288	164,288	164,288	164,288	164,288
	Panel B: Trip characteristics fixed effects								
UberX	0.9524***	0.0705***	−0.0356***	−0.2359***	−0.0378***	−0.2253***	0.0033***	0.0030***	0.0014***
	(0.0248)	(0.0137)	(0.0118)	(0.0143)	(0.0084)	(0.0075)	(0.0003)	(0.0001)	(0.0002)
Observations	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896
	Panel C: Trip characteristics and rider fixed effects								
UberX	0.9475***	0.0571***	−0.0520***	−0.2399***	−0.0074	−0.2090***	0.0037***	0.0030***	0.0017***
	(0.0261)	(0.0148)	(0.0128)	(0.0152)	(0.0096)	(0.0090)	(0.0003)	(0.0001)	(0.0002)
Observations	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896	7,849,896

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

All safety metrics are normalized to mean zero and variance one.

UberX drivers are more likely to mount their cell phones. The difference is almost one standard deviation. This is not surprising since Uber has led campaigns to ensure drivers mount their phones, sometimes giving away phone mounts. Drivers are, however, also more likely to handle their phones, despite the fact that we do not take into account handling at the beginning or end of a trip. This is also not too surprising, since they rely much 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.

We also see that 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' preferences.

Columns (7)-(9) report estimates of the ATE on driving scores. We see that UberX trips are better in terms of all three scores. The difference seems small when measured in rating stars, but it accounts for 0.16 standard deviations of score F, 0.22 standard deviations of score S, and 0.10 standard deviations of score NS. We will show in Section 5 that a large fraction of the variation in ratings is outside drivers' control, so these numbers are not small relative to what a driver can do to change ratings.

The main takeaway is that UberX trips look better than UberTaxi trips in terms of driving metrics. In the rest of this paper we will break down this difference and will explore the causal mechanisms that lead to it. We start by decomposing the effect across different types of trips. Since trips during rush hour are more time sensitive than other trips, as suggested in Table 3.2, it could be the case that riders hurry their drivers. We would like to see whether drivers respond to this, and whether this effect is stronger for UberX or UberTaxi. In order to do so, we define a set of dummies z_{ijkt} that represent whether a trip took place during the morning or afternoon rush hour and interact them with the UberX dummy.

Table 4.2: Heterogeneity in effect of UberX using a matching estimator.

	Dependent variable:								
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
UberX × off-peak	0.9491***	0.0567***	−0.0637***	−0.2749***	−0.0648***	−0.2725***	0.0039***	0.0032***	0.0018***
	(0.0243)	(0.0143)	(0.0124)	(0.0147)	(0.0089)	(0.0081)	(0.0003)	(0.0001)	(0.0002)
UberX × AM rush	0.9909***	0.0889***	−0.0188	−0.2111***	0.0190*	−0.1619***	0.0026***	0.0029***	0.0007**
	(0.0320)	(0.0146)	(0.0151)	(0.0176)	(0.0102)	(0.0093)	(0.0003)	(0.0001)	(0.0003)
UberX × PM rush	0.9200***	0.0804***	−0.0410***	−0.2031***	0.0216*	−0.1576***	0.0035***	0.0031***	0.0018***
	(0.0319)	(0.0160)	(0.0142)	(0.0179)	(0.0128)	(0.0091)	(0.0003)	(0.0001)	(0.0003)
Observations	164,288	164,288	164,288	164,288	164,288	164,288	164,288	164,288	164,288

Note: *p<0.1; **p<0.05; ***p<0.01

All safety metrics are normalized to mean zero and variance one.

Table 4.2 shows the result from this specification. UberX drivers handle their phones more relative to UberTaxi drivers during rush hour than during off-peak hours. The gap in hard accelerations and brakes shrinks during rush hour. Further, while during off-peak hours UberX drivers tend to be slower than UberTaxi drivers according to the speed low metric, during rush hour UberX drivers tend to be faster than UberTaxi drivers. This is consistent with UberX drivers paying more attention to riders who want to get to their destination on time by driving faster—therefore braking and accelerating more—and by handling the phone more to find better routes to avoid traffic. This results in a net decrease in the full score. However, as shown in Table 3.2, riders' preferences are different during the morning rush hour, so drivers might just be responding to changes in preferences that are not captured by our full score. 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.[16]

This behavior is consistent with UberX drivers paying more attention to riders who want to get to their destination on time by driving faster—therefore braking and accelerating more—and by handling the phone more to find better routes to avoid traffic. All these adjustments result in a net decrease in the full score. As shown in Table 3.2, riders' preferences are different during the morning rush hour, so drivers might just be responding to changes in preferences that are not captured by our full score. 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.[17]

5 Decomposition of driving behavior

We now start a more detailed analysis of the determinants of driving style. We focus our analysis on UberX trips given that we only have a small number of UberTaxi

[16]We also ran similar specifications with dummies for trips that end in airports. We found the unintuitive result that the interaction between this dummy and the UberX dummy has a negative coefficient for speed but a positive one for brakes and accelerations. We think this strange pattern arises from the fact that airport trips in Chicago are highway trips and are thus significantly different from the typical trip in our sample.

[17]We also ran similar specifications with dummies for trips that end in airports. We found the unintuitive result that the interaction between this dummy and the UberX dummy has a negative coefficient for speed but a positive one for brakes and accelerations. We think this strange pattern arises from the fact that airport trips in Chicago are highway trips and are thus significantly different from the typical trip in our sample.

observations. Broadly speaking, we would like to see how behavior varies by driver, by passenger, by origin and destination, and by time of the day.

We would first like to see whether riders have an impact on driving behavior. 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. In order to measure this, let $\bar{y}_i^{LO}$ be the average score of all trips taken by the rider who took trip i, whom we denote by $r(i)$, leaving out the current trip. We run regressions of the following form:

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \beta \bar{y}_i^{LO} + \epsilon_i \tag{5}$$

where the first two terms denote driver and trip characteristics fixed effects. The driver and trip fixed effects account for the fact that riders will tend to have similar trips (e.g. trips from their home downtown), and that drivers may focus their driving in certain areas. In some other specifications we also interact $\bar{y}_i^{LO}$ with dummies for whether a trip starts or ends in an airport and whether it took place during the morning or afternoon rush hour. One problem with this specification is that many riders only have a small number of trips in our database; we address this by restricting our sample to passengers with 20 or more trips.

Table 5.1 shows results for this exercise. We see a negative but small effect on mounting, and we see no evidence of a rider effect on handling. On the other hand, we see a strong and positive effect on brakes and accelerations, and especially on speed metrics. This means that there is a strong correlation of metrics within riders. We interpret this finding as establishing that riders influence driver behavior. The estimates in columns (7)-(9) are consistent with these findings, since the strong correlations in metrics should be reflected with correlations in scores.

In Table 5.2 we break down this correlation across different kinds of trips.[18] We see that it is especially strong for trips during the morning rush hour. This goes in line with riders being especially time sensitive when going to work and insisting on fast driving to get there early. In addition, riders who use UberX for the morning rush hour may do so consistently, and so their preferences are likely to be consistent as well. It is also possible that this coefficient is picking 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 a strong correlation among scores within a rider.

[18]We also ran similar specifications with dummies for trips that end in airports. We found that metrics and scores are less correlated for airport trips than for the rest of the trips, but we believe this is more closely related to the unusual road patterns from downtown Chicago to airports.

Table 5.1: Response to rider preferences. The average rider score does not include the current observation, and only riders with more than 20 trips are included in the sample.

	Dependent variable:								
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
LO mean by rider	-0.0049^{***}	-0.0046^{**}	0.1450^{***}	0.1168^{***}	0.3360^{***}	0.3564^{***}	0.1739^{***}	0.2805^{***}	0.0451^{***}
	(0.0015)	(0.0023)	(0.0027)	(0.0026)	(0.0024)	(0.0031)	(0.0026)	(0.0028)	(0.0024)
Trip Characteristics FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Driver FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.

Table 5.2: Heterogeneity in response to rider preferences. The average rider score does not include the current observation, and only riders with more than 20 trips are included in the sample.

	Dependent variable:								
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
LO mean by rider × Off-peak	-0.0013	-0.0023	0.1599^{***}	0.1266^{***}	0.3011^{***}	0.3265^{***}	0.1436^{***}	0.2334^{***}	0.0444^{***}
	(0.0018)	(0.0028)	(0.0033)	(0.0032)	(0.0029)	(0.0034)	(0.0032)	(0.0034)	(0.0030)
LO mean by rider × AM rush	-0.0202^{***}	-0.0224^{***}	0.1219^{***}	0.1101^{***}	0.5048^{***}	0.5226^{***}	0.3134^{***}	0.4800^{***}	0.0441^{***}
	(0.0037)	(0.0056)	(0.0069)	(0.0067)	(0.0059)	(0.0082)	(0.0068)	(0.0072)	(0.0061)
LO mean by rider × PM rush	-0.0051	0.0003	0.1105^{***}	0.0880^{***}	0.3126^{***}	0.3132^{***}	0.1536^{***}	0.2486^{***}	0.0482^{***}
	(0.0033)	(0.0051)	(0.0056)	(0.0055)	(0.0051)	(0.0061)	(0.0055)	(0.0060)	(0.0053)
Trip Characteristics FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Driver FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222	3,648,222

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.

Our next results decompose variance across different sources of variation. In order to do so, we run a model of the form

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \gamma_{r(i)} + \epsilon_i, \tag{6}$$

which decomposes y into a rider effect, a driver effect, a trip characteristics effect, and a residual.

One potential problem with this specification is that for drivers with a small number of trips $\mu_{d(i)}$ is going to take most of the variation. Similarly, for riders with a small number of trips $\nu_{c(i)}$ is going to take most of the variation. In order to avoid this we limit our sample to drivers and riders that have more than 20 trips.[19]

Table 5.3 shows the variance of each one of the four terms in equation (6). We see that for most variables the driver is responsible for a significant share of the variation. This is especially true for mounted and handling. Perhaps surprisingly, the variable for which driver effects are least important is rating. The rider is not responsible for much of the variation, with one notable exception: for rating, the rider is the most important source of variation after the residual. This highlights the limitations of the rating itself as opposed to the scores based on telematics we use to evaluate ride quality; our scores are applied systematically to the telematics from each ride, while the ratings have rider-specific noise that is unrelated to the driver's performance. We also see that trip characteristics are especially important for speed.[20]

Table 5.3: Variance decomposition (more than 20 trips)

	Driver	Rider	Trip characteristics	Residual
Mounted	0.830	0.005	0.003	0.164
Handling	0.591	0.013	0.008	0.386
Brakes	0.220	0.024	0.038	0.725
Accelerations	0.332	0.021	0.026	0.628
Speed low	0.099	0.045	0.083	0.787
Speed high	0.136	0.047	0.116	0.651
Rating	0.046	0.301	0.036	0.627
Score F	0.355	0.028	0.036	0.579
Score NS	0.511	0.017	0.018	0.454

In Appendix E we check whether the variance decomposition of score F looks

[19]We filter by riders, and then by drivers. After filtering by drivers, a few riders end up having fewer than 20 trips.

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

different by trip type (such as rush hour trips), and we find that the results look very similar across different trip types.

6 Incentives and behavioral nudges

Uber uses a variety of incentives and behavioral nudges intended to promote safety and quality on the platform. In this section we want to see whether these incentives affect drivers' behavior and whether the behavioral nudges provided by these messages promote better driving behavior.

6.1 Ratings, notifications, and deactivation

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 a certain threshold. In contrast, while riders can rate UberTaxi rides and UberTaxi drivers can access the ratings in the app, UberTaxi rides are likely to be a small share of a taxi driver's business, and there are no explicit incentives tied to the ratings. Thus, we expect the ratings to be both less salient and less important to UberTaxi drivers.

Table 6.1 summarizes the deactivation process, which follows each one of the steps in each row. In order to move to the next state, a driver has to satisfy both conditions on average ratings for the last 50 and 500 trips and the condition for the number of trips. 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.

Table 6.1: 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

Table 6.2 gives a sense of how many drivers are in each one of the ranges for the rating of the last 500 trips. This is the rating that drivers can observe in their app and which is shown to users when they are matched to a driver. We will call it the

app rating. 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.

Table 6.2: Number of trips during which driver ratings satisfy each condition

	Number	Fraction
Total	7,685,605	
lifetime trips >500	4,923,415	0.641
Rating <4.6	565,779	0.074
Rating <4.5	221,919	0.029
Rating <4.4	99,765	0.013

To get a sense of how strong these incentives are, Figure 6.1 shows the percentage of drivers that are at risk of falling below each one of these thresholds. It shows how many 3 star rated trips the driver would have to complete in order to fall below each threshold. We see that only a very small number of drivers are likely to eventually reach the 4.4 threshold for deactivation. On the other hand, a somewhat more important fraction of drivers are close and even below the 4.6 threshold for the first notification. This means that if this deactivation process has an effect on driving behavior it is most likely through behavioral nudges instead of through actual incentives a fully rational agent would react to.

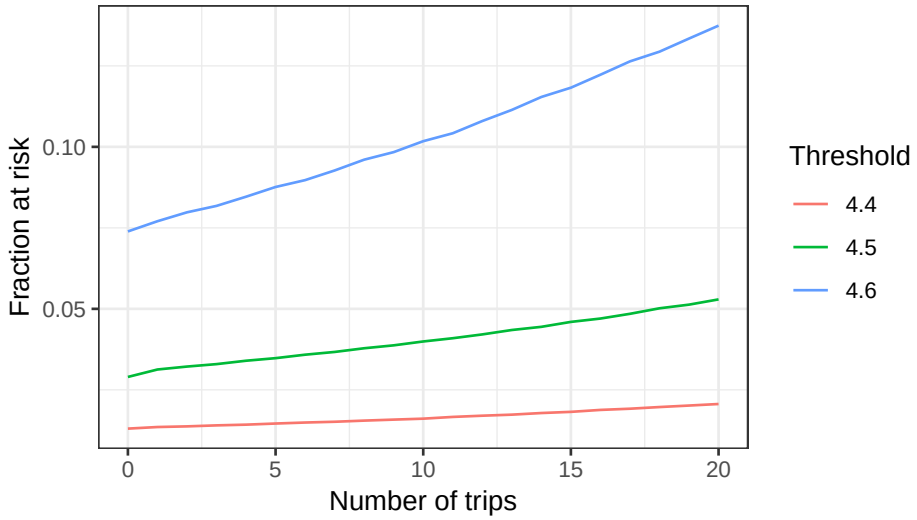


Figure 6.1: Fraction of drivers whose app rating would fall below a certain threshold if the next N consecutive trips received a 3-star rating, where N is displayed on the x-axis. The values corresponding to $N = 0$ represent the fraction whose app rating currently falls below the threshold.

Table 6.3 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 4.4, close to which there are not that many drivers.

Table 6.3: Number of threshold crossings

	From above		From below	
Threshold	Crossings	Unique drivers	Crossings	Unique drivers
4.6	7,405	4,742	7,135	4,495
4.5	4,913	3,331	4,343	2,874
4.4	3,613	2,690	2,808	2,046

We now consider a simple way to measure to what extent these incentives drive the results of Section 4. We estimate whether the differences between UberX and UberTaxi are the same for drivers with app rating above 4.6 (for whom there are no explicit incentives and behavioral nudges) and those with app rating below 4.6. We do so by matching trips as for Table 4.1, but we constrain matches to be only within trips with app rating above 4.6 and within those below it. We compute an overall effect of UberX, and additionally, an interaction of UberX and being below 4.6.

Table 6.4 shows the results. The UberX effect is stronger for low rated drivers. This result comes with two caveats: it goes the other way around for mounted and for the speed score (although the interaction coefficient for speed is very small), and not many of these results are significant. We also find similar results for thresholds different from 4.6. These findings support our claim that part of the difference between the way UberX and UberTaxi drivers behave can be attributed to the incentives set in place by Uber: the differences are largest when drivers' ratings are low and the incentives are strongest.

In order to further explore the role of incentives, we examine how previous ratings affect the behavior in new trips. We thus run regressions of the form

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \beta r_i^{\text{app}} + \epsilon_i, \tag{7}$$

where r_i^{app} is the main rating shown to rider $r(i)$ in the app at the beginning of trip i. The first two terms represent trip characteristics and, importantly, driver fixed effects. We are thus trying to measure how changes in ratings within a single driver are related to their driving behavior.

Table 6.4: UberX treatment effect by app rating, matching estimator

	Dependent variable:								
	Mounted	Handling	Brakes	Accels	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
UberX	0.0729***	0.9552***	−0.0327***	−0.2346***	−0.0359***	−0.2144***	0.0025***	0.0023***	0.0011***
	(0.0134)	(0.0257)	(0.0121)	(0.0146)	(0.0086)	(0.0075)	(0.0003)	(0.0001)	(0.0002)
UberX * Rating < 4.6	−0.0255	0.0220	−0.0740*	−0.0942*	−0.0344	−0.0475	0.0007	−0.0001	0.0009
	(0.0446)	(0.0726)	(0.0409)	(0.0496)	(0.0299)	(0.0302)	(0.0009)	(0.0003)	(0.0008)
Observations	164,031	164,031	164,031	164,031	164,031	164,031	164,031	164,031	164,031

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.

Table 6.5: Response to rating

	Dependent variable:									
	Rating	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
App rating	−0.3043***	0.0078	−0.0023	−0.0076**	−0.0395***	0.0424**	0.0492***	0.0005*	0.0001	0.0004
	(0.0149)	(0.0111)	(0.0127)	(0.0032)	(0.0027)	(0.0175)	(0.0093)	(0.0003)	(0.0001)	(0.0002)
# of trips quadratic	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Driver FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Trip Characteristics FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	2,286,796	7,656,222	7,656,222	7,656,222	2,286,796	7,656,222	7,656,222	7,656,222	7,656,222	7,656,222

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.

The results of this exercise are shown in Table 6.5. Column (1) shows that drivers get higher ratings when their app rating is lower, suggesting 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)-(10) measure to what extent this change is related to driving behavior. Except for higher speed and mounting with high ratings, the patterns on individual metrics are not clear.

We next conduct a similar exercise, where instead of exploring how drivers respond to app ratings, we look at how they respond to the last rating they received. Let r_i represent the rating given to driver $r(i)$ by the rider 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 that the driver received before the trip started. We run regressions of the form

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \alpha r_{l(i)} + P_2(n_i; \beta) + \epsilon_{ijkt} \tag{8}$$

where $P_2(n_i; \beta)$ is quadratic function of the number of trips completed by driver $d(i)$ before trip i. Panel A in Table 6.6 shows the result of these regressions.

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 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 (6). 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.[21]

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

Table 6.6: Response to last rating

	Dependent variable:									
	Rating	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: OLS									
Last rating	-0.0041^{***}	0.00003	-0.0003	0.0001	-0.0006^{***}	0.0012^{***}	-0.0004	0.00002	-0.00001	0.00002^{***}
	(0.0008)	(0.0006)	(0.0005)	(0.0003)	(0.0001)	(0.0004)	(0.0006)	(0.00001)	(0.00001)	(0.00001)
Observations	2,282,178	7,640,861	7,640,861	7,640,861	2,282,178	7,640,861	7,640,861	7,640,861	7,640,861	7,640,861
	Panel B: IV, average rating by rider									
Last rating	-0.0054^{**}	-0.0035^{**}	-0.0036^{**}	-0.0017^{**}	-0.0005	0.0017	0.0026	0.00002	-0.00002	0.00003
	(0.0021)	(0.0018)	(0.0017)	(0.0008)	(0.0004)	(0.0013)	(0.0017)	(0.00004)	(0.00003)	(0.00003)
Observations	2,087,525	6,994,328	6,994,328	6,994,328	2,087,525	6,994,328	6,994,328	6,994,328	6,994,328	6,994,328
	Panel C: IV, both instruments									
Last rating	-0.0133^{***}	-0.0012	-0.0023	-0.00003	-0.0012^{**}	0.0027^{*}	0.0011	-0.00003	-0.0001^{*}	0.00001
	(0.0026)	(0.0019)	(0.0019)	(0.0009)	(0.0005)	(0.0014)	(0.0019)	(0.00004)	(0.00003)	(0.00003)
Observations	1,014,732	3,416,989	3,416,989	3,416,989	1,014,732	3,416,989	3,416,989	3,416,989	3,416,989	3,416,989

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
All safety metrics are normalized to mean zero and variance one.
All regressions include rider and trip characteristics fixed effects.

Panels B and C in Table 6.6 shows the results of these 2SLS regressions. The results are consistent with the OLS results, and with the results in Table 6.5: ratings have a negative effect on new ratings, but they do not have a large effect on our metrics based on telematics.

We next explore how drivers' current app ratings affect the way they drive. Additionally, we want to see if the notification system influences their behavior. These notifications take place on the path towards deactivation according to the steps described in Table 6.1. Drivers move onto the next step when all three conditions are satisfied. In addition to the possibility that drivers move through this process towards deactivation, it is also possible to go back in the process. A driver that received the first notification can go back to normal if his average rating over his last 500 trips gets above the threshhold of 4.6. A driver that received notification 2 can go back to the same state right after notification 1 if his average rating over his last 500 trips goes above 4.5, and a driver that received notification 3 can go back to the state right after reactivation if his last 500 trip rating goes back above 4.45.

For this analysis, we run regressions of the following form:

$$y_i = \mu_{d(i)} + \nu_{c(i)} + \alpha r_i^{\text{app}} + \beta w_i + \epsilon_i, \tag{9}$$

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. Since we are including driver fixed effects, this regression exploits the variation due to drivers that crossed some threshold and got a notification. The number of such events is summarized in Table 6.3.

Table 6.7 shows the results of this exercise. Panel A measures the effect of 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 until the end. We see that notifications have a positive effect on ratings. They also affect every driving metric in the direction that riders prefer according to Table 3.1, although not all coefficients are significant. Notifications also have a positive effect on scores, which is consistent with the direction in which metrics change.

Panel B separates the effect of notifications across different notification states. They 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.

Table 6.7: Response to ratings and notifications

	Dependent variable:									
	Rating	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Panel A: General effect of notifications									
Has received notif.	0.1124***	0.0587***	−0.0452***	−0.0021	−0.0131*	0.0090*	−0.0031	0.0006***	0.0002***	0.0004***
	(0.0070)	(0.0117)	(0.0118)	(0.0070)	(0.0074)	(0.0052)	(0.0057)	(0.0002)	(0.0001)	(0.0002)
App rating	−0.3098***	0.0459***	−0.0246	0.0190*	0.0076	0.0424***	0.0196**	0.0005**	0.0002	0.0003
	(0.0144)	(0.0174)	(0.0188)	(0.0111)	(0.0116)	(0.0087)	(0.0093)	(0.0003)	(0.0001)	(0.0002)
Trip Characteristics FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Driver FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	2,286,772	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161
	Panel B: Decomposition of effect of notifications									
1st notification	0.1039***	0.0549***	−0.0416***	−0.0071	−0.0153**	0.0038	−0.0127**	0.0006***	0.0002**	0.0005***
	(0.0074)	(0.0116)	(0.0118)	(0.0071)	(0.0076)	(0.0055)	(0.0059)	(0.0002)	(0.0001)	(0.0002)
2nd notification	0.1345***	0.0294	−0.0394**	−0.0135	−0.0265**	0.0012	0.0145	0.0003	0.00003	0.0003
	(0.0133)	(0.0204)	(0.0188)	(0.0117)	(0.0118)	(0.0090)	(0.0098)	(0.0003)	(0.0001)	(0.0003)
3rd notification	0.2470***	0.0893*	−0.1199*	−0.0586**	−0.0321	0.0249	−0.0362	0.0027***	0.0011***	0.0017*
	(0.0373)	(0.0476)	(0.0660)	(0.0281)	(0.0274)	(0.0235)	(0.0260)	(0.0009)	(0.0004)	(0.0008)
1st notification exp.	0.1078***	0.0747***	−0.0461***	0.0192**	0.0010	0.0191***	0.0144*	0.0005**	0.0002**	0.0002
	(0.0081)	(0.0154)	(0.0140)	(0.0093)	(0.0096)	(0.0066)	(0.0075)	(0.0002)	(0.0001)	(0.0002)
2nd notification exp.	0.0964***	0.0142	−0.0070	0.0207	0.0184	0.0298*	0.0267	−0.0003	0.0004	−0.0006
	(0.0234)	(0.0319)	(0.0313)	(0.0197)	(0.0185)	(0.0165)	(0.0190)	(0.0005)	(0.0003)	(0.0004)
3rd notification exp.	0.2686***	0.1336*	−0.1937	−0.0829**	−0.1264**	0.0697**	−0.0475	0.0049***	0.0021***	0.0029*
	(0.0504)	(0.0791)	(0.1259)	(0.0410)	(0.0593)	(0.0343)	(0.0330)	(0.0018)	(0.0006)	(0.0015)
App rating	−0.3110***	0.0409**	−0.0222	0.0155	0.0059	0.0392***	0.0170*	0.0005*	0.0002	0.0003
	(0.0145)	(0.0173)	(0.0188)	(0.0110)	(0.0115)	(0.0087)	(0.0093)	(0.0003)	(0.0001)	(0.0002)
Trip Characteristics FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Driver FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	2,286,772	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161	7,656,161

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.

The takeaway from Table 6.7 is that although ratings do not seem to have a direct effect on driver's behavior, once drivers enter the deactivation process and start receiving notifications they do make substantial changes to their behavior. Most importantly, these effects are persistent, since they remain after notifications expire, i.e., when they start driving better and receive enough high ratings to get out of the deactivation process.

We now analyze in a nonparametric way the effect associated with the app rating, to see whether it occurs mostly at low or high ratings. If we just compared trips with high or low app rating we would obtain a spurious mechanical effect, since people with high ratings are systematically different from those with low ratings. We pool drivers into buckets of width 0.02 of their lifetime average rating. We then subtract to each outcome variable the average of the outcome variable for drivers in their bucket *that took place when the observed rating was between 4.7 and 4.8*. The main idea is that, after demeaning, we are measuring changes within groups of drivers with similar ratings, relative to trips when ratings were between 4.7 and 4.8 as a reference point. We then plot this difference in the vertical axis against the observed rating in the vertical axis. We see that there is a strong negative relationship between current app rating and the rating on the given ride, but there is little relationship between the current app rating and our scores.

More precisely, let $A_{r(i)}$ denote the set of all trips by drivers whose lifetime rating falls in the same bucket as driver $r(i)$ and during which the observed rating was in $[4.7, 4.8)$. If we are analyzing outcome y_i, we compute $\hat{y}_i = y_i - \frac{1}{|A_{r(i)}|} \sum_{i \in A_{r(i)}} y_i$. The term we subtract takes as a reference trips from similar drivers when their rating was between 4.7 and 4.8. We then average $\hat{y}_{ij}$ by buckets of observed rating of width 0.1; results are shown in Figure 6.2. Thus, every point in these plots is an indication of the effect of observed rating with respect to the effect that would have taken place if the observed rating was between 4.7 and 4.8. Therefore, it allows us to observe any heterogeneity in the effect of observed ratings on the various outcome variables we analyze.

6.2 Information on past behavior

The information that Uber discloses to drivers about their past behavior has the potential to influence their future behavior. One example is a driver's app rating: as we showed in the previous section, drivers with lower ratings behave in a way that increases their rating. Uber also previously informed drivers of their past

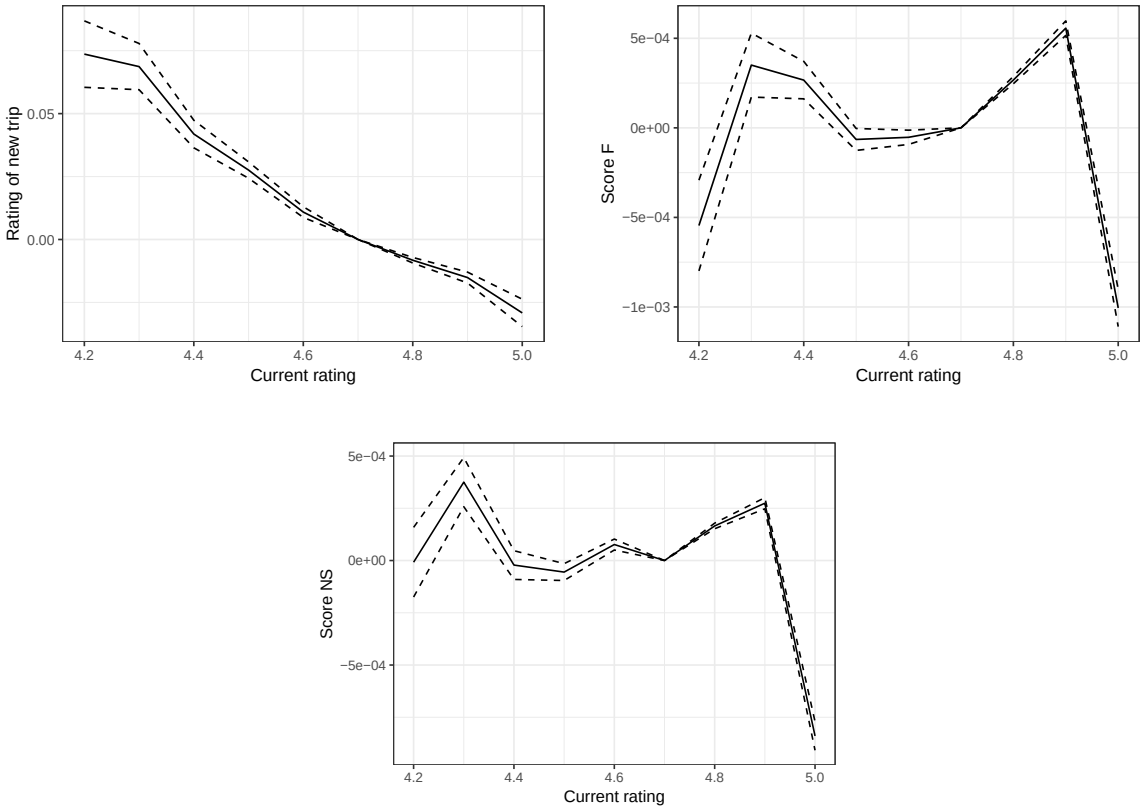


Figure 6.2: Effect of current observed rating on the rating and scores of new trips.

performance according to driving metrics[22]. Every week, drivers received a simple report, which fit in a smartphone screen, summarizing drivers' 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.

Every driver in our sample received this report, so there is no direct way to tell to what extent it affected their behavior. However, Uber ran a closely related experiment that is informative about how information of past behavior affects future behavior. 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 their past behavior. The dashboard included information on individual trips and on specific segments within each trip. Images of the dashboard are in Figure 6.3.

Uber initially treated a random set of drivers with the upgraded dashboard, and

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

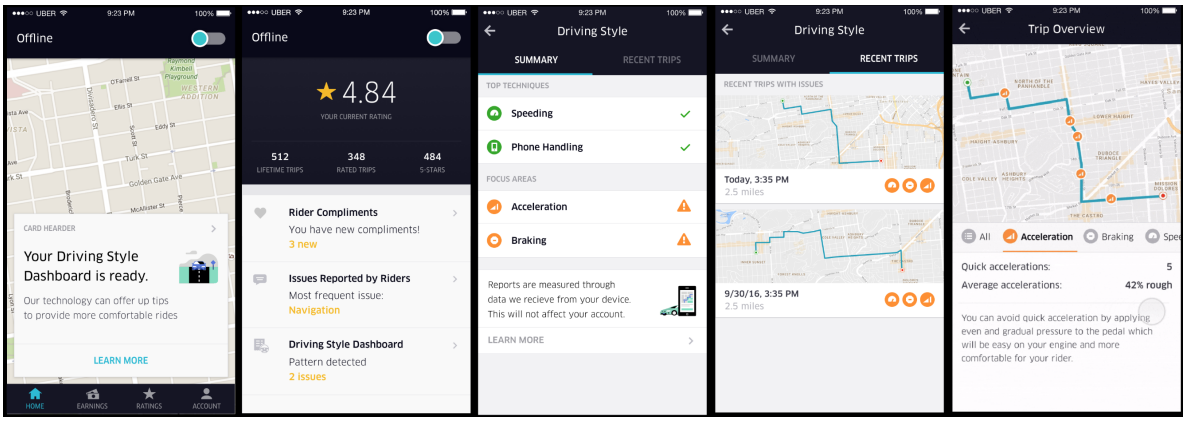


Figure 6.3: Images of the dashboard with detailed information about past driving behavior.

kept on sending the original report to a control group. We use this experimental setting to measure how additional information on past behavior influenced their behavior. If treated drivers improved their driving metrics, then it is natural to conclude that the original report is responsible for some of the difference in behavior between UberX and UberTaxi drivers.[23]

In Appendix F we report results of a balance test of pre-experiment averages by driver for each of the metrics and scores. We include only drivers who took at least ten trips in the month before the experiment. We find statistically insignificant differences for each of the non-speed metrics and scores, but we find significant differences for the speed variables, particularly for the low speed 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

$$y_i = \alpha + \tau T_{d(i)} + \gamma X_i + \epsilon_i, \tag{10}$$

where $T_{d(i)}$ 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). 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 a set of trip characteristics fixed effects.

We are also interested in measuring the effect of interacting with the new dashboard on outcomes y_i. We construct an indicator variable I_i which is equal to one if driver $d(i)$ interacted with the dashboard in the week prior to trip i. We are then

[23] Appendix F shows that the sample is balanced.

interested in estimating the following regression:

$$y_i = \alpha + \theta I_i + \gamma X_i + \epsilon_i. \tag{11}$$

We estimate this regression by 2SLS, instrumenting I_i with the treatment dummy and treatment interacted with the driver's pre-treatment mean for each of the telematics metrics and scores. Treatment status and the pre-treatment outcomes are demeaned for interpretability.[24] Our estimate of θ is thus an estimate of the average treatment effect, where being treated means interacting with the dashboard.

Table 6.8 shows results for regressions of the form in equations 10 and 11 where the dependent variables are our scores. Interacting with the dashboard leads to an improvement in all three scores, which means there is also an effect of having access to the dashboard. The effect is especially clear for the full and no-speed scores. We do not measure significant effects on individual metrics (see Appendix G).[25]

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. We exclude drivers with fewer than ten trips in the month before the experiment launched. We then code a dummy variable, referred to in Table 6.9 as "Bottom 10th Perc. Before," which indicates whether the driver was in the worst-performing 10% of drivers in the pre-period. For example, in column 1 "Bottom 10th Perc. Before" means the driver was below the 10th percentile for the full score, while in columns 2 and 3 it indicates the driver was below the 10th percentile for the speed score and non-speed score, respectively.[26]

Table 6.9 shows results from regressions similar to equations 10 and 11, but where access and interaction with the dashboard are interacted with our pre-treatment dummy. We see that the effects of the Dashboard are driven mainly by drivers that performed worst in the pre-treatment period. This means that informing poorly-performing drivers about their performance results in increased efforts and an improvement in their performance. There also seems to be a small improvement for drivers that do not perform poorly, suggesting that more detailed information does not lead to worse performance for previously good-performing drivers.

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

[25] 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.

[26] 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 6.8: Results of experiment

	Dependent variable:		
	Score F	Score S	Score NS
	(1)	(2)	(3)
	Panel A: Intent to treat estimator		
Treatment	0.0002*	0.00005	0.0001*
	(0.0001)	(0.00004)	(0.0001)
Pre-Period Mean	0.8199***	0.6876***	0.8430***
	(0.0055)	(0.0113)	(0.0054)
Observations	4,254,109	4,254,109	4,254,109
	Panel B: 2SLS estimator		
Interaction	0.0008***	0.0003**	0.0006**
	(0.0003)	(0.0001)	(0.0003)
Observations	4,254,109	4,254,109	4,254,109

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.

Table 6.9: Results of experiment, heterogeneity

	Dependent variable:		
	Score F	Score S	Score NS
	(1)	(2)	(3)
	Panel A: Intent to treat estimator		
Bottom 10th Perc. Before	−0.0271***	−0.0082***	−0.0255***
	(0.0005)	(0.0004)	(0.0004)
Treatment x Not Bottom 10th Perc.	0.0001	0.0001	0.00004
	(0.0002)	(0.0001)	(0.0001)
Treat x Bottom 10th Perc.	0.0015**	0.0006	0.0014**
	(0.0006)	(0.0005)	(0.0005)
Observations	4,254,109	4,254,109	4,254,109
	Panel B: 2SLS estimator		
Bottom 10th Perc. Before	−0.0008*	−0.0005**	0.0002
	(0.0005)	(0.0002)	(0.0004)
App Int. x Not Bottom 10th Perc.	0.0003	0.0001	0.0002
	(0.0002)	(0.0001)	(0.0002)
App Int. x Bottom 10th Perc.	0.0028***	0.0007	0.0027***
	(0.0010)	(0.0005)	(0.0009)
Observations	4,254,109	4,254,109	4,254,109

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.

7 Conclusions

Observers have expressed concern that ridesharing platforms might reduce quality by allowing inexperienced drivers in their platform, a consequence of their streamlined screening process that enables a flexible workforce. Using objective measures of driving quality, we find that UberX in fact provides better driving quality than taxis. We are not able to fully explain the forces shaping this finding, but we provide empirical evidence that the ratings, incentives, nudges, and information systems set up by Uber explain part of this difference.

Our paper raises the issue of what exactly the channels are that contribute to the difference 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 answer 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.

Acland, Dan and Matthew R. Levy, "Naiveté, Projection Bias, and Habit Formation in Gym Attendance," *Management Science*, 2015, *61* (1), 146–160.

Allcott, Hunt and Judd B Kessler, "The welfare effects of nudges: A case study of energy use social comparisons," Technical Report, National Bureau of Economic Research 2017.

_ **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. and Sydnee Caldwell, "Uber vs. Taxi: A Driver's Eye View," *Working Paper*, 2017.

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," *arXiv preprint arXiv:1603.09326*, 2016.

Buessing, Marric and Mauricio Soto, "Getting to the Top of Mind: How Reminders Increase Saving," *Center for Retirement Research at Boston College*, 2006.

Charness, Gary and Uri Gneezy, "Incentives to Exercise," *Econometrica*, 2009, *77* (3), 909–931.

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

Chen, M Keith, Judith A Chevalier, Peter E Rossi, and Emily Oehlsen, "The value of flexible work: Evidence from Uber drivers," Technical Report, National Bureau of Economic Research 2017.

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.

Cook, Cody, Rebecca Diamond, Jonathan Hall, John A. List, and Paul Oyer, "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers," *Working paper*, 2018.

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 in the Long-Run," Technical Report, Working Paper 2017.

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, "Labor Market Equilibration: Evidence from Uber," *Working paper*, 2017.

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.

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.

Leonard, Thomas C, "Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness," *Constitutional Political Economy*, 2008, *19* (4), 356–360.

Liu, Meng, Erik Brynjolfsson, and Jason Dowlatabadi, "Do Digital Platforms Reduce Moral Hazard? The Case of Uber and Taxis," Technical Report, National Bureau of Economic Research 2018.

Luca, Michael, "Reviews, reputation, and revenue: The case of Yelp.com," *Harvard Business School NOM Unit Working Paper*, 2011, *12-016.*

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," Technical Report, National Bureau of Economic Research 2015.

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.

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 and Cass Sunstein, *Nudge: Improving decisions about health, wealth, and happiness*, Yale University Press, 2008.

Appendix A Selection of metrics

In order to select our main variables, we run a lasso regression 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 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 A.1: Penalty at which variables are dropped by a lasso regression.

Metric	Penalty
Cont. speed 30	0.00001
Cont. speed 0	0.00001
Cont. speed mean	0.00002
Cont. speed 40	0.00003
Cont. speed 50	0.00004
Accelerations 2.5 m/s^2	0.0002
Brakes 3 m/s^2	0.0002
Accelerations 3 m/s^2	0.0002
Cont. speed 20	0.0003
Cont. speed 100	0.0004
Cont. speed 60	0.0005
Mounted	0.0008
Cont. speed 70	0.0011
Brakes 2.5 m/s^2	0.0027
Avg. speed when moving	0.0027
Cont. speed 10	0.0027
Handling	0.0036
Cont. speed 80	0.0040
Cont. speed 90	0.0040
Brakes 2 m/s^2	0.0044
Accelerations 2 m/s^2	0.0044
Excess distance	0.0146
Excess duration	0.0212

Table A.1 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, 90, and 10. Percentiles 80 and 90 measure similar information, so we choose only percentile 80 because it has a distribution with higher variance. We also choose percentile 10, despite being less predictive, because it captures speed during the slowest parts of a trip.

Appendix B 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).

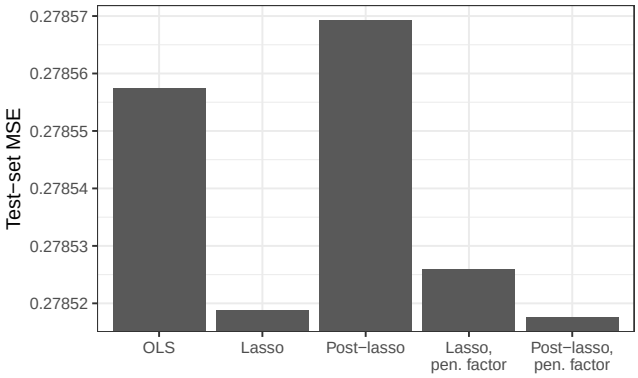


Figure B.1: Performance of different models for score.

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 B.1 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).

The coefficients we measure throughout our paper change very little if we use different methodologies, and the interpretation of all our results stays the same.

Appendix C 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 C.1 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. Furthermore, we do not observe any major changes from the main coefficients in Table 3.1.

We also explore how the UberX effect varies by car type. One major challenge is that only 5% of UberTaxi trips in our sample have a car model. However, we do observe the car year. This gives us a good idea of the car quality, given that taxi models tend to be relatively homogeneous. We split UberX trips into four groups. The first group ("unknown") includes trips with a car we did not find a price for. We split trips with a known car into three price quantiles ("low", "medium", and "high"). We split UberTaxi into those newer than the median and those older than the median.

Table C.1: Response to ratings and notifications, including car prices

	Dependent variable:					
	Rating		Rating is 5		Rated	
	(1)	(2)	(3)	(4)	(5)	(6)
Mounted	0.0049***	0.0037	0.0020**	0.0017	−0.0011*	−0.0014
	(0.0017)	(0.0034)	(0.0008)	(0.0016)	(0.0006)	(0.0011)
Handling	−0.0070***	−0.0044*	−0.0022***	−0.0009	−0.0002	−0.0010
	(0.0011)	(0.0023)	(0.0005)	(0.0010)	(0.0004)	(0.0007)
Brakes	−0.0047***	−0.0057***	−0.0016***	−0.0018***	0.0018***	0.0018***
	(0.0008)	(0.0015)	(0.0004)	(0.0007)	(0.0003)	(0.0005)
Accelerations	−0.0030***	−0.0032**	−0.0014***	−0.0008	0.0020***	0.0024***
	(0.0008)	(0.0016)	(0.0004)	(0.0008)	(0.0003)	(0.0005)
Speed low	0.0091***	0.0097***	0.0036***	0.0041***	−0.0039***	−0.0037***
	(0.0007)	(0.0014)	(0.0003)	(0.0007)	(0.0002)	(0.0005)
Speed high	−0.0015*	−0.0031**	−0.0018***	−0.0027***	−0.0032***	−0.0029***
	(0.0008)	(0.0015)	(0.0004)	(0.0007)	(0.0003)	(0.0005)
Price	0.0013	0.0013	0.0006	0.0006	0.0003	0.0003
	(0.0009)	(0.0009)	(0.0004)	(0.0004)	(0.0003)	(0.0003)
Price × Mounted		0.0002		0.00005		0.00005
		(0.0004)		(0.0002)		(0.0001)
Price × Handling		−0.0004		−0.0002		0.0001
		(0.0003)		(0.0001)		(0.0001)
Price × Brakes		0.0001		0.00003		0.000001
		(0.0002)		(0.0001)		(0.0001)
Price × Accels.		0.00002		−0.0001		−0.0001
		(0.0002)		(0.0001)		(0.0001)
Price × Speed low		−0.0001		−0.0001		−0.00003
		(0.0002)		(0.0001)		(0.0001)
Price × Speed high		0.0002		0.0001		−0.00004
		(0.0002)		(0.0001)		(0.0001)
Trip characteristics FE	✓	✓	✓	✓	✓	✓
Driver FE	✓	✓	✓	✓	✓	✓
Observations	1,413,283	1,413,283	1,413,283	1,413,283	4,773,516	4,773,516

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
All safety metrics are normalized to mean zero and variance one.

Table C.2: Comparison between UberX and UberTaxi by car quality

	Dependent variable:								
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
UberX, low	0.9454***	0.0836***	−0.0371**	−0.2743***	−0.0437***	−0.2448***	0.0026***	0.0025***	0.0011***
	(0.0353)	(0.0219)	(0.0176)	(0.0206)	(0.0117)	(0.0107)	(0.0004)	(0.0001)	(0.0004)
UberX, medium	0.9963***	0.0453**	−0.0088	−0.2084***	−0.0208*	−0.2068***	0.0026***	0.0024***	0.0011***
	(0.0352)	(0.0220)	(0.0176)	(0.0207)	(0.0117)	(0.0106)	(0.0004)	(0.0001)	(0.0004)
UberX, high	1.0044***	0.0314	0.0060	−0.2027***	−0.0228*	−0.2001***	0.0028***	0.0023***	0.0014***
	(0.0352)	(0.0220)	(0.0176)	(0.0208)	(0.0117)	(0.0106)	(0.0004)	(0.0001)	(0.0004)
UberX, unknown	0.9547***	0.0570***	−0.0912***	−0.2999***	−0.0821***	−0.2833***	0.0033***	0.0023***	0.0020***
	(0.0338)	(0.0204)	(0.0168)	(0.0196)	(0.0112)	(0.0100)	(0.0004)	(0.0001)	(0.0003)
UberTaxi, new	0.0368	−0.0292	−0.0364*	−0.0589**	−0.0139	−0.0105	0.0006	−0.0001	0.0007*
	(0.0450)	(0.0236)	(0.0213)	(0.0250)	(0.0151)	(0.0133)	(0.0004)	(0.0002)	(0.0004)
Trip Characteristics FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	8,530,144	8,530,144	10,459,336	10,459,336	9,511,238	9,511,238	7,855,743	7,855,743	7,855,743

Note: *p<0.1; **p<0.05; ***p<0.01

Table C.2 shows the results of regressions of the form of equation 4, where we measure treatment effects for the four UberX groups and for trips with new UberTaxi cars. In other words, we measure effects relative to trips with old UberTaxi cars. These coefficients do not show much heterogeneity from our main results in Table 4.1 . The main differences we observe are that drivers with more expensive UberX cars seem to handle their phone less and have more hard brakes. The net effect is a slightly higher score for expensive cars. UberX cars with unknown price brake significantly less, which leads to a somewhat more noticeable increase in scores.

Appendix D Controlling for routing

We create a score similar to score F, following equation (2), where $s(m_{ijkt};\theta)$ includes the sum of both terms in score F and a routing component, which is a quartic function of the distance metric interacted with a quartic function of the duration metric. We run the same cross validation procedure we did for our main scores to obtain estimates for θ. We then create a *residualized score* equal to $s(\tilde{m}_{ijkt};\theta)$, where $\tilde{m}_{ijkt}$ is the same as $\tilde{m}_{ijkt}$, except that the distance and duration metrics are set to zero. This score thus accounts for cell phone usage, speed, brakes, and accelerations, after residualizing any effect that may be taking place through routing. Both scores are very similar, with a correlation of 0.89.

Table D.1: Response to ratings and notifications

	Dependent variable:	
	Full score	Res. score
	(1)	(2)
	Panel A: Matching estimator	
UberX	0.0035***	0.0032***
	(0.0003)	(0.0002)
Observations	164,288	164,288
	Panel B: Trip characteristics fixed effects	
UberX	0.0033***	0.0031***
	(0.0003)	(0.0002)
Observations	7,849,896	7,849,896
	Panel C: Trip chars. and rider FEs	
UberX	0.0037***	0.0033***
	(0.0003)	(0.0002)
Observations	7,849,896	7,849,896

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.

Table D.1 shows results for our main UberX vs UberTaxi comparison, using both the full score and this new residualized score. As we can see, both scores give very similar results, with slightly smaller differences using the residualized score.

One might be concerned that our speed variables are capturing the effect of trip duration. However, if that was the case, we would see a smaller coefficients with our original scores, given that UberX drivers are slower.

Appendix E Variance decomposition by trip type

Table E.1 is a similar exercise to Table 5.3, but it focuses on the full score. It splits the sample across different trip characteristics: trips to and from airports, and trips during morning and afternoon rush hour.

Table E.1: Variance decomposition of score F, by different subsamples of the data (more than 20 trips).

	Driver	Rider	Trip characteristics	Residual
All	0.355	0.028	0.035	0.579
Beg. airport	0.304	0.028	0.036	0.465
End airport	0.332	0.029	0.029	0.513
Morning rush	0.344	0.030	0.041	0.576
Afternoon rush	0.353	0.027	0.032	0.548

Appendix F Balance of experimental sample

Table F.1 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 F.1: Balance Test for Experiment

	Dependent variable:								
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high	Score F	Score S	Score NS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Constant	0.6915***	0.1020***	0.2196***	0.1875***	23.7304***	86.0890***	−0.0152	−0.0170	−0.0089
	(0.0050)	(0.0022)	(0.0011)	(0.0013)	(0.0592)	(0.0525)	(0.0123)	(0.0117)	(0.0124)
Treatment	0.0009	−0.0021	−0.0001	−0.0015	0.2367***	0.1491**	0.0262	0.0292*	0.0153
	(0.0066)	(0.0028)	(0.0014)	(0.0018)	(0.0772)	(0.0682)	(0.0160)	(0.0153)	(0.0161)
Observations	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix G Effect of Dashboard Experiment on Metrics

Table G.1 shows results for regressions of the form in equations 10 and 11 where the dependent variables are the quality metrics.

Table G.1: Results of experiment, metrics

	Dependent variable:					
	Mounted	Handling	Brakes	Accels.	Speed low	Speed high
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Intent to treat estimator					
Treatment	−0.0036	−0.0157***	−0.0005	0.0019	0.0030	−0.0013
	(0.0060)	(0.0058)	(0.0042)	(0.0047)	(0.0028)	(0.0032)
Pre-Period Mean	2.0863***	3.6528***	4.6098***	4.7673***	0.0501***	0.0755***
	(0.0085)	(0.0352)	(0.0305)	(0.0291)	(0.0004)	(0.0005)
Observations	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109
	Panel B: 2SLS estimator					
Interaction	−0.0059	−0.0408**	−0.0087	−0.0193	0.0220***	0.0086
	(0.0257)	(0.0173)	(0.0123)	(0.0140)	(0.0080)	(0.0100)
Observations	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109	4,254,109

Note: *p<0.1; **p<0.05; ***p<0.01
All safety metrics are normalized to mean zero and variance one.