Measuring Psychological Effects of Peer-to-Peer Reputation Systems Involving In-Person Exchanges

A Thesis
SUBMITTED TO THE FACULTY OF UNIVERSITY OF MINNESOTA
BY

Mark Yousef

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

Advisor: Haiyi Zhu

July, 2019
Abstract

Reputation systems such as those used by peer-to-peer services have proven significant in helping companies better understand and manage their users. Seemingly the new credit scores for the digital economy, these personal rating systems have unexplored consequences on human psyche. Using a case study of Uber passengers and drivers, this study examines stress and control levels associated with personal rating scores. We found that while drivers indicated more difficult experiences in response to the control of their scores, passengers had issues with distress in relation to factors commonly associated with bias, such as age and ethnicity. Both groups exhibited lower perceptions of distress the more times they had used Uber. Overall, the use of peer-to-peer reputation systems can be improved to provide users a higher level of control and lower distress in response to ratings.
# Table of Contents

List of Tables ........................................................................................................................................... iii

List of Figures .............................................................................................................................................. iv

Introduction .................................................................................................................................................. 1

Theory and Hypothesis ............................................................................................................................... 2

Method ......................................................................................................................................................... 3

Results ......................................................................................................................................................... 6

Discussion ................................................................................................................................................... 13

Implications ............................................................................................................................................... 14

Conclusion .................................................................................................................................................. 15

Limitations ................................................................................................................................................. 15

References .................................................................................................................................................. 16
List of Tables

i. Table 1: Control questionnaire scores for both drivers and passengers; page 9

ii. Table 2: Distress questionnaire scores for both drivers and passengers; page 9
List of Figures

i. Figure 1: Ranked ANOVA between times total driven for Uber and distress questionnaire scores for the driver group; Page 9

ii. Figure 2: Ranked ANOVA between times total ridden with Uber and distress questionnaire scores for the passenger group; Page 10

iii. Figure 3: Ranked ANOVA between time since last driving for Uber and distress questionnaire scores for the driver group; Page 11

iv. Figure 4: Ranked T-Test between ethnicity and distress questionnaire scores for the passenger group; Page 11

v. Figure 5: Ranked ANOVA between Uber passenger ratings and their distress questionnaire scores; Page 12

vi. Figure 6: Ranked ANOVA between age and distress questionnaire scores for the passenger group; Page 13
Introduction

Peer-to-peer social reputation systems or trust systems, are heavily used throughout various online applications and are often the basis where sharing economy apps hinge. As these trust systems continue to develop, companies have come forward about their use previously unbeknownst to the user. Apple uses these systems to check for fraudulent app store purchases as mentioned in their updated privacy policy [24–26]. Facebook uses them to filter for fake news [7,11,14,28]. Furthermore, the rise in these systems have been predicted since 2010 along with a future aggregate of social reputation scores across various platforms, called reputation banks [18]. While users may currently have the option of not using such systems, completely signing off social media applications can yield unexpected consequences for many—57 percent of employers who screen their candidates said they were less likely to hire a candidate if they had no social media presence [5]. In addition, as social reputation scores become commonplace, a position is open for a company that can win the public's approval in order to aggregate all their scores. As an example, in 2015 an official “yelp for people” application was announced called, Peeple, allowing users to rate anyone publicly for the first time [6,27,29,30]. The application was met with huge controversy and shut down, yet is projected to relaunch within this year [31]. Currently there still exists, and is in use, an application called, MyLife.com, which provides similar features but is not as well-known [32].
To date, no study has measured how these rating systems can negatively affect user psyche. For example, users are often rated by others on a variety of subjective criteria, such as personality, looks, etc. Previous research has already shown bias in the popular ride hailing app, Uber’s, rating system among older drivers, female drivers paired with a majority of male passengers, female drivers between the ages of 30-50, as well as driver and passenger combinations which exhibit the lowest age differences [13]. Other researchers have specified that the inherent design of a social reputation system is itself not without some form of bias [20]. While it may be clear for producers in such systems that they are primarily rated on the quality of their service, the rating of consumers is a more recent phenomenon [19,22,23]. Urged by the consequences of low ratings, consumers may be implicitly placed in a role as a type of service provider. For example, there have been reports of some drivers on the Uber ridesharing service withholding five-star ratings from passengers who fail to make adequate conversation [10,15,17]. Similar implicit types of non-monetary payments required by service providers and affecting consumer reviews, has been noted among the home sharing service, Couchsurfing [12]. Further highlighting some possible undesired responses to its consumer rating system, Uber initially did not openly notify passengers of their ratings [1].

Theory and Hypothesis

This study aims to test for negative psychological effects associated with the use of peer-to-peer reputation systems. Due to the novelty of reputation systems aimed at consumers, it is expected that users in this group will be
caught off guard when receiving low ratings. In-person interactions are sought in this case, due to the increased amount of personal details exchanged. Such details are thought to influence bias along with the misattribution of low ratings. Uber will serve as a suitable application to study for this purpose, given its widespread use and strong consequences associated with its low ratings [33,34].

Distress and control have been chosen as two likely indicators of a type of negative psychological effect in this context.

**H1. Passengers feel less in control to improve their scores than drivers.**

As opposed to drivers, who are aware that their ratings stem from their driving habits or route chosen, passengers have fewer external factors to which they can attribute their low ratings. Internal factors like personality or looks can be more difficult to change.

**H2. Negative personal ratings cause more acute psychological distress for passengers than for drivers.**

Negative ratings, likely to be perceived as a result of personal characteristics, can be very offensive to passengers and in direct opposition to their identity. Additionally, having provided monetary payment for their service, passengers will feel as though they are the judge, and be internally reluctant to alter behaviors that negatively affect their rating.

**Method**

To measure both participants perception of control and psychological distress in response to their Uber rating, two validated scales were found suitable and adapted for use in this study. The first, a six-item subscale taken from
Lumpkin et al. designed to measure locus of control [16,21]. The second scale was taken from Mardi Horiwitz et al. and designed to measure subjective distress in response to a specific event [9]. Seven items were chosen from the validated distress scale and reworded to measure distress in response to a recent negative Uber rating. Participants who indicated they could not recall a recent negative Uber rating were not administered this part of the survey. The items from both questionnaires were scored on a five-point Likert scale. Questions five and six of the control questionnaire were reverse coded. For the control scale, a lower score indicated a higher perception of control. For the distress scale, a lower score indicated a higher perception of distress.

**Locus of Control Questionnaire:**

1) When I try, I am almost certain that I can improve my rating.

2) Getting a good rating depends upon ability; luck has nothing to do with it.

3) Whatever rating I get is my own doing.

4) Many of the negative ratings people receive are partly due to bad luck.

5) Getting a good rating depends mainly on being in the right place at the right time.

6) Many times I feel that I have little influence over the rating I receive.

**Subjective Distress Questionnaire:**

1) I thought about it when I didn't mean to.

2) I avoided letting myself get upset when I thought about it or was reminded of it.
3) I tried to remove it from memory.

4) I had waves of strong feelings about it.

5) Other things kept making me think about it.

6) I was aware that I still had a lot of feelings about it, but I didn’t deal with them.

7) I tried not to think about it.

Follow up questions were asked regarding participants’ statistical use of the system and their demographics, in order to better understand factors which may influence their responses. The statistical use questionnaire requested participants’ 1) current rating scores 2) time elapsed since last using the system 3) total number of uses, and 4) experience as dual users. Participants were provided a link to reference their user statistics on Uber.

Two different versions of the survey were created on Qualtrics with specific instructions related to each group. The control and distress questionnaire portions of the survey were administered in mixed order. The two surveys were posted as HITS on Amazon Mechanical Turk between 04/09/2019–04/12/2019. Scholarly literature regarding MTurk’s efficacy as a representative sample were reviewed prior [2–4,8]. HIT details were tailored to each group. The requirements set for all Turkers to be eligible to participate were that they be located in the U.S., have greater than 1000 completed HITs, and a greater than 97% HIT approval rate. The following information was used in the driver recruitment HIT:

**Title:** Uber Driver Survey

**Description:** Tell us about your experiences as an Uber driver.
**Survey Link Instructions:** We are conducting an academic survey about P2P social reputation systems. We would like to understand your opinion about Uber driver ratings. You must be an Uber driver to participate. Select the link below to complete the survey. At the end of the survey, you will receive a code to paste into the box below to receive credit for taking our survey.

Further instructions were provided in the consent form at the beginning of each survey. As both passengers and drivers were required to self-identify, verification questions were included to filter for ineligible responses. Ranked ANOVA and T-Tests were run between all variables in the final analysis to test relations.

**Results**

**Demographics and reported statistical use**

We received 143 responses for the passenger survey. Of these, 119 respondents indicated that they were eligible to participate by correctly answering a verification check question of whether they had ever ridden with Uber. Of these, 34 indicated that they could recall receiving a low rating from an Uber driver making them eligible to complete the distress portion of the survey. For race, 68.1% of respondents indicated White, 18.6% Asian, 8.0% Black or African American, 2.7% Other, 1.8% American Indian or Other Alaska Native, and 0.9% Native Hawaiian or Other Pacific Islander. For Ethnicity, 28.3% indicated they were Hispanic or Latino. 57.5% indicated an age range of 25-34, 17.7% of 18-24, 14.2% of 35-44, 6.2% of 45-54, and 4.4% of 55-64.
Respondents in the passenger group had an average Uber rating of 4.52, with 34.2% indicating they had also driven for Uber in the past. When asked how many times total they had ridden with Uber 25.0% responded 5-10 times, 24.2% 10-20 times, 14.2% 1-4 times, 12.5% 10-30 times, 12.5% greater than 50 times, and 11.7% 30-50 times. When asked how long ago was the last time they had ridden with Uber, 35.0% of respondents answered 1-3 weeks ago, 26.7% less than 1 week ago, 15.0% 1 month ago, 13.3% 1-3 months ago, 5.0% 3-6 months ago, and 5.0% > 6 months ago.

We received 143 responses for the driver survey. Of these, 126 respondents indicated that they were eligible to participate by correctly answering a verification check of whether they had ever driven for Uber. Of these, 76 indicated that they could recall receiving a low rating from an Uber driver making them eligible to complete the distress portion of the survey. For race, 67.9 % identified as White, 16% as Asian, 10.5% as Black or African American, 2.8% as American Indian or Alaskan Native, 0.9 % as Native Hawaiian or Other Pacific Islander, and 1.9% as Other. For ethnicity, 36.8% identified as Hispanic or Latino. 60.4% of participants in this group indicated an age range of 25-34, 20% of 18-24, 24.2% of 35-44, and 4.7% of 45-54.

Respondents in the driver group had an average Uber rating of 4.29, with 91.2% indicating they had also ridden for Uber in the past. When asked how many times total they had driven for Uber 23.9% of participants responded 50 times, 26.5% 5-10 times, 0.9% 30-50 times, 18.6% 20-30 times, 18.6% 10–20 times, and 11.5% less than 5 times. When asked how long ago was the last time
they had ridden with Uber, 28.3% of respondents answered less than a week ago, 23.9% 1 month ago, 8.8% 1-3 months ago, 29.2% 1-3 weeks ago, 8% 3-6 months ago, and 1.0% greater than 6 months ago.

**Locus of control and subjective distress questionnaire**

For locus of control, drivers scored an average of 16.4 with a median of 17 and standard deviation of 4.41, while passengers scored an average of 15.3 with a median of 17 and standard deviation of 5.04. For subjective distress, drivers scored an average of 15.0 with a median of 15.0 and standard deviation of 4.95, while passengers scored an average of 15.1 with a median of 14.5 and standard deviation of 5.77. The computed scores between groups had a striking resemblance, not supporting the hypotheses.

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Average</th>
<th>Confidence Interval of Average</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drivers</td>
<td>17</td>
<td>16.4</td>
<td>15.62 to 17.17</td>
<td>4.41</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>Passengers</td>
<td>17</td>
<td>15.3</td>
<td>14.44 to 16.11</td>
<td>5.04</td>
<td>0</td>
<td>27</td>
</tr>
</tbody>
</table>

**Table 1:** Control questionnaire scores for both drivers and passengers.

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Average</th>
<th>Confidence Interval of Average</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drivers</td>
<td>15</td>
<td>15</td>
<td>13.88 to 16.14</td>
<td>4.95</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>Passengers</td>
<td>14.5</td>
<td>15.1</td>
<td>13.05 to 17.07</td>
<td>5.77</td>
<td>7</td>
<td>35</td>
</tr>
</tbody>
</table>

**Table 2:** Distress questionnaire scores for both drivers and passengers.

**Balanced significant relationships**
For Drivers, a ranked ANOVA test showed a strong statistically significant relationship between the number of times Drivers had driven for Uber and their levels of distress, with a P-value of 0.0291 and an Effect Size of 0.427 (Figure 1).

This can be seen in the table below, where the P-value is 0.0291 and the Effect Size (Cohen’s f) is 0.427.

<table>
<thead>
<tr>
<th>Count</th>
<th>Average</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5 times</td>
<td>8</td>
<td>10.9</td>
</tr>
<tr>
<td>10-20 times</td>
<td>13</td>
<td>14.1</td>
</tr>
<tr>
<td>20-30 times</td>
<td>14</td>
<td>14.7</td>
</tr>
<tr>
<td>5-10 times</td>
<td>25</td>
<td>14.8</td>
</tr>
<tr>
<td>&gt; 50 times</td>
<td>16</td>
<td>18.5</td>
</tr>
<tr>
<td>Total (5)</td>
<td>76</td>
<td>15.0</td>
</tr>
</tbody>
</table>

Figure 1: Ranked ANOVA between times total driven for Uber and distress questionnaire scores for the driver group.

Drivers that indicated driving more times with Uber had on average higher distress scores. For passengers, a ranked ANOVA also showed a strong statistically significant relationship between how many times total they had ridden with Uber and their distress score with a P-value of 0.00762, and Effect Size of 0.545 (Figure 2). The longer they had indicated riding with Uber the higher their distress score (and therefore lower perception of distress).

This can also be seen in the table below, where the P-value is 0.00762 and the Effect Size (Cohen’s f) is 0.545.

<table>
<thead>
<tr>
<th>Count</th>
<th>Average</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5 times</td>
<td>8</td>
<td>10.9</td>
</tr>
<tr>
<td>10-20 times</td>
<td>13</td>
<td>14.1</td>
</tr>
<tr>
<td>20-30 times</td>
<td>14</td>
<td>14.7</td>
</tr>
<tr>
<td>5-10 times</td>
<td>25</td>
<td>14.8</td>
</tr>
<tr>
<td>&gt; 50 times</td>
<td>16</td>
<td>18.5</td>
</tr>
<tr>
<td>Total (5)</td>
<td>76</td>
<td>15.0</td>
</tr>
</tbody>
</table>
Imbalanced significant relationships

For drivers a ranked ANOVA test also showed a statistically significant relationship between the last time they had driven for Uber and their Control score with a P-Value of < 0.00001 and an Effect Size of 0.319 (Figure 3). A steadily higher control score (decrease in overall perception of control) was found with increased reported time since driving for Uber. For passengers, a ranked T-
Test showed that those who indicated Hispanic or Latino for their ethnicity tended to have lower distress scores (higher levels of perceived distress) compared to Non-Hispanics or Latinos, with a P-Value of 0.0168 and Effect Size of 0.910 (Figure 4). Ranked ANOVA tests also showed significant relationships for certain passenger ratings and their distress scores. Those with a 5.0 overall rating had among lowest score on the survey (and therefore highest distress). The highest scores ranged from 4.7 to 4.9. Ranked ANOVA tests also returned statistically
Figure 5: Ranked ANOVA between Uber passenger ratings and their distress questionnaire scores.

significant results between age and distress for passengers, with a P-value of 0.0194 and Effect Size of 0.382 (Figure 6). Those in the oldest age range sampled (55-64) had the lowest distress scores (highest perceptions of distress), while those in the 25-34 age group had the highest scores towards negative ratings.

Figure 6: Ranked ANOVA between age and distress questionnaire scores for the passenger group.

Discussion
Driver and passenger scores for both control and distress surveys were very close, with nearly identical averages. Both drivers and passengers indicated a perception of some level of control over their Uber ratings, although minimal, with their average scores being less than two points below a neutral score of 18, out of a maximum of 30 points. Drivers and passengers also both indicated moderate perceptions of distress surrounding negative ratings, with average scores being approximately six points from a neutral score of 21, out of a maximum of 35 points.

From these results, it can be observed that passengers and drivers tested did not exhibit differences in their overall perceptions of control or distress regarding their ratings associated with the app. Drivers and passengers also both exhibited lower perceptions of distress the more times they had used Uber indicating that these perceptions of distress could be the result of users becoming accustomed to the system.

Conversely, drivers but not passengers exhibited lower perceptions of control the longer they had been away from the system. This could possibly indicate a greater negative experience for drivers which is gradually felt as they are removed from the experience. Furthermore, drivers had an average of 0.29 points lower Uber rating than passengers. Although it is not expected for these two groups to have directly comparable ratings, the difference is high and the consequences for both groups are the same in the case their scores drop below a certain level. Also, a higher proportion of drivers indicated that they could recall negative Uber ratings, possibly indicating a greater importance of these ratings to
drivers. Overall, these results point towards a slightly more difficult experience for drivers in maintaining higher ratings.

Unlike drivers, Hispanic and Latino passengers indicated higher levels of perceived distress in response to their negative ratings, which could be the result of real or perceived discrimination. This study found that passengers in the oldest age group exhibited the highest perceptions of distress in response to negative ratings. No significant relationships however were found between passengers’ age and their overall rating score, which could point to a greater difficulty for older passengers in dealing with these negative ratings.

**Implications**

Implications from this study show that while there may exist some differences between passengers and drivers on Uber—in cases such as their overall rating, perception of control with time spent away from the system, and distress in relation to age and ethnicity—they still exhibit surprisingly similar scores for their overall perceptions of distress and control. More research might be useful to understand why ethnic and older passengers have higher perceptions of distress as opposed to drivers, or why drivers show decreased perceptions of control the longer they have been away from the system but not passengers. In addition, it may be beneficial to research ways in increasing users’ perceptions of control in response to the system, as currently levels are close to baseline. Distress towards negative ratings is also moderate but not severe and could be reduced to avoid any potential harmful effects.
**Conclusion**

The main imbalance observed between passengers and drivers is that while drivers indicated more difficult experiences in response to the control of their scores, passengers seemed to have issues with distress in relation to factors commonly associated with bias, such as age and ethnicity. Although some slight negative effects were found overall, there does not appear to be any alarming signs with the use of closed system peer-to-peer reputation systems on user-psyche. Overall, the use of peer-to-peer reputation systems can be improved to provide users a higher level of control and lower distress in response to ratings.

**Limitations**

Participants in this study were tasked with self-identifying. Verification questions posed successfully filtered 16.8% and 11.9% of participants in the passenger and driver group, respectively. Researchers further encountered difficulty recruiting drivers from other verifiable sources due to budgeting constraints. Still important is to investigate distress surrounding the implementation of reputations systems in general, as well as those utilizing public visibility.

**References**


[28] Facebook is judging how trustworthy you are: What you need to know. *USA TODAY*. Retrieved July 2, 2019 from


[32] Background Report: Search, Check & Edit Public Background Reports. Retrieved July 2, 2019 from https://www.mylife.com/?utm_source=google&utm_medium=cpc&utm_campaign=1742049248&gclid=EAIaIQobChMIn7u_rJeW4wIaVDrACh1dSQFBEEAYASAEgKanvD_BwE
