How do Platform Participants respond to an Unfair Rating? An Analysis of a Ride-Sharing Platform using a Quasi-Experiment

Anuj Kapoor and Catherine Tucker*

May 15, 2017

Abstract

The rise of peer-to-peer marketplaces has led to many previously unrated commercial transactions being rated. Of course, online rating systems can lead, on occasion, to reviews that are unfair or unrepresentative of the true quality provided. On the one hand, receiving an unfairly low rating once, might induce participants to exert more effort and receive a better rating the next time. On the other hand, it might dispirit participants and make them exert less effort. We use data from a ride-sharing platform in India where driver ratings were made particularly salient to the driver after each trip. We use instrumental variables to isolate the causal effect of receiving an unfairly bad rating. As our exogenous shifter of a driver's rating, we use whether or not that driver was a replacement for another driver who had previously canceled that particular customer's ride. We show that if a customer experiences a ride cancellation, they are more likely to unfairly blame the replacement driver. We show that drivers are more likely to respond negatively to a bad rating and receive subsequently bad ratings if they were blameless for the previous negative rating. We show that this effect is larger in contexts where there is a higher potential for an emotional response and when there is a greater need for driver skill in the subsequent ride. Finally, we show that these potentially unfair ratings can lead drivers to leave the platform, suggesting a broader negative effect of unfair negative ratings on platform participation.

Keywords: The Sharing Economy, User Generated Content, Ratings
JEL Codes: L86, M37

*Anuj Kapoor is a doctoral student at the University of Utah. Catherine Tucker is Sloan Distinguished Professor of Management and Professor of Marketing at MIT Sloan School of Management, MIT, Cambridge, MA 02139 and Research Associate at NBER. We thank an anonymous vendor who worked with us on this project. All mistakes are ours alone.

1 Introduction

Digitization has facilitated the emergence of many peer-to-peer platforms which try to match consumers better to underused resources. Interactions on these platforms requires a high level of trust and quality assurance, so ratings evaluating platform participants have become a core feature. Ratings received by service providers have become a key input into their ability to continue with and profit from the platform.

Bias in review systems of online marketplaces has been an active topic of research since at least 1999 Avery et al. (1999). The number of empirical studies on this topic has exploded with the advent of field experiments conducted by economists working with the platforms themselves (Nosko and Tadelis, 2015; Fradkin et al., 2015). Existing work focuses more on the reliability of reviews from a consumer welfare perspective, but little is known about the effect of bias in reviews on the platform participant who receives the review.

It is not clear what happens when participants receive a negative rating which may not be necessarily fair. On the one hand they may be induced to exert more effort to obtain better ratings in the future so that they can continue to participate profitably in the platform. On the other hand, unfairly low ratings may dispirit workers and lead them to exert less effort Dickinson and Villeval (2008), especially if the reviews are perceived as outside the control of the platform participant (Stanton and Barnes-Farrell, 1996).

Of course, trying to understand this question is riddled with identification challenges. In particular, ratings are unlikely to be unrelated to the characteristics of the worker receiving the rating, and these characteristics in turn may affect subsequent behavior. To overcome this challenge, we collected data on ratings behavior in a unique setting that allows us to identify the causal effect of an unfair rating on subsequent behavior. We collect data from a new taxi-sharing platform. This platform's app informed each driver after each ride (before they could take a new passenger) what rating they received from the earlier ride. To identify the causal effect of an unfair bad rating, we looked for an exogenous source of bad ratings that were unrelated to specific driver characteristics. We found such an exogenous shifter in the form of cancellations by *other* drivers. We show that if a customer had been assigned to another driver who canceled (thereby causing delay and inconvenience to the passenger), that customer tended to take out their aggravation by transferring the blame to the driver who was reassigned to their ride and giving them a substantially lower rating. This empirical pattern reflects an earlier literature on misattribution (Schwarz and Clore, 1983; Payne et al., 2005), that shows that judgments about one's satisfaction with an experience are heavily influenced by mood at the time of judgment.

Our results suggest that frustration caused by delays and inconvenience caused by the behavior of *other* drivers leads to a more negative rating for the replacement driver. The driver's negative response to the customer's unfairly negative rating in turn leads to worse performance in a subsequent ride. We provide a battery of robustness checks to ensure that our identification strategy is valid. We then turn to try and understand why we see this effect. We show that the effect we measure is stronger when taking place in a context where an emotional response to an unfair review is more likely, such as proximity in time to receiving the negative review, when the driver is young, and when the unfair rating represents a larger deviation from the driver's normal rating. Finally, we show that the cumulative effects of these unfair negative ratings can lead drivers to quit the platform, in a manner which simply having a lower average rating does not explain.

Through this research, we contribute to three main literature streams:

First, we contribute to the growing literature that studies ride-sharing platforms and peer-to-peer marketplaces. Cramer and Krueger (2016) paper examines the efficiency of ride-sharing services relative to traditional taxis and found that Uber drivers spend a significantly higher fraction of their time, and drive a substantially higher share of miles, with a passenger in their car than do taxi drivers. Hall and Krueger (2015) provide a thorough analysis of Uber's driver-partners, based on both survey data and anonymized, aggregated administrative data. Another welfare analysis of the sharing economy is Zervas et al. (2014), who explore the economic impact of the sharing economy on incumbents by studying the case of Airbnb, a prominent platform for short-term accommodations, and find negative but heterogeneous revenue effects. Zhang et al. (2016) find that drivers benefit significantly from their ability to learn from not only demand information directly observable in the local market, but also aggregate information on demand flows across markets. We contribute to this literature by studying the potential consequences of emotional reactions to elements of the customer experience that are outside the participants' control.

Second, we contribute to a literature on review dynamics. Muchnik et al. (2013) designed and analyzed a large-scale randomized experiment on a social news aggregation website and found that prior ratings created significant bias in individual rating behavior, creating asymmetric herding effects. Negative social influence inspired users to correct manipulated ratings, and positive social influence increased the likelihood of positive ratings. Looking more at level effects, Zervas et al. (2015) found that nearly 95% of Airbnb properties boast an average user-generated rating of either 4.5 or 5 stars (the maximum); virtually none have less than a 3.5 star rating. They contrast this with the ratings of approximately half a million hotels worldwide, collected on TripAdvisor, where there is a much lower average rating of 3.8 stars and more variance across reviews. Though these papers have established in general the possibility for correlation in ratings, our paper establishes a new finding where the correlation in ratings is actually driven by the participant's negative emotional response to what they perceive as an unfair rating.

Third, we contribute to a literature that studies how monitoring affects effort in a labor context. Dickinson and Villeval (2008) show that at low levels of monitoring, an increase in monitoring can improve effort, but that at a high level of monitoring there may be crowding-out effects. Stanton and Barnes-Farrell (1996) show that electronic monitoring which is highly visible, by reducing feelings of control, can inadvertently reduce worker satisfaction and associated effort. Outside of the lab, researchers have used survey research to establish a link between constant monitoring and review and emotional stress and performance Holman (2004). Other authors (Alge, 2001) have raised questions of how persistent electronic monitoring can affect the privacy and rights of workers. We contribute to this behavioral literature by showing the negative effects for workers if they receive an unfairly low rating.

Our results are important for managers as they help inform reputation system design in online platforms. Many papers have documented how much reviews, and negative reviews in particular, affect sales (Chevalier and Mayzlin, 2006; Chen et al., 2011). Our paper however, suggests that the potential for unfair negative reviews can also affect participation decisions and platform performance, as we show that an unfairly negative review can actually lead to worse subsequent performance, and that such reviews are more likely to lead that participant to leave the platform. This is important because while industry reports have documented individual anecdotes regarding platform participants being angered and 'burned out' by unfair reviews[1], to our knowledge we are the first paper to document that these unfair reviews can have real and harmful effects on platform participant performance and dedication to the platform.

This means that platform operators need to be careful when designing their platform so that there is no potential for misattribution of errors that are due to factors outside the platform participants' control.

[1]See for example this report - 'Being an Airbnb host is a lot of fun, but it can also be emotionally exhausting when you feel you've done everything you can to cater to an ungrateful guest. Including language that clears up this little issue with the way hosts and guests see reviews and ratings would go a long way towards fighting host burnout, and encouraging us to open our doors for years to come.' https://www.forbes.com/sites/sethporges/2016/06/29/the-one-issue-with-airbnb-reviews-that-causes-hosts-to-burnout/#6938b8f1eb3b

2 Background and Data

We obtained data from an anonymous ride sharing platform. The travel aggregator provided us data for six months for the city of Chandigarh in India. Unlike Uber and Lyft, the focus of this ride-sharing platform was to enable taxi drivers and taxi firms to provide a service more akin to those of better-known ride-sharing platforms. This means that the platform is less focused on the peer-to-peer aspects of the platform - for example, drivers do not rate customers. The firm's aim is more to professionalize taxi service so they can compete with firms such as Uber and Lyft.

There are three features of the ride-sharing platform which are important for our empirical analysis.

The first feature of the platform is the particular salience of how ratings are given to drivers. The taxi-aggregator platform provides the rating for the driver's previous ride, before the start of the next ride. Figure 1 below is a (slightly altered to protect the identity of the firm) screen shot of the app interface at driver's end.

Reputation is particularly important for transactions for taxi aggregator services because guest and driver interact with each other, in the car of the driver. Riders review drivers at the end of each session. Unlike many other peer-to-peer platform, drivers did not (during the time period we study) rate consumers. Each review includes ratings on a 1 to 5 scale. Given the fluidity of such platforms, it is important to note that during our observation period, our taxi-aggregator kept the incentives and rating system.

It is likely that the driver's own behavior can influence this rating. Drivers who can establish 'micro-relationships that make customers feel good' Rogers (2015) earn better ratings. To earn a five star rating, drivers may need to be friendly, and perhaps a little servile. Such emotional labor may impose a disparate burden on some drivers over others.[2]

[2]The driver training manual is mainly focused on the extent to which a driver's own behavior will influence their customer rating. They are exhorted to be 'humble', practice good personal hygiene, have a clean car,

The second feature of the platform is that around 3% of the time, a customer is reassigned to a new driver because the previous driver cancels the ride.

The third feature of this platform is that the location of the platform in Chandigarh presents us with unusual sources of variation in potential traffic flows. We exploit, in particular, traffic problems caused by the sacredness of cows in India when investigating the mechanism behind the effect. Specifically, we use the fact that in Chandigarh there are several roundabouts which attract cows to graze on them as an exogenous source of the need for driver skill.

2.1 Data Description

We received six months of data from June 2015 to December 2015 on 2,197 drivers from the platform. Table 1 provides summary statistics for our data at the trip-driver data. The average rating (if given) was 4.35 stars but many trips were not rated. Taking that into account only 23% of rides earned a 5 star rating.

Customers were reassigned to a new driver for 3% of trips. We have over 500,000 rides in our data, meaning that more than 15,000 rides were reassigned. 56% of rides were from areas which were not in the city center (rural). The average fare was 73 rupees which is around $1.15.

Table 2 provides summary statistics of our data at the driver level, which is our primary unit of observation when we move to explore driver platform exit behavior. On average each of our 2197 drivers made 241 trips during the 6 months of our data (sometimes they exited prior to the end of the data). On average this translated to just over 5 average trips per day.

2.2 Data

As in other online marketplaces, reviews on our taxi aggregator platform are predominantly positive. More than 70% of reviewers leave a 5 star rating. Figure 2 below plots the average

open doors for customers, and not make unaccompanied women uncomfortable.

ratings that the drivers received during our observation period.

Peer-to-peer platforms do not usually make it mandatory for the platform users to rate each other. Most of the time, drivers are not rated, so we focus on times that drivers actually get a five star rating. Past research has also emphasized that buyers and sellers with mediocre experiences review less than 3% of the time Dellarocas and Wood (2008). On similar lines,Nosko and Tadelis (2015); Fradkin et al. (2017) suggest that users with mediocre experiences are less likely to review. This means that a major indicator of a potentially bad rating is the withholding of a review. Given the strength of evidence presented in the field experiments of Fradkin et al. (2017); Nosko and Tadelis (2015) that a bad review is often an absent review, we therefore focus our study positively on the presence of a 5* review, and by implication treat occasions where a 5* review was not being given as negative. Earlier versions of the paper used whether or not the driver received an explicitly negative review, with similar effects.

We enriched the data we received from the taxi aggregating company with additional data. We used geo-tagging techniques to generate latitudes and longitudes for the pickup areas in a similar manner to Ghose et al. (2012). We use this locational data later on in our study to identify areas which may pose larger driving challenges than others, due to both the original design of the roads which dates from an era where cars were far narrower than they are today, and also the presence of sacred cows on the road. To account for heterogeneity across locations, we collect demographic information by pickup areas from the Indian Census[3] We used this data to classify the areas as rural versus urban.

3 Regression Approach

We first ask whether an unfair rating affects the rating for the subsequent ride. We build our main empirical specification at the individual driver level. For driver i for ride j and on

[3]2011 data available at`www.censusindia.gov.in`.

day d and time t, their likelihood of getting a top rating is a function of:

$$\begin{aligned} 5 * \text{Rating}_j = \alpha &+ \beta_1 5 * \text{Rating}_{j-1} \\ &+ \beta_2 \text{Fare}_j \\ &+ \beta_3 \text{Rural}_j \\ &+ \beta_4 \text{TripLength}_j \\ &+ \beta_5 \text{CustomerReassigned}_j \\ &+ \beta_6 \text{Day}_d \\ &+ \beta_7 \text{Time}_t \\ &+ \beta_8 \text{Driver}_i + \epsilon_k \end{aligned}$$

β_1 captures the key coefficient of interest for this paper, which is whether driver's previous rating has an impact on driver's rating for the current ride. $5 * \text{Rating}$ is a binary measure that captures whether or not the driver receive a 5* review. Therefore, $5 * \text{Rating}_{j-1}$ is the rating for the previous ride for each driver i for ride $j - 1$,

Fare_j is the fare for the ride j. Similarly, TripLength_j is the length of the ride in terms of time taken to complete the ride j. $\text{CustomerReassigned}_j$ is a binary variable that captures if the ride j was reassigned because it was canceled by another driver.

We also include a vector of fixed effects for the hour of the day and day of the week in Day_d and Time_t which allows us to capture how the traffic flow differs across the days which may in turn affect a driver's rating. For our specifications, we defined day of the week as either Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday and time of the day has been defined on hourly basis so there are fixed effects for each hour time window starting from 1 AM, 2 AM, 3 AM and so on. Finally, and importantly we also include a vector of fixed effects for each of the drivers in our dataset which should control for non-time varying differences in inherent driver quality.

4 Results

Though our research focuses on potentially unfair ratings, as a baseline we think it useful to present results from a straightforward correlation between an earlier rating and subsequent rating.

Accordingly, Table 3 provides ordinary least square estimates of the relationship between the rating a driver receives and the previous rating they had received.

Column (1) of Table 3 indicates that there is a positive relationship between receiving a 5* rating and receiving a 5* rating on a subsequent trip. This controls for driver characteristics via a driver fixed effect.

Column (2) of Table 3 shows that this holds when controlling for the characteristics of that trip, such as how long and expensive it was and whether it was rural in origin. We show that the probability of receiving a positive review is negatively related to that ride having being reassigned from another driver, which is a source of variation we will exploit in our instrumental variables specification.

One concern is that traffic on the roads varies and that this is driving the results. Since volume of traffic on the roads in general is also a function of time of day, controlling for time helps us rule out traffic volume on roads in general as the underlying mechanism driving our results. We present these controls in Column (3) of Table 3. In Column (4) of Table 3, we also control for the day of the week to reflect potential differences in congestion across weekends and weekdays, and also variation in traffic police activity across days.

4.1 Instrument Variable Estimation: Moving to a Causal Mechanism

Though the results of Table 3 provide evidence of a consistent correlation between ratings, they do not demonstrate a causal relationship. Therefore, we turn to instrumental variables.

To identify the causal relationship between ratings, we need a plausibly exogenous source that may influence ratings that is not correlated with the drivers' and consumers' unobserved attributes. We use as our instrument whether or not the customer was reassigned from another driver who had canceled their ride.

Figure 4 below presents initial model-free evidence for the strength of this potential instrumental variable. It reports the average chance of receiving a 5* rating for drivers who were the first driver allocated to a ride compared to drivers who were allocated to a ride where another driver had previously canceled They suggest that when customers have been delayed by the cancellation of another driver they tend to misdirect that anger towards the replacement driver and rate them lower.

Table 4 provides instrumental variables estimates of the presumptively causal relationship between the rating a driver receives and the previous rating they had received using whether or not the driver was allocated to a ride where a driver had previously canceled as our independent variable.

Table 5 presents first-stage results for our specification. The instrumental variable for prior rating is whether or not the prior ride was reassigned from another driver. As might be expected, given Figure 3, we see that if a customer had been assigned to another driver who canceled (thereby causing the passenger delay and inconvenience), that customer tended to take out their aggravation by transferring the blame to the driver who was reassigned to their ride and giving them a substantially lower rating.

Column (1) of Table 4 indicates that there is a positive relationship between receiving a 5* rating and receiving a 5* rating on a subsequent trip. Compared to Table 3, the point estimate of this effect is substantially larger. We interpret this increase in magnitude as suggesting that if the rating received was influenced by external forces - and consequently potentially unfair - drivers are far more likely to receive a negative rating than otherwise.

Column (2) shows that this result holds when controlling for the characteristics of that trip such as how long and expensive it was and whether it was rural in origin. We show that the probability of receiving a positive review is negatively related to that ride having being reassigned from another driver, which is in line with our findings in Figure 3.

One concern is that there is a varying amount of traffic on the roads and that this is driving the results. Since volume of traffic on the roads in general is also a function of time of day, controlling for the time and day of the week helps us rule out traffic volume on roads in general as the underlying mechanism driving our results. We present these controls in column (3). Further in Column (3), we also control for the day of the week to reflect potential differences in congestion across weekends and weekdays, and also variation in traffic police activity across days.

Another concern about our results is our use of a linear model with binary endogenous and exogenous variables. The proper technique to address this has been debated extensively (Horrace and Oaxaca, 2006; Angrist and Pischke, 2008). To address this debate, we report similar results using a probit functional form in Table A1. Though it is reassuring the results are similar, we emphasize that one downside of the probit functional form is that we cannot include driver fixed effects due to potential issues of bias (Greene et al., 2002).

In general, as we would expect given the model-free evidence in Figure 3, our first-stage F-tests suggest that our first stage and the associated instrumental variables are strong. The equation is exactly identified, meaning we do not run tests for over-identification. In the next section, we turn to consider challenges to our exclusion restriction.

5 Checking the Exclusion Restriction

Figure 3 and Table 5 suggests that our instrument meets the first criterion of being a valid instrument, that it is highly correlated with the endogenous variable. We have not yet investigated whether it also meets the exclusion restriction. The exclusion restriction states that the instrumental variable must not be related to the dependent variable except via its effect on the endogenous variable.

In any novel instrument such as ours, there are many potential challenges to the exclusion restriction. Table 6 investigates whether our results change when we remove instances where the instrument might not meet the exclusion restriction.

Column (1) of Table 6 investigates whether it is only new (and inexperienced) drivers who are unlucky enough to be reassigned, or accept reassignments to customers. We show that when we exclude these drivers, our results hold. In Column (2) of Table 6, we investigate whether it is only the drivers who drive late at night who accept reassignments to customers. We show that when we exclude these drivers, our results still hold.

Column (3) of Table 6 investigates whether exacting customers are more likely to be reassigned. We show that when we exclude such customers, our results hold. Last, Column (4) of Table 6 investigates whether it is only new customers who are most likely to be reassigned. We show that when we exclude these new customers, our results hold.

6 Mechanism

The next question is why drivers react negatively to receiving a rating which is unfair.

Rationally, a driver who receives a bad rating should invest more to compensate for that bad rating in subsequent trips. However, that is not we observe. Therefore, a potential explanation of our results is that what we are measuring is an emotional response to what is perceived as an unfair rating.

To tease this apart we look at instances where there is more likely to be emotion involved. The results of this investigation are reported in Table 7.

We stratify our results across different driver and trip characteristics. One potential shifter of emotional state is the amount of time which has elapsed between the current and the earlier trip. The idea is that the closer the previous rating is in time, the more likely it is that the driver would be still in the midst of an emotional response.

We investigate this assumption by segregating our rides into rides with a short interval between them versus a long interval. Column (1) of Table 7 investigates our results for short intervals; Column (2) of Table 7 investigates our results for long intervals. The coefficient on the previous rating of 5* variable is large and statistically significant for Column (1) and for the short-interval trips but not the long interval trips in column (2). This suggests that temporal proximity enhances the effect we measure, suggesting potentially that it is to do with the driver's current emotional state.

Along similar lines, it is possible that driver age determines their ability to have a non-emotional response to feedback (Fishbach and Finkelstein, 2012) and also that younger people react more strongly to feedback (Wang et al., 2015). Past research also suggests that experts seek more negative feedback than novices (Finkelstein and Fishbach, 2012) Column (3) of Table 7 displays the results for old drivers whereas Column (4) of Table 7 investigates our results for young drivers. The coefficients suggest that the effect is indeed larger for younger drivers.

In Columns (1) and (2) of Table 8, we looked to see whether the extent of the deviation that the rating represented matters. We test this assumption by segregating our rides into drivers with low deviation versus high deviation rides. We find that it is instances where the previous unfair rating represents a larger deviation from the driver's average rating where we tend to see the largest negative effect.

We then turn to consider why it is that an emotional response can lead to a negative rating. One hypothesis is that a negative emotional response can lead to a lower level of effort by the driver concerned. Therefore, we turn to look for instances in our data where more driver skill might be needed. Columns (3) and (4) of Table 8 investigate a potential source of variation in potential need for driver skill in our context.

Roundabouts are a standout cultural marker of the city of Chandigarh. However, these roundabouts were designed many decades ago by a French architect for a city with very different traffic patterns and vehicle types. They are now too narrow and cannot cope with the current width of average vehicular traffic. However, City residents have a strong affinity with the city's beautifully landscaped roundabouts and have opposed their removal.[4]] However, in addition to these roundabouts no longer being adapted for current traffic patterns, the presence of green grass on these roundabouts attracts cows and aggravates the existing traffic congestion and difficulties at the roundabouts.

Due to the geography of the city, drivers choose whether to navigate around the difficulties presented by these roundabouts. GPS generally reduces the extent to which driver effort can affect the smoothness of the ride, but in this case the driving challenges posed by these roundabouts require greater driver effort and thoughtfulness.

We created a categorical variable which indicated the presence of four large roundabouts around the pick up areas. These areas of four large roundabouts are the hardest to navigate around in Chandigarh city as there is not an easy route that allows the driver to avoid them. The presence of cows on and by roundabouts increases the difficulty of driving,[5] as does the fact that the roundabouts were designed for cars which were of far smaller width and number.

Columns (7) and (8) of Table 8 investigate this potential source of variation in potential need for driver skill in our context. We find that it is instances where the rides were difficult, i.e. where they were more likely to encounter roundabouts, where we see the largest negative effect. This is suggestive that the mechanism is that the driver is less likely to exert themselves to avoid potential traffic hazards if they receive a potentially unfair review.

[4]See for example - `http://chandigarh.gov.in/cmp2031/open-space.pdf`

[5]According to the University of Washington, cows sleep for approximately 3.9 hours per day, meaning that they spend about 16.9 percent of the day sleeping. Therefore they are often active and potentially causing traffic problems.

7 Why does it matter?

One obvious question is whether the pattern we observe in the data of drivers responding negatively to a negative rating matters. We investigate this by examining whether drivers are more likely to leave the platform as a result of an unfair negative rating.

We constructed a driver level panel database where the dependent variable is now a binary variable for whether it is the driver's last trip on the platform - that is their terminal trip. Table 9 displays the results.

Column (1) of Table 9 investigates our results using a discrete time hazard model (Allison, 1982), underpinned by a linear probability model and ordinary least squares. These results which have the raw average rating as the explanatory variable indicate that drivers with a low rating are not more likely to leave the platform than the drivers with higher ratings just in terms of the raw correlation. Column (2) looks at the relationship between leaving the platform and a cumulative average of our instrument, which is the proportion of rides the driver receives have been reallocated from another driver that canceled. The more rides a driver has been allocated that were reassigned, the more likely they are to leave the platform. We then turn to estimate a version of the specification which focuses on the exogenous variation in cumulative average rating, which can be explained by having been allocated customers whose driver had previously canceled.

Column (2) and Column (3) of Table 9 report results from this instrumental variables specification. The results indicate that low ratings that are a function of unfair consequences such as accepting reassigned customers are more likely to make a driver leave the platform. Last, as a robustness check, Column (4) of Table 9 repeats the results excluding rides from the final month of our data - so that we can be sure we are not accidentally identifying something as a final ride when actually the driver would return to the platform.

8 Conclusion

The rise of peer-to-peer marketplaces has led to a new need to provide quality information in platforms by ensuring that users continually rate the performance of other users. However, little is known about how the effect of being constantly rated affects user performance, and in particular how the potential for unfair negative ratings affects user performance.

We find evidence that negative (unfair) ratings lead drivers to have less good ratings subsequently and eventually be less likely to participate in the platform. We present suggestive evidence that this is due to a negative emotional response to the negative rating by the platform participant. To identify our effects we use plausibly exogenous variation in rating generated by negative spillovers from other drivers canceling on that customer prior to the ride.

Our results are important because though there has been much discussion and protest about unfair ratings in the industry, until now there has not been a systematic evaluation of how they might actually affect behavior.[6] Our results suggest that there are real and negative effect from platform participants from receiving an unfair rating, that affects the quality of their subsequent performance and also their dedication to the platform.

There are of course limitations to our findings. First, our results are limited to one city in the country India. It is not clear the extent to which our results may generalize internationally. Third, our vendor partner provides rating feedback to the driver straight after the end of the ride, and other vendors may not do so. Therefore, ratings may be more salient in our platform than in other platforms. Notwithstanding these limitations, we believe our paper provides a useful first step in studying how consistent rating affects the

[6]For example as reported by business insider 'The [rating] system in unfair,' Lotfi Benyedder, whose been an Uber driver for about three years, said to the crowd. 'A driver was given one star and was deactivated from the system for five days, but the guy has kids to feed, has family, has bills, and he was not able to drive because a difficult client gave him one star. He sent several emails to Uber and they did not respond until after six days and then they wanted him to take a class, when it was not even his fault. This rating system needs to stop. No more rating!' http://www.businessinsider.com/uber-drivers-protesting-2014-6

performance of users in a platform.

References

Bradley J Alge. Effects of computer surveillance on perceptions of privacy and procedural justice. *Journal of Applied Psychology*, 86(4):797, 2001.

Paul D Allison. Discrete-time methods for the analysis of event histories. *Sociological methodology*, 13:61–98, 1982.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Christopher Avery, Paul Resnick, and Richard Zeckhauser. The market for evaluations. *American Economic Review*, pages 564–584, 1999.

Yubo Chen, Qi Wang, and Jinhong Xie. Online social interactions: A natural experiment on word of mouth versus observational learning. *Journal of marketing research*, 48(2): 238–254, 2011.

Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.

Judd Cramer and Alan B Krueger. Disruptive change in the taxi business: The case of uber. *The American Economic Review*, 106(5):177–182, 2016.

Chrysanthos Dellarocas and Charles A Wood. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science*, 54(3): 460–476, 2008.

David Dickinson and Marie-Claire Villeval. Does monitoring decrease work effort?: The complementarity between agency and crowding-out theories. *Games and Economic behavior*, 63(1):56–76, 2008.

Stacey R Finkelstein and Ayelet Fishbach. Tell me what i did wrong: Experts seek and respond to negative feedback. *Journal of Consumer Research*, 39(1):22–38, 2012.

Ayelet Fishbach and Stacey R Finkelstein. How feedback influences persistence, disengagement, and change in goal pursuit. *Goal-directed behavior*, pages 203–230, 2012.

Andrey Fradkin, Elena Grewal, Dave Holtz, and Matthew Pearson. Bias and reciprocity in online reviews: Evidence from field experiments on airbnb. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 641–641. ACM, 2015.

Andrey Fradkin, Elena Grewal, and David Holtz. The determinants of online review informativeness: Evidence from field experiments on airbnb. 2017.

Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, 31(3):493–520, 2012.

William Greene, C Han, and P Schmidt. The bias of the fixed effects estimator in nonlinear models. *Unpublished Manuscript, Stern School of Business, NYU*, 29, 2002.

Jonathan V Hall and Alan B Krueger. An analysis of the labor market for ubers driver-partners in the united states. 2015.

David Holman. Employee well-being in call centres. In *Call Centres and Human Resource Management*, pages 223–244. Springer, 2004.

William C Horrace and Ronald L Oaxaca. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321–327, 2006.

Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

Chris Nosko and Steven Tadelis. The limits of reputation in platform markets: An empirical analysis and field experiment. Working Paper 20830, National Bureau of Economic Research, January 2015. URL `http://www.nber.org/papers/w20830`.

B Keith Payne, Clara Michelle Cheng, Olesya Govorun, and Brandon D Stewart. An inkblot for attitudes: affect misattribution as implicit measurement. *Journal of personality and social psychology*, 89(3):277, 2005.

Brishen Rogers. The social costs of uber. *University of Chicago Law Review Dialogue, Forthcoming*, 2015.

Norbert Schwarz and Gerald L Clore. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology*, 45(3):513, 1983.

Jeffrey M Stanton and Janet L Barnes-Farrell. Effects of electronic performance monitoring on personal control, task satisfaction, and task performance. *Journal of Applied Psychology*, 81(6):738, 1996.

Mo Wang, Gabriela Burlacu, Donald Truxillo, Keith James, and Xiang Yao. Age differences in feedback reactions: The roles of employee feedback orientation on social awareness and utility. *Journal of Applied Psychology*, 100(4):1296, 2015.

Georgios Zervas, Davide Proserpio, and John W Byers. The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research*, 2014.

Georgios Zervas, Davide Proserpio, and John Byers. A first look at online reputation on airbnb, where every stay is above average. 2015.

Yingjie Zhang, Beibei Li, and Krishnan Ramayya. Learning individual behavior using sensor data: The case of gps traces and taxi drivers. 2016.

Figure 1: Screen shot of the App from the Driver's Perspective

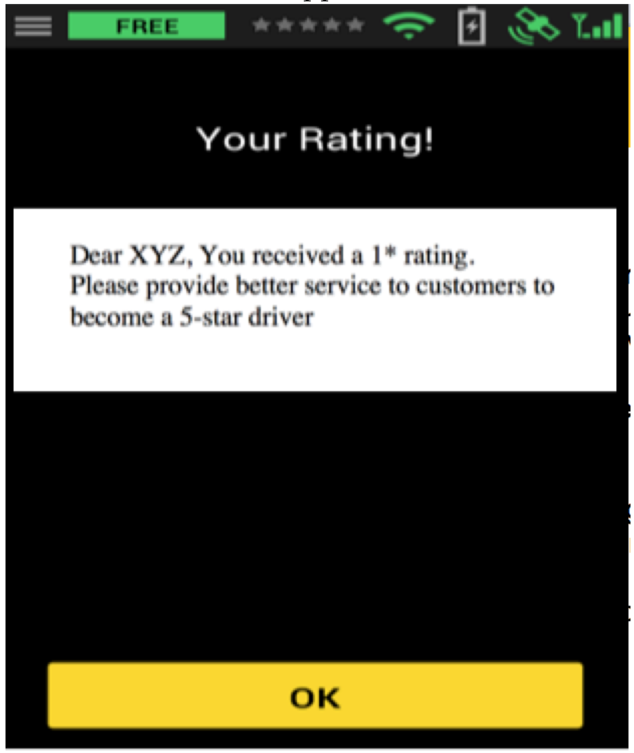

Figure 2: Distribution of Ratings that the Drivers Received during our Observation Period

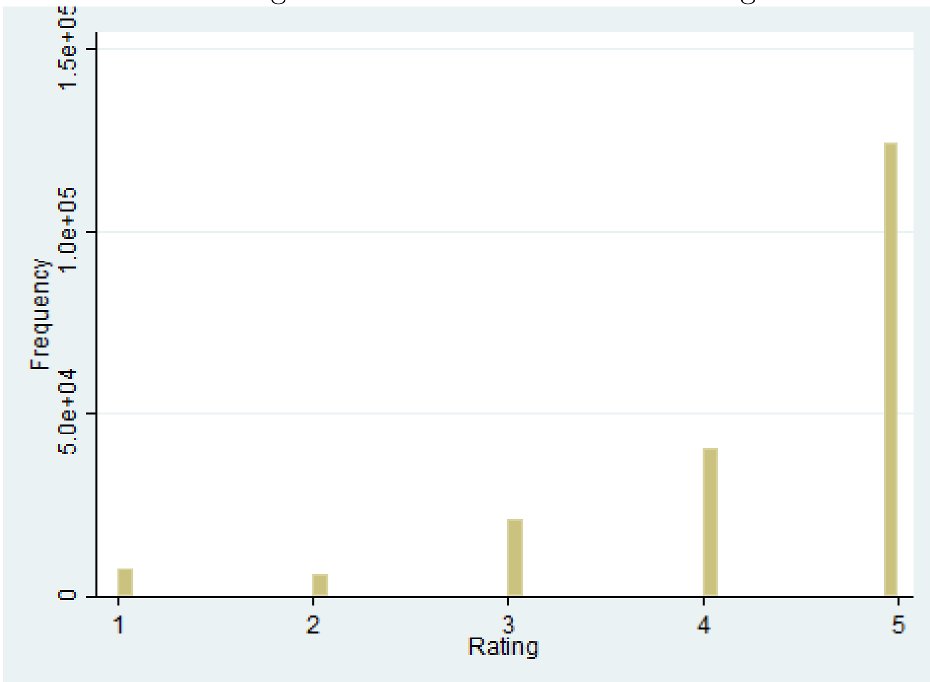

Figure 3: Model Free Evidence for Instrumental Variables Approach

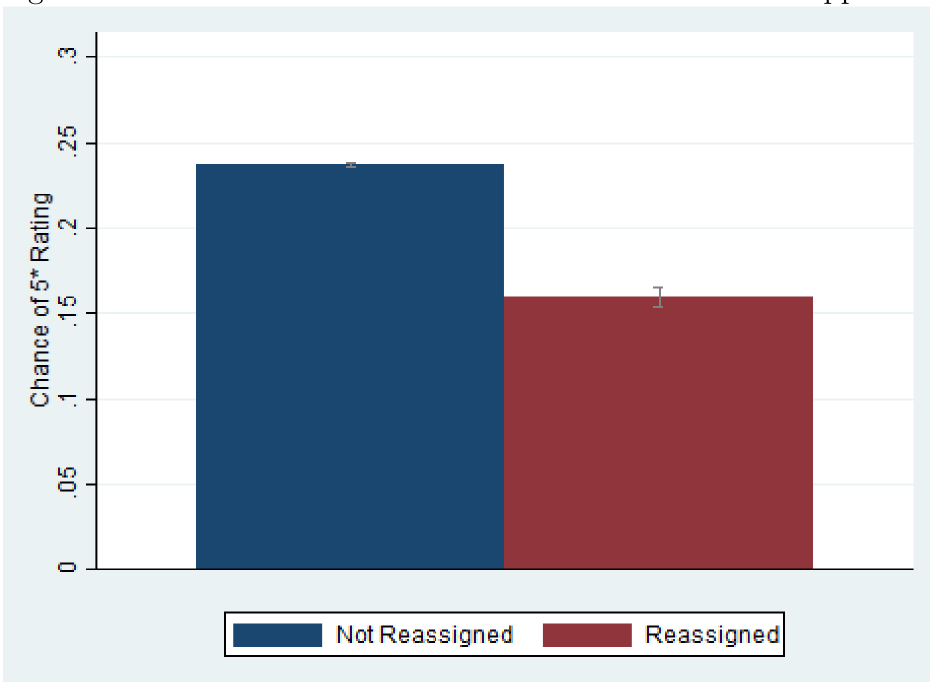

Table 1: Panel Summary Statistics

	Mean	Std Dev	Min	Max	Observations
Rating	4.35	1.03	0	5	198578
5* Rating	0.23	0.42	0	1	529469
Customer Reassigned	0.028	0.17	0	1	529469
Trip Length	24.2	127.2	-713.8	59216.6	529469
Rural	0.56	0.50	0	1	529469
Fare	73.0	34.1	40	3738	529469
Observations	529469				

Table 2: Driver Level Summary Statistics

	Mean	Std Dev	Min	Max	Observations
Cumulative Rating	4.34	0.40	1	5	2142
Trips During Dataset	241.0	276.0	1	1559	2197
Average Daily Trips	5.23	2.69	1	16.0	2197
Observations	2197				

Table 3: Ordinary Least Squares Estimates: Correlation in Ratings Over Time

	(1)	(2)	(3)	(4)
	5* Rating	5* Rating	5* Rating	5* Rating
Previous Rating 5*	0.0325***	0.0315***	0.0310***	0.0305***
	(0.0014)	(0.0014)	(0.0014)	(0.0014)
Customer Reassigned		-0.0633***	-0.0637***	-0.0637***
		(0.0035)	(0.0035)	(0.0035)
Trip Length		-0.0000***	-0.0000***	-0.0000***
		(0.0000)	(0.0000)	(0.0000)
Rural		-0.0026**	-0.0028**	-0.0027**
		(0.0012)	(0.0012)	(0.0012)
Fare		-0.0006***	-0.0006***	-0.0006***
		(0.0000)	(0.0000)	(0.0000)
Day of Week FE	No	No	No	Yes
Time Fixed Effects	No	No	Yes	Yes
Driver Fixed Effects	Yes	Yes	Yes	Yes
Observations	527272	527272	527272	527272
Log-Likelihood	-291056	-290148	-289875	-289750

Notes: Dependent variable is a binary indicator for whether the driver receives a 5* rating. Ordinary Least Square Estimates. Robust Standard Errors reported in parentheses. *** p <0.01, ** p <0.05, * p <0.1

Table 4: Instrumental Variables: The Correlation in Ratings is More Pronounced if Shifted by Something Outside the Driver's Control

	(1)	(2)	(3)
	5* Rating	5* Rating	5* Rating
Previous Rating 5*	0.1469***	0.1232**	0.1282***
	(0.0493)	(0.0495)	(0.0493)
Customer Reassigned		-0.0626***	-0.0629***
		(0.0035)	(0.0035)
Trip Length		-0.0000***	-0.0000***
		(0.0000)	(0.0000)
Rural		-0.0025**	-0.0026**
		(0.0012)	(0.0012)
Fare		-0.0006***	-0.0006***
		(0.0000)	(0.0000)
Day of Week FE	No	No	Yes
Time Fixed Effects	No	No	Yes
Driver Fixed Effects	Yes	Yes	Yes
Observations	527244	527244	527244
Log-Likelihood	-294485	-292363	-292261
Anderson Rubin F-Stat	8.99	6.24	6.82
Anderson Rubin p-value	0.0027	0.013	0.0090
Anderson canonical correlations LR	416.5	409.8	413.7
Anderson canonical correlations LR p-value	1.4e-92	4.0e-91	5.7e-92

Notes: Dependent variable is a binary indicator for whether the driver receives a 5* rating. Instrumental Variable Estimates. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 5: First-Stage Results for Table 4

	(1)	(2)	(3)
	L.5* Rating	L.5* Rating	L.5* Rating
L.Times Reassigned	-0.0460***	-0.0456***	-0.0458***
	(0.0023)	(0.0023)	(0.0022)
Customer Reassigned		-0.0079**	-0.0080**
		(0.0035)	(0.0035)
Trip Length		0.0000	0.0000
		(0.0000)	(0.0000)
Rural		-0.0012	-0.0012
		(0.0012)	(0.0012)
Fare		-0.0002***	-0.0002***
		(0.0000)	(0.0000)
Day of Week FE	No	No	Yes
Time Fixed Effects	No	No	Yes
Driver Fixed Effects	Yes	Yes	Yes
Observations	527244	527244	527244

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Dependent variable is a binary indicator for whether the driver received a 5* rating in Prior Trip. First stage estimates from 2SLS squared specification. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 6: Investigating whether our Results are Robust to Various Challenges to the Exclusion Restriction

	Not First Month	Not Late Night	Not Exacting Customer	Not New Customer
	(1)	(2)	(3)	(4)
	5* Rating	5* Rating	5* Rating	5* Rating
Previous Rating 5*	0.1401***	0.1473***	0.1290**	0.1059*
	(0.0508)	(0.0527)	(0.0511)	(0.0576)
Customer Reassigned	-0.0631***	-0.0611***	-0.0635***	-0.0642***
	(0.0037)	(0.0038)	(0.0037)	(0.0042)
Trip Length	-0.0000***	-0.0000***	-0.0000***	-0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Rural	-0.0029**	-0.0014	-0.0026**	-0.0024*
	(0.0012)	(0.0012)	(0.0012)	(0.0013)
Fare	-0.0006***	-0.0006***	-0.0006***	-0.0005***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Day of Week FE	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Driver Fixed Effects	Yes	Yes	Yes	Yes
Observations	507315	473826	507513	431444
Log-Likelihood	-281708	-262610	-287174	-252067
Anderson Rubin F-Stat	7.69	7.92	6.43	3.40
Anderson Rubin p-value	0.0055	0.0049	0.011	0.065
Anderson canonical correlations LR	390.7	363.5	393.5	316.9
Anderson canonical correlations LR p-value	5.9e-87	4.8e-81	1.5e-87	6.9e-71

Notes: Dependent variable is a binary indicator for whether the driver receives a 5* rating. Instrumental Variable Estimates. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 7: Suggestive Evidence that the Mechanism is Driven by Emotional Response (Part 1)

	Long InterRide Interval	Short InterRide Interval	Old	Young
	(1)	(2)	(3)	(4)
	5* Rating	5* Rating	5* Rating	5* Rating
Previous Rating 5*	0.0705	0.1790**	0.0196	0.1345*
	(0.0633)	(0.0827)	(0.0878)	(0.0811)
Customer Reassigned	-0.0639***	-0.0600***	-0.0564***	-0.0700***
	(0.0048)	(0.0053)	(0.0059)	(0.0061)
Trip Length	-0.0000***	-0.0001***	-0.0000***	-0.0000**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Rural	-0.0038**	-0.0013	-0.0010	-0.0033
	(0.0016)	(0.0017)	(0.0020)	(0.0020)
Fare	-0.0006***	-0.0006***	-0.0007***	-0.0006***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Day of Week FE	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Driver Fixed Effects	Yes	Yes	Yes	Yes
Observations	261304	265877	184935	184687
Log-Likelihood	-137626	-153545	-100418	-103942

Notes: Dependent variable is a binary indicator for whether the driver receives a 5* rating. Instrumental Variable Estimates. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 8: Suggestive Evidence that the Mechanism is driven by the Potential for an Emotional Response (Part 2)

	Small Deviation (1) 5* Rating	High Deviation (2) 5* Rating	Easy Driving (3) 5* Rating	Difficult Driving (4) 5* Rating
Previous Rating 5*	0.1063**	0.2827*	0.0826	0.1716**
	(0.0486)	(0.1620)	(0.0689)	(0.0725)
Customer Reassigned	-0.0580***	-0.0750***	-0.0604***	-0.0663***
	(0.0033)	(0.0114)	(0.0048)	(0.0053)
Trip Length	-0.0000***	-0.0005***	-0.0000***	-0.0000***
	(0.0000)	(0.0001)	(0.0000)	(0.0000)
Rural	-0.0031**	-0.0055	0.0050***	-0.0045**
	(0.0012)	(0.0035)	(0.0017)	(0.0018)
Fare	-0.0006***	-0.0006***	-0.0007***	-0.0005***
	(0.0000)	(0.0001)	(0.0000)	(0.0000)
Day of Week FE	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Driver Fixed Effects	Yes	Yes	Yes	Yes
Observations	445646	78791	262235	264940
Log-Likelihood	-225980	-55592	-146122	-145416

Notes: Dependent variable is a binary indicator for whether the driver receives a 5* rating. Instrumental Variable Estimates. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 9: It is only Situations where Ratings are Unfair which Lead Drivers to Leave the Platform

	OLS		IV	IV: Exclude Final Month
	(1)	(2)	(3)	(4)
	Last Trip	Last Trip	Last Trip	Last Trip
Cumulative Average Rating	0.0007	-0.0635**	-0.0635**	-0.0493**
	(0.0013)	(0.0251)	(0.0251)	(0.0214)
Date Fixed Effects	Yes	Yes	Yes	Yes
Observations	89158	89158	89158	80187
Log-Likelihood	56598	55387	55387	63405
Anderson Rubin F-Stat		6.58	6.58	5.41
Anderson Rubin p-value		0.010	0.010	0.020
Anderson canonical correlations LR		245.0	245.0	225.1
Anderson canonical correlations LR p-value		3.2e-55	3.2e-55	7.0e-51

Notes: Dependent variable is a binary indicator for whether it is the driver's final trip on the platform. Column (1) and (2) investigates our results using a discrete time hazard model and Ordinary Least Squares. Columns (3)-(4) reports results from a 2SLS squared approach. Full Driver Panel in Columns (1)-(3). Column (4) excludes final month of data. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A1: Robustness to Binary Functional Form

	Probit	Bi Probit
	(1)	(2)
	5* Rating	5* Rating
5* Rating		
Previous Rating 5*	0.1368***	1.3019***
	(0.0044)	(0.0808)
Customer Reassigned	-0.2336***	-0.2082***
	(0.0126)	(0.0118)
Trip Length	-0.0008***	-0.0008***
	(0.0001)	(0.0001)
Rural	-0.0095**	-0.0082**
	(0.0038)	(0.0034)
Fare	-0.0024***	-0.0021***
	(0.0001)	(0.0001)
L.five_rating		
L.Times Reassigned		-0.1824***
		(0.0086)
Day of Week FE	Yes	Yes
Time Fixed Effects	Yes	Yes
Observations	527272	527272
Log-Likelihood	-284652	-571354

Notes: Dependent variable is a binary indicator for whether the driver receives a 5* rating. Column (1) displays probit estimates. Column (2) displays bivariate probit estimates. Robust Standard Errors reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$