NET Institute*

www.NETinst.org

Working Paper #17-19

September 2017

**How do Platform Participants Respond to an Unfair Rating?
An Analysis of a Ride-Sharing Platform Using a Quasi-Experiment**

Anuj Kapoor
University of Utah

Catherine Tucker
MIT Sloan School of Management

* The Networks, Electronic Commerce, and Telecommunications (“NET”) Institute, <http://www.NETinst.org>, is a non-profit institution devoted to research on network industries, electronic commerce, telecommunications, the Internet, “virtual networks” comprised of computers that share the same technical standard or operating system, and on network issues in general.

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*

October 24, 2017

Abstract

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 someone a platform supplier to exert more effort and receive a better rating the next time. On the other hand, it might dispirit suppliers 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. Our data suggests that if a customer experiences a ride cancellation, they are more likely to unfairly blame the replacement driver. We use this as a exogenous source of unfair negative ratings for the 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. 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.

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. We also thank the NET Institute, www.netinst.org for funding. All mistakes are ours alone.

1 Introduction

Digitization has facilitated the emergence of many digital platforms which match customers with service providers. Interactions on these platforms requires a high level of trust and quality assurance, making ratings a core component of consumers' decisions. Correspondingly, ratings received by service providers have become a key input into their ability to continue with and profit from the platform.

It is not clear what happens when participants receive a rating which does not reflect accurately their performance. 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 service providers and lead them to exert less effort (Dickinson and Villeval, 2008), especially if the reviews are perceived as outside their control (Stanton and Barnes-Farrell, 1996).

We use data from a new ride-sharing platform launched in India. Reviews were unusually salient as this platform informed each driver after each ride (before they could take a new passenger) what rating they received from the earlier ride via a smartphone application. 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 a customer’s unfairly negative rating in turn leads to a worse driver rating in the 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 deterioration in ratings stemming from an unfair rating. 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. We present suggestive evidence that it is a negative emotional response to the unfair negative rating which leads to the deterioration in driver performance we see in our data. Finally, we show that there is asymmetry in how drivers respond to a potentially unfair rating. They appear to react more profoundly to a truly negative rating than a truly positive one.

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

First, we contribute to the growing literature that studies ride-sharing platforms. Much of this literature has been focused on calibrating potential welfare gains from ride sharing (Hall and Krueger, 2015; Cramer and Krueger, 2016; Chen et al., 2017; Zhang et al., 2016). There has of yet been little focus on the effect of ratings within these platforms, and we contribute to this literature by studying the potential consequences of emotional reactions to elements of the platform ratings experience that are outside the participants’ control.

Second, we contribute to a literature on review dynamics and bias. Bias in review systems of online marketplaces has been an active topic of research since the advent of the internet (Avery et al., 1999). Muchnik et al. (2013) found that prior ratings created significant bias in individual rating behavior, where too negative ratings inspired users to correct manipulated ratings, but positive social influence increased the likelihood of positive ratings. Though the potential for correlation in reviewing behavior (or lack of it) has been demonstrated, our paper is the first to suggest that this may stem from supply-side rather than demand-side

behavior.

Third, we contribute to a literature which tries to understand the consequences of the lack of variation in reviews. Zervas et al. (2015) emphasize the lack of heterogeneity in ratings on some platforms - nearly 95% of Airbnb properties boast an average user-generated rating of either 4.5 or 5 stars (the maximum). 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) to tackle the problem that often when customers have a negative experience they do not share a review rather than posting a negative review. Generally, the lack of non 5* reviews has been perceived as a problem for platforms, but our results suggest a countervailing effect which is that an undeserved negative review can reduce incentives for future performance.

Our results are important for managers as they help inform reputation system design in online platforms. Many papers have documented how 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.

This is important because while industry reports have documented individual anecdotes regarding platform participants being angered and ‘burned out’ by unfair reviews, 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.¹ This means that platform operators need to be careful when designing their platforms so that there

¹See 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>

is no potential for misattribution of errors due to factors outside the platform participants' control.

2 Background and Data

2.1 Firm Characteristics

We obtained data from an anonymous ride sharing platform. The platform provided us with 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. The firm aims to professionalize existing taxi services so they can compete with firms such as Uber and Lyft. The service just provides car-based taxi services and not alternatives such as autorickshaws. The firm was based in India and Chandigarh was one of the first markets in which it launched.

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. Riders review drivers at the end of each session. The platform provides the rating for the driver's previous ride, a short interval after the review is submitted but always before the next ride. Prior to our data, the platform allowed drivers to see the rating just as the rider was leaving the platform. This led to some unfortunate confrontational incidents, so the platform ensured that the rating was not presented before the passenger giving the rating had the opportunity to have decisively left the vehicle. Figure 1 (slightly altered to protect the identity of the firm) below is a screen shot of the interface at driver's end.

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, the platform kept its incentives and rating system constant. Unlike many other peer-to-peer platforms, drivers did not (during the time period we study) rate consumers. The platform in its communications to

the driver articulated that the likely main influence of each review was the driver’s own behavior. Indeed, 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, open doors for customers, and not make unaccompanied women uncomfortable by looking at them via the driver mirror. In the time period of our study, low ratings did not actually lead the platform to reject drivers (only serious direct complaints) but in their driver communications they indicated that low ratings were desirable in terms of avoiding rejection by the platform in the future.

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 assignment to a new driver was based on proximity of a driver to that customer at the time the other driver cancelled. The driver was not informed at the time of the ride that they were picking up a customer whose previous driver had canceled on them.

The third feature of this platform is that the location of the platform in Chandigarh presents us with unusual sources of variation in terms of difficulty of driving condition. Roundabouts are a standout cultural marker of the city of Chandigarh. These roundabouts were designed many decades ago, by Swiss architect Le Corbusier, 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. 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. This reflects the sacred status of cows in India. However, City residents have a strong affinity with the city’s beautifully landscaped roundabouts and have opposed their removal.² Crucial for our study is that these roundabouts present driving challenges which require greater driver effort and thoughtfulness.

²See for example - <http://chandigarh.gov.in/cmp2031/open-space.pdf>

2.2 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 level. 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. As in other online marketplaces, reviews on the platform are predominantly positive if they are actually given. More than 70% of reviewers leave a 5 star rating. Figure 2 below plots the average ratings that the drivers received, conditional on receiving a rating, during our observation period.

The average fare was 73 rupees which is around \$1.15. The average length of trip was around 25 minutes. Customers were reassigned to a new driver for 3% of trips. We have over 500,000 rides in our data, and more than 15,000 rides were reassigned.

We enriched the data we received from the platform 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 and roundabouts dating 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 information by pickup areas from the Indian Census.³ We used this data to classify the areas as suburban versus urban. As shown in Table 1, 56% of rides were from areas which were not in the city center (rural).

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.

³2011 data available at www.censusindia.gov.in.

3 Regression Approach

We build our main empirical specification at the individual driver level. We define our dependent variable as whether the driver does not receive our five star rating. 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 these papers that a bad review is often an absent review, we therefore focus our study on the absence of a five star review, and by implication treat occasions where a five star review was not being given as negative.

Therefore, we model that for driver i for ride j and on day d and time t , their likelihood of getting not getting a five star rating is a function of:

$$\begin{aligned}
\text{Not5} * \text{Rating}_{i,j} = & \alpha + \beta_1 \text{Not5} * \text{Rating}_{i,j-1} \\
& + \beta_2 \text{Fare}_j \\
& + \beta_3 \text{Surburban}_j \\
& + \beta_4 \text{TripLength}_j \\
& + \beta_5 \text{CustomerReassigned}_j \\
& + \beta_6 \text{Day}_d \\
& + \beta_7 \text{Time}_t \\
& + \beta_8 \text{Driver}_i + \epsilon_k
\end{aligned}$$

β_1 captures the key coefficient of interest for this paper, which is whether a driver's previous rating affects that driver's rating for the current ride. $\text{Not5} * \text{Rating}_{i,j}$ is a binary measure that captures whether or not the driver did not receive a 5* review. Therefore, $\text{Not5} * \text{Rating}_{i,j-1}$ 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 day of the week in Day_d and hour of the day in 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, the time of the day has been defined on an hourly basis so there are fixed effects for each hour time window starting from 1 AM, 2 AM, 3 AM and so on. Finally, we 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 not receiving a five star rating and not receiving a five star rating on a subsequent trip. This controls for persistent driver characteristics via a driver fixed effect.

Column (2) of Table 3 shows 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. In general longer, rural and more expensive rides have lower ratings. We show that the probability of receiving a non-positive review is positively 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 Effect

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 3 presents initial model-free evidence for the strength of this potential instrumental variable. It compares the average chance of receiving a five star 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. The estimates 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. Table 5 presents first-stage results for our specification. The instrumental variable for the endogenous variable of 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. In general, as we would expect given the model-free evidence in Figure 3, our first-stage F-tests suggest that our first stage instrument is indeed correlated with the endogenous variable. The equation is exactly identified, meaning we do not run tests for over-identification.

Column (1) of Table 4 indicates that there is a positive relationship between not receiving a five star rating and not receiving a five star 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 for the current trip.

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, in a manner similar to Table 3. Similarly, Column (3) and (4) show our results hold when controlling additionally for variation in ratings attributable to time and day of the week.

Another concern about our results is our use of a linear probability model with binary endogenous and instrumental 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 probit and bivariate probit functional forms 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).

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.

One challenge to the exclusion restriction is whether the decision to accept a ‘reassigned’ passenger could also be related to the underlying experience of the driver. If, for example, inexperienced drivers were the only ones willing to accept reassigned passengers, then this could directly be correlated with their expected rating for subsequent trips. Column (1) of Table 6 investigates whether our results hold if we exclude new (and inexperienced) drivers.

We show that when we exclude these drivers, our results hold.

A similar story is that reassigned customers have some type of undesirable characteristic (such as being drunk) which is also correlated with the kind of ratings a driver might expect in the next trip. Conversations with the firm suggest that some of the most undesirable characteristics of customers are drunkenness and that this drunkenness tends to be most prevalent late at night. In Column (2) of Table 6, we investigate whether our results hold if we exclude the drivers who drive late at night. We show that when we exclude these late night drivers, our results still hold.

Another challenge to the exclusion restriction is that reassigned customers are systematically different, and more likely to be negative in their rating behavior - for example if there was negative word of mouth about a particular customer which led more experienced drivers to reject them. Potentially if these more "hard to place" customers tended to be dropped off and picked up in similar places this could provide an alternative explanation of the effects we measure. Column (3) of Table 6 investigates whether exacting customers, that is customers who on average give a rating below three stars, are more likely to be reassigned. We show that when we exclude such customers, our results hold.

Another potential challenge is that it is only new customers, who potentially be less experienced about the need to give five star ratings, are more likely to be reassigned. Again if these new customers tend to share pickup and drop off points this could lead to the kind of correlation that we observe in our data. To investigate this, Column (4) of Table 6 investigates whether our results hold if we exclude new customers. We show that when we exclude these new customers, our results hold.

Aside from these tests, a more direct test is to try and see whether there any easily observable differences in the types of areas that taxi drivers canceled rides. This is the crucial variable of importance because when deciding whether to cancel a ride, the only information that the taxi drivers had accessible to them was the pickup-address. One concern is that

cancellation rates may be concentrated in pickup zones which were considered somewhat undesirable, and that undesirability was in turn linked with the customer rating.

To investigate this we looked at the distribution of cancellation across different zones in the city. Table 7 shows the distribution of cancellations. What is striking is that every single zone experienced cancellations and that the variation in these cancellation rates does not suggest a particular concentration in anyone region but instead kept within percentage points of the mean.

6 Why does a low rating provoke this negative response?

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 what we observe. Therefore, a potential explanation of our results is that what we are measuring is an emotional response to the perception of 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 8 where 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, as the literature on emotional intensity suggests that as time passes emotions become less intensely felt (Sonnemans and Frijda, 1994; Bruehlman-Senecal and Ayduk, 2015). We investigate this assumption by segregating our rides into rides with a short interval between ride j and $j - 1$ versus a long interval between j and $j - 1$. Column (1) of Table 8 investigates our results for short intervals; Column (2) of Table 8 investigates our results for long intervals. The coefficient on the previous five star rating 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, which provides suggestive evidence that the negative shift in subsequent rating is to do with the driver’s current emotional state.

A rating may be judged to be more unfair if it represents a bigger deviation from what the driver expects. In Columns (3) and (4) of Table 8, we looked to see whether the extent of the deviation of that day’s rating from the average rating of the driver matters. We test this assumption by separating daily ratings into low deviation versus high deviation from the average driver cumulative rating score. 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.

The existing literature also suggests that 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 (1) of Table 9 displays the results for old (above median age) drivers whereas Column (2) of Table 9 investigates our results for young (below median age) drivers. The coefficients suggest that the effect is indeed larger for younger drivers, again suggesting it is the potential for a negative response to negative feedback which drives our result. Another potential interpretation of this effect is that it reflects experience - it is only the older (and more experienced) drivers who have the reflexive skill to keep on driving well even if they are feeling upset at an unfair negative rating. We turn to explore this more in the next specifications.

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 9 investigate the presence of historic roundabouts as a potential source of variation in need for driver skill.

We created a categorical variable which indicated the presence of four large roundabouts around the pick up area. These areas are the hardest to navigate around in Chandigarh, as there is no easy way to avoid them, so instead the driver has to exert skill and exercise patience to navigate them. The presence of cows on and by roundabouts increases the difficulty of driving,⁴ 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 9 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.

7 Is the effect symmetric over different types of ratings

So far paper has only focused on the single case of whether or not the taxi driver reacting negatively to receiving a rating which wasn't five stars. A natural question is whether this is symmetric across different star ratings - that is does a driver react as positively to receiving a 4* rating as they do to receiving a 3* rating unfairly. Our current specification where we use exogenous variation caused by reassignment does not necessarily give us the power we need to contrast across these different ratings types.

Therefore, to investigate this we decided to use an additional type of exogenous variation which is the fact that some customers persistently give the same rating. We exploit this persistence in rating behavior how drivers behave when they are given a customer that has

⁴According 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.

an overarching tendency to give five ratings gives them a five rating, and compare this to their behavior when they have a customer who has an overarching tendency across all rides to give a rating that is at the lower end of the scale - at this case 3* and below. We create a new instrumental variable where we measure whether that customer has in the past given a 5* rating or a 'bad' rating of 3*. Table 10 gives our results. A comparison of Column (1) and (2) suggest that indeed there is an asymmetric effect - that the effect of getting a customer who has a tendency to gives bad ratings and getting a bad rating as a result of that - has a larger effect on the next rating the driver receives, relative to the effect of a 5* rating. However, we emphasize this is just a difference in the point estimate and the point estimates are quite close in size. In column (3) we investigate the effect of receiving no-rating. We see that the point estimate of this effect is somewhat between the other two - it has more of an effect than getting a positive rating but less of an effect than getting a truly negative rating.

8 Conclusion

The rise of peer-to-peer platforms 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. 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. We present suggestive evidence that this is due to a negative emotional response to the negative rating by the platform participant.

Our results are important because though there has been much discussion and protest about unfair ratings in the sharing economy, until now there has not been a systematic

evaluation of how they might actually affect participant behavior.⁵ Our results suggest that there are real and negative effects from platform participants from receiving an unfair rating, and that this 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 India. It is not clear how much our results may generalize internationally. Second, 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. Third, the consumer behavior we rely on for exogenous variation, which is the transfer of blame for a delay from the canceling driver to the replacement driver, may not be universal. As consumers become more sophisticated, there is the chance that they will not transfer blame across platform providers in this way. Notwithstanding these limitations, we believe our paper provides a useful first step in studying how consistent rating affects the performance of users in a platform.

⁵For example as reported by business insider ‘The [rating] system is 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>

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.
- Emma Bruehlman-Senecal and Ozlem Ayduk. This too shall pass: Temporal distance and the regulation of emotional distress. *Journal of personality and social psychology*, 108(2): 356, 2015.
- M Keith Chen, 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.
- 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. *Mimeo, MIT*, 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 Uber’s driver-partners in the united states. *Princeton University Industrial Relations Section Working Paper*, 587, 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.
- 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.
- Joep Sonnemans and Nico H Frijda. The structure of subjective emotional intensity. *Cognition & Emotion*, 8(4):329–350, 1994.
- 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. *Mimeo, Boston University*, 2015.
- Yingjie Zhang, Beibei Li, and Krishnan Ramayya. Learning individual behavior using sensor data: The case of GPS traces and taxi drivers. *Mimeo, Carnegie Mellon University*, 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	1	5	198578
Not 5* Rating	0.77	0.42	0	1	529469
Trip Length (Hr)	0.40	2.12	0	986.9	529469
Fare	73.0	34.1	40	3738	529469
Customer Reassigned	0.028	0.17	0	1	529469
Suburban	0.56	0.50	0	1	529469
Observations	529469				

Table 2: Driver Level Summary Statistics

	Mean	Std Dev	Min	Max
Cumulative Rating	4.34	0.40	1	5
Trips During Dataset	241.0	276.0	1	1559
Average Daily Trips	5.23	2.69	1	16.0
Birth Year	1982.2	7.77	1950	1996
Observations	2197			

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

	(1)	(2)	(3)	(4)
	Not 5* Rating	Not 5* Rating	Not 5* Rating	Not 5* Rating
Previous Rating Not 5*	0.0325*** (0.0014)	0.0315*** (0.0014)	0.0310*** (0.0014)	0.0305*** (0.0014)
Trip Length (Hr)		0.0014*** (0.0003)	0.0013*** (0.0003)	0.0013*** (0.0003)
Fare		0.0006*** (0.0000)	0.0006*** (0.0000)	0.0006*** (0.0000)
Customer Reassigned		0.0633*** (0.0035)	0.0637*** (0.0035)	0.0637*** (0.0035)
Suburban		0.0026** (0.0012)	0.0028** (0.0012)	0.0027** (0.0012)
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 rating which is not five stars. . 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) Not 5* Rating	(2) Not 5* Rating	(3) Not 5* Rating
L.Not 5* Rating	0.1469*** (0.0493)	0.1231** (0.0495)	0.1281*** (0.0493)
Trip Length (Hr)		0.0014*** (0.0003)	0.0014*** (0.0003)
Fare		0.0006*** (0.0000)	0.0006*** (0.0000)
Customer Reassigned		0.0626*** (0.0035)	0.0629*** (0.0035)
Suburban		0.0025** (0.0012)	0.0026** (0.0012)
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	-292360	-292258
Anderson Rubin F-Stat	8.99	6.23	6.81
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 rating which is not five stars. . 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.Not 5* Rating	L.Not 5* Rating	L.Not 5* Rating
L.Times Reassigned	0.0460*** (0.0023)	0.0456*** (0.0023)	0.0458*** (0.0022)
Trip Length (Hr)		-0.0001 (0.0003)	-0.0001 (0.0003)
Fare		0.0002*** (0.0000)	0.0002*** (0.0000)
Customer Reassigned		0.0079** (0.0035)	0.0080** (0.0035)
Suburban		0.0012 (0.0012)	0.0012 (0.0012)
Day of Week FE	No	No	Yes
Time Fixed Effects	No	No	Yes
Driver Fixed Effects	Yes	Yes	Yes
Observations	527244	527244	527244

Notes: Dependent variable is a binary indicator for whether the driver received a five star 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 (1) Not 5* Rating	Not Late Night (2) Not 5* Rating	Not Exacting Customer (3) Not 5* Rating	Not New Customer (4) Not 5* Rating
Previous Rating Not 5*	0.1401*** (0.0508)	0.1472*** (0.0527)	0.1289** (0.0511)	0.1059* (0.0576)
Trip Length (Hr)	0.0014*** (0.0003)	0.0013*** (0.0003)	0.0014*** (0.0003)	0.0013*** (0.0003)
Fare	0.0006*** (0.0000)	0.0006*** (0.0000)	0.0006*** (0.0000)	0.0005*** (0.0000)
Customer Reassigned	0.0631*** (0.0037)	0.0611*** (0.0038)	0.0635*** (0.0037)	0.0642*** (0.0042)
Suburban	0.0029** (0.0012)	0.0014 (0.0012)	0.0025** (0.0012)	0.0024* (0.0013)
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	507524	431444
Log-Likelihood	-281706	-262607	-287175	-252065
Anderson Rubin F-Stat	7.69	7.91	6.42	3.39
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 rating which is not five stars. . Instrumental Variable Estimates. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 7: Geographic Variation in Pickup Zone by Cancellation Raate7

(1)			
	Not Reassigned	Reassigned	Total
1	95.53	4.47	100.00
2	95.96	4.04	100.00
3	95.96	4.04	100.00
4	96.86	3.14	100.00
5	98.18	1.82	100.00
6	97.72	2.28	100.00
7	97.68	2.32	100.00
8	97.78	2.22	100.00
9	97.64	2.36	100.00
10	96.82	3.18	100.00
11	97.42	2.58	100.00
12	97.94	2.06	100.00
13	97.38	2.62	100.00
14	97.58	2.42	100.00
15	97.94	2.06	100.00
16	97.81	2.19	100.00
17	97.60	2.40	100.00
18	97.35	2.65	100.00
19	97.63	2.37	100.00
20	97.58	2.42	100.00
21	97.19	2.81	100.00
22	96.73	3.27	100.00
23	97.14	2.86	100.00
24	97.10	2.90	100.00
25	97.35	2.65	100.00
26	96.59	3.41	100.00
27	97.67	2.33	100.00
28	97.33	2.67	100.00
29	93.72	6.28	100.00
30	97.83	2.17	100.00
31	96.59	3.41	100.00
32	97.89	2.11	100.00
33	96.46	3.54	100.00
34	95.16	4.84	100.00
35	97.77	2.23	100.00
36	96.79	3.21	100.00
37	96.79	3.21	100.00
38	97.19	2.81	100.00
39	97.16	2.84	100.00
40	96.79	3.21	100.00
41	97.20	2.80	100.00
42	97.37	2.63	100.00
43	97.18	2.82	100.00
44	95.88	4.12	100.00
45	96.48	3.52	100.00
46	97.66	2.34	100.00
47	97.13	2.87	100.00
48	97.17	2.83	100.00
49	96.14	3.86	100.00
50	96.46	3.54	100.00
51	97.87	2.13	100.00
52	97.07	2.93	100.00
53	96.28	3.72	100.00
54	96.43	3.57	100.00
55	97.71	2.29	100.00
56	97.30	2.70	100.00
57	96.48	3.52	100.00
58	96.33	3.67	100.00
59	96.65	3.35	100.00
60	94.93	5.07	100.00
Total	97.16	2.84	100.00

Notes: Distribution of cancellation percentage rates by 60 zones within Chandigarh.

Table 8: Suggestive Evidence that the Mechanism is Driven by Emotional Response (Part 1)

	Long Interval (1) Not 5* Rating	Short Interval (2) Not 5* Rating	Small Deviation (3) Not 5* Rating	High Deviation (4) Not 5* Rating
Previous Rating Not 5*	0.0704 (0.0633)	0.1790** (0.0827)	0.1063** (0.0486)	0.2826* (0.1620)
Trip Length (Hr)	0.0010*** (0.0003)	0.0049*** (0.0009)	0.0012*** (0.0004)	0.0284*** (0.0058)
Fare	0.0006*** (0.0000)	0.0006*** (0.0000)	0.0006*** (0.0000)	0.0006*** (0.0001)
Customer Reassigned	0.0639*** (0.0048)	0.0600*** (0.0053)	0.0580*** (0.0033)	0.0750*** (0.0114)
Suburban	0.0038** (0.0016)	0.0013 (0.0017)	0.0031** (0.0012)	0.0055 (0.0035)
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	445646	78791
Log-Likelihood	-137625	-153545	-225978	-55590

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

Table 9: Suggestive Evidence that the Mechanism is driven by the Potential for an Emotional Response (Part 2)

	Old (1) Not 5* Rating	Young (2) Not 5* Rating	Easy Driving (3) Not 5* Rating	Difficult Driving (4) Not 5* Rating
Previous Rating Not 5*	0.0196 (0.0878)	0.1344* (0.0811)	0.0825 (0.0689)	0.1716** (0.0725)
Trip Length (Hr)	0.0011*** (0.0004)	0.0012** (0.0005)	0.0012*** (0.0003)	0.0017*** (0.0005)
Fare	0.0007*** (0.0000)	0.0006*** (0.0000)	0.0007*** (0.0000)	0.0005*** (0.0000)
Customer Reassigned	0.0564*** (0.0059)	0.0699*** (0.0061)	0.0604*** (0.0048)	0.0662*** (0.0053)
Suburban	0.0010 (0.0020)	0.0033 (0.0020)	-0.0050*** (0.0017)	0.0045** (0.0018)
Day of Week FE	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Driver Fixed Effects	Yes	Yes	Yes	Yes
Observations	184935	184687	262235	264940
Log-Likelihood	-100418	-103940	-146120	-145416

Notes: Dependent variable is a binary indicator for whether the driver receives a rating which is not five stars. . Instrumental Variable Estimates. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 10: Instrumental Variables: Using state persistence in customer rating behavior to tease apart asymmetry in reactions to different types of ratings

	(1) 5* Rating	(2) 3* and below Rating	(3) No Rating
L.5* Rating	0.0611*** (0.0025)		
L.3* and below Rating		0.0690*** (0.0028)	
L.No Rating			0.0674*** (0.0027)
Trip Length (Hr)	-0.0013*** (0.0003)	0.0016*** (0.0003)	0.0016*** (0.0003)
Fare	-0.0006*** (0.0000)	0.0007*** (0.0000)	0.0006*** (0.0000)
Customer Reassigned	-0.0634*** (0.0035)	0.0775*** (0.0038)	0.0682*** (0.0040)
Suburban	-0.0027** (0.0012)	0.0011 (0.0013)	-0.0018 (0.0013)
Day of Week FE	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes
Driver Fixed Effects	Yes	Yes	Yes
Observations	527244	527244	527244
Log-Likelihood	-289996	-336088	-360217
Anderson Rubin F-Stat	204.5	223.9	231.0
Anderson Rubin p-value	1.5e-132	3.7e-145	8.4e-150
Anderson canonical correlations LR	155677.4	126201.4	136476.7
Anderson canonical correlations LR p-value	0	0	0

Notes: Dependent variable is a binary indicator for whether the driver receives a rating which is five stars in column (1), less than 3 stars in Column (2). Instrumental Variable Estimates and no rating being given in Column (3). Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table A1: Robustness to Binary Functional Form

	Probit (1) Not 5* Rating	Bi Probit (2) Not 5* Rating
Not 5* Rating		
Previous Rating 5*	0.1368*** (0.0044)	1.3062*** (0.0797)
Trip Length (Hr)	0.0497*** (0.0045)	0.0472*** (0.0042)
Fare	0.0024*** (0.0001)	0.0021*** (0.0001)
Customer Reassigned	0.2333*** (0.0126)	0.2076*** (0.0118)
Suburban	0.0096** (0.0038)	0.0082** (0.0034)
L.not_five_rating		
L.Times Reassigned		0.1822*** (0.0086)
Day of Week FE	Yes	Yes
Time Fixed Effects	Yes	Yes
Observations	527272	527272
Log-Likelihood	-284648	-571350

Notes: Dependent variable is a binary indicator for whether the driver receives a rating which is not five stars. . Column (1) displays probit estimates. Column (2) displays bivariate probit estimates. Robust Standard Errors reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$