How Unbecoming of You: Gender Biases in Perceptions of Ridesharing Performance

Brad Greenwood
University of Minnesota
wood@umn.edu

Idris Adjerid
Virginia Tech
iadjerid@vt.edu

Corey Angst
University of Notre Dame
cangst@nd.edu

Abstract

It has been suggested that the gig-economy's elimination of traditional arm's-length transactions may introduce bias into perceptions of quality. In this work, we build upon research that has identified biases based on ascriptive characteristics in rating systems, and examine gender biases in ridesharing platforms. In doing so, we extend research to consider not simply willingness to transact, but post transaction perceptions of quality. We also examine which types of tasks may yield more biased ratings for female drivers. We find no differences in ratings across gender in the presence of a high quality experience. However, when there is a lower quality experience, penalties for women accrue faster, notably when poorly performed tasks are perceived to be highly gendered.

1. Introduction

The advent of the Internet and the digitization of commerce have provided more efficient mechanisms by which goods and services are exchanged [1], as well as an improved ways for consumers to voice their opinions about retailers and service providers [2, 3]. Online ratings systems, a key component of matching platforms, have been widely heralded for obviating the *Lemons Market* issues that emerge in markets characterized by a lack of trust and quality uncertainty [4]. Yet, just as evidence is beginning to emerge suggesting that reviews are strongly predictive of sales, increase product salience, and are useful to consumers [2, 5], research has also revealed that bias can emerge during the review process [3, 5, 6].

Concomitantly, digital platforms have increasingly made personal information about transacting parties available, thereby reducing the anonymity that has characterized online transactions. Airbnb and Uber, for example, provide photos, names, and quality information. One might expect that this decreased anonymity may introduce additional bias into perceptions of the quality [7-9]. Yet, as researchers have delved further into this phenomenon, the majority has focused on how factors like race affect the willingness to transact *ex ante*, rather than the actual evaluation of the quality of service. For example, the likelihood of a guest being accepted or an entrepreneur receiving capital based on their name and picture, as opposed to an assessment of the experience or service they receive [10, 11].

We extend this body of research by examining how gender biases in online platforms influence not simply the willingness to transact, but a consumer's evaluation of the service. Further, we examine how these evaluations are moderated by the ratée's historic quality, the ascriptive characteristics of the rater, and various facets of the service provided, (e.g. pickup, navigation). We draw upon a rich literature discussing gender roles and bias [12, 13] and develop theory which posits that because driving is typically a male dominated profession [12], the incongruence with professional roles will cause a significant *a priori* penalty for female drivers. We then argue female drivers will be disproportionately penalized for poorer service. Finally, we decompose the effect and examine which types of service failures are penalized more.

Empirically, we execute a two-phase experiment. In the first phase, we present a mock mobile application, in which the gender and historic quality data about the driver are manipulated. Respondents then proceed to the second phase, where we use a structured narrative to provide a salient experience. This experience may also be of high or low quality. Thus, while Phase 1 is used to establish a baseline of bias, Phase 2 allows us to mimic the decision point of consumers, and assess the degree of bias after a salient transaction. In particular, we assess whether gender biases exist in the *ex ante* perception of driver quality, how quality of the transaction influences bias, and if historic quality of the driver, and/or characteristics of the rater moderate these effects.

Important findings stem from this study. Prior to being exposed to a salient experience with the driver, and conditional on prior quality, gender offers no additional predictive power. Further, we find no evidence of gender bias when the experience is high

URI: https://hdl.handle.net/10125/60092
ISBN: 978-0-9981331-2-6

quality. Yet, as quality deteriorates, the penalty for women is larger, suggesting that errors of attribution may be at play [14]. Interestingly, this effect is primarily driven by Caucasian male raters.

Notable contributions for theory and practice stem from these findings. First, to the degree that prior literature has highlighted the biased nature of online reviews [3], our work provides additional insights into mechanisms which drive such biases, namely errors of attribution [15]. At the same time, the finding that penalties accrues when historical quality is high suggests that providing such data is unlikely to ameliorate the problem, even if it does increase initial willingness to transact [11].

Second, our work begins to push the boundary of bias in management research beyond the traditional workplace. Digital platforms, where buyers and sellers rate each other, are estimated to contribute $335B to the world's economy by 2025 [16], and these new organizational forms create intriguing interpersonal dynamics that warrant the attention of scholars.

2. Related Literature

Since the emergence of the internet and electronic commerce, IS researchers have embraced the topic of user generated content and ratings [2, 5, 17]. Our focus is on biases that exist within the ratings' systems themselves. Two distinct streams of work exist in this domain. The first argues that aspects of the ratings' process might contribute to bias [17]. The second investigates the impact of rater and ratée characteristics on willingness to transact [10, 18, 19].

In the first stream, researchers argue that there are selection issues associated with rating a product online [3]. If a consumer's experience is not notable, then the rater may not feel compelled to inform others of her experience, thus limiting the number of reviews [17]. Further, some consumers may be positively inclined towards a product, thereby creating a selection bias in terms of who has the opportunity to rate. For example, fans of a popular book (e.g., Harry Potter) may be more likely to purchase a sequel. Thus, the quality of the product may be exaggerated, because an excess of consumers who are positively predisposed to the product initially rate it [6]. Finally, there is often an impulse to exaggerate quality at the end of the quality spectrum [3]; which pushes a marginally negative review more negative, or vice versa.

The second stream of literature suggests that factors like race and gender may influence the willingness of agents to interact with each other. Research shows that African-American renters on Airbnb are less likely to be accepted by hosts and more likely to be subject to cancellations [10]; a finding also observed in ridesharing [18] and job search [20]. Similarly, Muslim job applicants are less likely to be called back than identically qualified Christian candidates [19]. Racial and ethnic biases have also been observed against service providers. Research finds that biases exist on crowdfunding websites in the form of discrimination against African-American project founders, evident by a decreased willingness to fund such campaigns [11]. The study closest to our own [18] finds that women who utilize ridesharing services are taken for longer, more expensive trips.

While this research provides critical insights into how ascriptive characteristics influence party willingness to transact *ex ante*, it provides minimal insights into how ratings are be affected by the characteristics of service providers. Coupled with the fact that extant research rejects the notion that simply allowing sub-groups to access markets will ensure equality [7, 12, 13, 21], it is incumbent upon researchers to quantify such biases; not simply because they are unknown, but because such information is critical to the design of effective interventions which may ameliorate such biases.

In what follows, we discuss how literature may inform our understanding of these gaps, both in terms of expectations of performance, and the evaluation of actual performance. In doing so, we focus specifically upon gender biases. We do this for two reasons. First, while gender discrimination has been studied extensively in offline contexts (see [22-24]), limited work has delved into such biases in the gig-economy; with a notable exception [18] that examines the role of gender discrimination in ridesharing, albeit not from a ratings perspective. Second, theoretically, deep streams of literature in psychology, sociology, economics, and organizational theory exist examining perceptions of women in the workplace, as well as perceptions of their performance [22-29]. As a result, we are able to glean insights into how and when women may be more or less subject to bias. Finally, we are able to connect these disparate streams of literature with active research in digital platforms, thereby creating a richer picture of the conditions under which gender discrimination may manifest.

3. Hypothesis Development

3.1. Performance expectations

Why might women be subject to biased expectations of performance in digital platforms? As is well established, the majority of riders and drivers who participate on ridesharing platforms are men [30]. This creates two potential problems for female drivers.

First, women may be cast as a social outgroup, which opens them up to taste based discrimination [31]. Taste based discrimination is premised on the notion that an individual may have a preference, on the margin, for dealing with one group over another (e.g., men over women or Caucasians over African Americans) despite no observable difference in quality. From an economic perspective, this would create an aversion to cross-gender interactions. And, despite criticisms that this irrationality should equilibrate in the long run because markets are competitive, research in the space of workplace discrimination has uncovered many places where bias persists [13, 32]. Moreover, there may be significant ingroup and homophily preferences, where individuals favor those who look and act like them [8, 15].

Second, continuing the logic of an ingroup preference, it could be argued that women entering a field like driving, i.e. a male dominated profession [30], could be seen as violating traditional gender roles [7, 12]. To date, scholars have argued that social perceptions often cast occupations in terms of "men's work" and "women's work" [7, 33]. While this is often seen as an attempt by men to ensure their status within an occupation, it can also be a result of the occupation being male dominated [33]. Empirically, this has been shown in many ways, such as an decreased probability of women being promoted when fewer women hold the sought after position [29] or an embedded belief in gender based qualities which are needed to succeed in an occupation [34]. As a result of perceived lack of fit with the position, i.e. driving, women may be expected to perform at a lower rate [12].

In sum, these two literature streams suggest there might be an intrinsic penalty for female drivers, even prior to observation of quality, despite unambiguous evidence that women are safer drivers than men [35].

Hypothesis 1 (H1): Female gender status will result in lower perceived quality of service, as compared with men, all else equal.

3.2. Evaluation of performance

Inasmuch as ridesharing passengers possess the ability to directly observe the quality of their ride, it is plausible that such biases would be reduced by the resulting amelioration of the information asymmetry which accompanies riding with the driver. Yet, research in social psychology would challenge such a clean economic view of bias in perceptions of quality. Scholars have argued that outgroup biases may manifest in numerous ways, including: employment decisions [22, 23], performance appraisals [14], compensation [36], and ratings of quality [25]. Researchers have also suggested that while members of an ingroup typically do not penalize the outgroup for exceptional or acceptable service [26], they are likely to penalize members of the outgroup more severely for deficiencies in service [22, 28].

What does this mean in the context of online reviews when quality can be observed? Potentially, this suggests that absent anything out of the ordinary about the product or service being rendered, there may be little additional bias in evaluations of service (over H1). However, it also suggests that if there is something notable about the product or service, from a random stroke of luck to some sort of preventable poor service on the part of the driver, women (the outgroup) would be penalized to a greater degree than men (the ingroup) [22, 26, 28].

Hypothesis 2 (H2): Female drivers will be penalized to a greater degree, as compared with male drivers, for performance shortfalls, all else equal.

3.3. Heterogeneity in performance penalty based on task type

While our second hypothesis relates to evaluation penalties which may unduly accrue to women for performance shortfalls [37], our final hypothesis relates to conditions under which women are more likely to be disproportionately penalized [22].

Occupations are often broadly cast as "men's" or "women's" work [7, 22]. Intuitively, this notion of the "gendered" work can be extended to the task itself. For example, although the notion of the "good-provider" role as male has steadily decreased in recent decades, some tasks remain viewed as more feminine (e.g. cleaning, cooking) or masculine (e.g. home repair, yard work) [38]. Even in the workplace, women are often cautioned against "playing house," by providing baked goods or bringing treats, because such actions can lead to feminine traits crowding out perceptions of professional abilities [39]. In the context of ridesharing, these observations are particularly salient. Within the broader occupation of "driver," there are heterogeneous tasks which vary in the degree to which they are gendered. For example, cleanliness of the vehicle, a task traditionally associated with femininity [38], and street smarts, a task traditionally associated with masculinity [40], are both identified by ridesharing firms as critical to receiving top ratings[1].

As a result of disparity in the degree to which tasks are gendered, we propose that women will be more strongly penalized for failing to perform female-gendered tasks well. We also expect to see that females will be rated lower on male-gendered tasks because

[1] https://www.uber.com/drive/philadelphia/resources/5-star-rating-tips/

women persist as the social outgroup of the broader occupation. Put another way, because women are expected to be more competent at traditionally feminized tasks, disconfirmation of this expectation should lead to a greater penalty. Importantly, it is unlikely that similar penalties would accrue for men, because of their status as the social ingroup [14].

Hypothesis 3 (H3): Female drivers will be penalized to a greater degree, as compared with male drivers, for performance shortfalls when performing highly gendered tasks, all else equal.

4. Experiment Overview and Design

We take an experimental approach to identify the biases which may emerge in quality perceptions of platform enabled transactions. Our participants were sampled from Amazon Mechanical Turk (AMT), which has been shown to be at least as representative as other Internet samples, and more representative than student samples [41]. Although a field experiment would be preferable in some respects (e.g., realism), it is difficult to randomly manipulate quality information in a real-world setting, and feasible approaches for doing so introduce significant ethical issues (e.g., purposefully providing a rider a dangerous or low quality experience or inaccurate quality information about their driver).

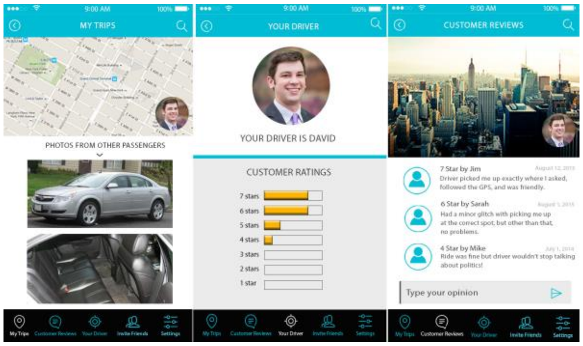
Our experiment employed a 2 (gender) x 2 (race) x 2 (Historical Quality) x 2 (Experience Quality), between-subjects design. Our first two dimensions (gender and race), were manipulated in the study by presenting the subject with driver photographs that varied across gender (Male, Female) and race (Caucasian, African American). We included manipulations of Caucasians and African American as prior work shows significantly different dynamics for African American women vs. white women in the workplace [42, 43].

Consistent with prior literature [25, 27], we manipulate race in order to evaluate robustness of gender effects across racial lines. We manipulated quality by altering the information subjects were given about the driver. Our experiment had two distinct phases and quality was manipulated over both phases in the study. In Phase 1, historical quality was manipulated and subjects were provided an overview of the drivers' past performance. Between subjects, we manipulated whether the driver presented to the raters had high or low historical quality information. In Phase 2, subjects were asked to imagine a detailed experience with the driver (based on another customer's recent experience with the driver) and then update the rating of the driver on the same dimensions from Phase 1. Again, we manipulated whether the rater was presented a high or low quality experience with the driver. Manipulations of race and gender persist through Phase 1 and Phase 2 (i.e., the driver that participants reviewed is the same across phases). Quality, on the other hand, was allowed to change between Phase 1 and Phase 2, since participants were assigned to either high or low historical quality in Phase 1, and then again assigned to either high or low experience quality in Phase 2. In Phase 2, the objective is to determine whether race and/or gender bias emerge in the rating of a single salient ride experience, how the quality of this transaction modifies this bias, and whether high versus low historical quality and characteristics of the rater ameliorate or exacerbates these effects.

4.1. Procedure

Participants were told that we represent a new ride sharing service, called "Agile Rides," and that we are in the process of launching our service. We employed this deception (with IRB approval) to increase the external validity of our experimental setting and have participants believe that their assessments would have real impact. We also created and published a publicly available mock website to further reinforce our existence as a new ride sharing company. Participants were then told that we required their assistance in understanding what makes a good rider experience.

Following this, participants provided general demographic data and answered a series of general questions about their experience with ride sharing services. Participants were then set to begin Phase 1 of the study, in which they were provided information about the driver's gender, race, and aggregate historical quality in three panels (Figure 1). The purpose of Phase 1 was to introduce our various experimental manipulations and establish a baseline rating for each driver before the subject was exposed to any salient information about the ride experience itself. The first panel shows images of the driver's car (interior and exterior) taken by other riders, the second panel shows aggregate rating information for the driver, and the final panel shows three detailed reviews left by other riders of the driver. All panels include an image of the face of the driver. After reviewing the information in the panel, participants are asked to rate the driver (using a seven-star rating scale) on several distinct dimensions (e.g., timeliness, safety, etc.). The participants were then asked to provide an overall rating of the driver. Photos of all drivers are available upon request.



7 Star by Jim — August 12, 2015
Driver picked me up exactly where I asked, followed the GPS, and was friendly.

6 Star by Sarah — August 5, 2015
Had a minor glitch with picking me up at the correct spot, but other than that, no problems.

4 Star by Mike — July 1, 2014
Ride was fine but driver wouldn't stop talking about politics!

Type your opinion

Figure 1. High Historical Quality Driver

Participants then proceeded to Phase 2, where they were asked to imagine going through a detailed customer experience which, they were told, was based on a recent customer experience with that driver. Participants were then asked to rate the driver on the same dimensions as those in Phase 1. In this hypothetical scenario, five dimensions of the ride experience were described to participants: i) pick-up, ii) how luggage was handled, iii) the condition of the car, iv) the driving style of the driver, and v) the route taken. For each of these dimensions, either a high or low quality experience was described (descriptions of the experiences, omitted in the interest of space, are available upon request). Finally, participants answered a number of exit questions, were provided a debrief to inform them that they had just participated in a research study, i.e. that Agile Rides was not a ride sharing firm, and were given the option to exclude their responses from the study without penalty.

4.2. Pre-Studies

Prior to running our main experiment, we ran two additional pre-studies. These were intended to refine and validate the manipulations used in it. In the first pre-study, we sought to identify individuals with faces that there was agreement with the intended race and gender of the driver to avoid introducing unintended bias into the experiment. We also sought to validate that the faces of the individuals used in our manipulations of race and gender were not eliciting unintended differences in other factors (e.g., warmth, professionalism, attractiveness, etc.), which could subsequently bias the results. This was done because extant research highlights the importance of appearance as a powerful behavioral influencer [44].

To accomplish the above validation, we recruited 18 students from a small North American university that were approximately the same age at the time of the study (early 20s) and varied in gender and race. [2] Names for the individuals were chosen from a 2014 online repository of names from Johnson & Johnson. To reduce the bias associated with names, we used the most popular names for African Americans and Caucasians; "David" for men and "Kayla" for women.

We recruited 48 participants from AMT and asked them to provide their input on the students based solely on the student's photograph.[3] From the original 18 student participants, we selected the 8 individuals (2 African American men, 2 Caucasian Men, 2 African American women, and 2 Caucasian women) who had the highest agreement with their intended race and gender (~ 98% agreement for each chosen individual) as well as agreement that the individual was born in the United States (~95%). Moreover, initial perceptions of individuals were found to be nearly identical across all dimensions captured, i.e., individuals rated equally on perceived trustworthiness, kindness, welcoming, and attractiveness.

In the second pre-study, our objective was to validate that the manipulations of high and low quality from the rider's experience were effectively triggering differing perceptions of quality. Recall that we manipulate quality in both Phase 1 and Phase 2 of the experiment. In Phase 1, we manipulate quality in a binary fashion, with participants receiving either a high or low quality driver (Quality = 0,1). This was done by altering the content in each of the panels from Figure 1. In the first panel, the interior of the car was clean and without clutter for the high quality condition. In the low quality condition, a small amount of debris was present. In the center panel, the high quality condition had a top-skewed distribution of reviews with most ratings at 6 or 7 out of 7. In the low quality condition, the driver had a normal distribution with most reviews clustered at 4 or 5 out of 7. In the final panel, the high quality condition had three written reviews with ratings of 7, 6, and 4 stars out of 7. In the low quality condition, the driver had the identical 6 and 4 star reviews, but also had a critical 3-star review in lieu of the 7-star review.[4]

Our intent in Phase 2 of the study was to manipulate experience quality by altering the narrative presented to participants, i.e. the description of the experience of a previous rider. Therefore, in our validation test, it is incumbent upon us to evaluate how

[2] All 18 individuals were professionally photographed (head and shoulders), had nearly identical backdrops in their images, wore semi-professional attire (common for drivers on ridesharing platforms), and were asked to smile (so as to have similar facial expressions).

[3] This type of evaluation of a person based on the presentation of only a photograph is known as a zero-acquaintance study of judgment and its reliability and consistency relative to in-person, face-to-face evaluations has been tested in a variety of contexts .

[4] We avoided manipulations that we perceived as too extreme and thus not believable (e.g., a driver with only 1 or 2 stars, or a filthy and cluttered car). To avoid potential bias, the driver's face was replaced with a gender-neutral silhouette in the pre-study.

introducing negative experiences, with respect to various dimensions of the ride, affected perceptions of quality. To accomplish this, we randomly manipulated (between subjects) each of the five dimensions of quality. Thus, participants in our pre-study were presented with different versions of quality ranging from five negative quality narratives to five positive quality narratives (Quality=1..5).

We recruited 236 subjects to take the study and they either assessed the quality information provided in Phase 1 or Phase 2. We found evidence that our manipulations of quality had the anticipated impact on perceptions of the quality of the driver in both phases. In Phase 1, drivers with “high quality” panels had a significantly higher star rating relative to those with the low quality panels (5.65 vs. 4.37, $t(97)=7.28$, $p<.0001$). Similarly, a higher proportion of positive narratives when describing a ride experience significantly and strongly correlated with a higher overall rating ($p=.8, p<.0001$). Results are confirmed using an OLS (results provided upon request).

4.4. Measures and estimation approach

The main measure of interest in our experiment is the overall rating given to drivers by study participants. To conduct this estimation, we use a triple difference (DDD) model. We estimate this model as an OLS with robust standard errors. Our estimated model is described below:

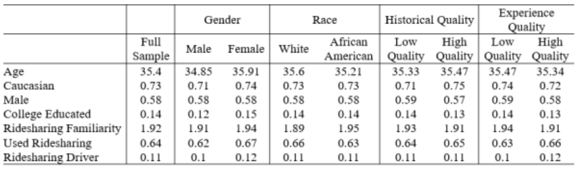
$$\begin{aligned} OverallRating_i = {} & \beta_1*LowQuality_i + \beta_2*AA_i + \\ & \beta_3*Female_i + \beta_4*LowQuality*AA_i + \\ & \beta_5*LowQuality*Female_i + \beta_6*Female*AA_i + \\ & \beta_7*LowQuality*AA_i*Female_i + u_i \qquad (1) \end{aligned}$$

$OverallRating_i$ is a continuous measure from 1-7 that captures the overall star rating given to the driver by a rater *i*. While we also ask participants to evaluate more specific dimensions of the ride (e.g. safety, timeliness), our focus in the analysis is the overall rating given to drivers. *$LowQuality_i$* is a binary indicator for whether the driver presented to the participant was of high or low quality (depending on the phase of the study, the quality may be either be historical or experiential in nature). *AA_i* is a binary indicator for whether the driver was African American (1 – yes / 0 – no), and *$Female_i$* is a binary indicator or whether the driver was female (1 – yes / 0 – no). In this specification, the omitted category (i.e., comparison group) is Caucasian male drivers with high quality. This means that the constant term in all models is interpretable as the average rating provided to Caucasian male drivers of high quality. Thus, β_1 identifies the difference in overall rating when quality is low and the driver is a Caucasian male. β_2 and β_3 identify the difference in overall rating when quality is high and the driver is an African American male or a Caucasian female, respectively. A significant and negative coefficient of β_2 would provide evidence of *H1*, and suggest that women accrue a penalty on account of their gender. β_4 and β_5 are interaction terms, and identify whether the overall rating differs for African Americans men and Caucasian women when quality is low. A significant coefficient of β_5 would provide evidence for *H2*, and suggests that women accrue a more severe penalty when quality is low. β_6 captures any difference in rating for African American women relative to Caucasian women. Finally, β_7 is a triple interaction which captures whether the penalty for low quality differs for African American women. A significant β_7 would suggest a different penalty for African American women while an insignificant coefficient would suggest that African American and Caucasian women accrue this penalty to a similar degree. An insignificant coefficient implies broad support for *H2* and suggests that the observed effect spans both Caucasian and African American women. To evaluate H3, we estimate an identical main model while condition only on observations where the drivers had a low-quality performance on a gendered dimension of the ride.

4.5. Sample

There were 919 participants who completed the full experiment (sample descriptive statistics are provided in Table 1). To ensure high quality date, we utilized validated questions commonly used in experimental research to identify and exclude inattentive participants. Our sample had an average age of 34, was 73% Caucasian, 58% male, and fourteen percent had a college education. Asked to indicate their familiarity with ride sharing services on a Likert scale ranging from 1-Very Familiar to 5-Very Unfamiliar, our sample had a mean of 1.92. Specifically, 86% of our sample indicated being either “Very Familiar” or “Somewhat Familiar” with the ride sharing context. Finally, 11% of our sample were ride sharing drivers themselves. Importantly, we find no significant differences in these demographics across our various manipulations with nearly identical and averages across the main manipulations in our experiment. This suggests that the randomization in our experiment was effective.

Table 1. Descriptive Data



		Gender		Race		Historical Quality		Experience Quality	
	Full Sample	Male	Female	White	African American	Low Quality	High Quality	Low Quality	High Quality
Age	35.4	34.85	35.91	35.6	35.21	35.33	35.47	35.47	35.34
Caucasian	0.73	0.71	0.74	0.73	0.73	0.71	0.75	0.74	0.72
Male	0.58	0.58	0.58	0.58	0.58	0.59	0.57	0.59	0.58
College Educated	0.14	0.12	0.15	0.14	0.14	0.14	0.13	0.14	0.13
Ridesharing Familiarity	1.92	1.91	1.94	1.89	1.95	1.93	1.91	1.94	1.91
Used Ridesharing	0.64	0.62	0.67	0.66	0.63	0.64	0.65	0.63	0.66
Ridesharing Driver	0.11	0.1	0.12	0.11	0.11	0.11	0.11	0.1	0.12

5. Results

We first analyze the impact of race, gender, and quality on the baseline assessments of our drivers in Phase 1 (Table 2, Column 1). In this phase, we introduced our manipulation of race, gender, and high or low historical quality using three panels from our mobile application. We find that, as expected, quality is a strong predictor of the driver's baseline rating ($\beta_{LowQuality}$ = -1.2, p<.01). However, this effect does not seem to differ by gender in the first phase. Specifically, we do not identify a significant coefficient of *Female*, the interaction between *LowQuality* and *Female*, or the three-way interaction between *LowQuality*, *Female*, and *AA* (Column 1). These results suggest that the baseline rating for participants is not being biased by gender. All else equal, this suggests that baseline ratings for all drivers in Phase 1 are only driven by normative factors, viz. quality, and not gender (or racial) biases.

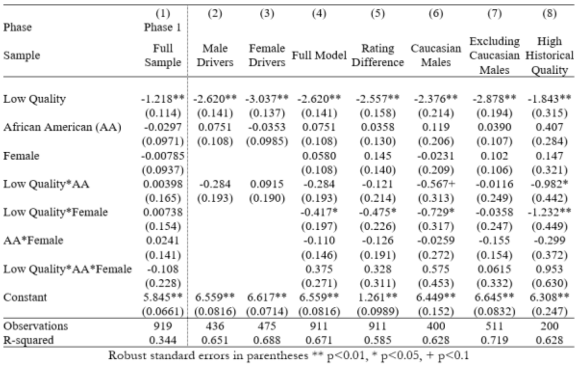
Next, we analyze the ratings of the drivers from Phase 2 (Columns 2-9). Recall, in this phase, participants were provided information on a specific experience with the driver, which they believed was based on a recent customer experience. This experience was then randomly assigned to either a high or low quality manipulation. The race and gender, i.e. the picture, of the drive was held constant across the phases. In this phase, we again find a strong impact of quality for both male ($\beta_{LowQuality}$ = -2.6, p<.01, Columns 2) and female drivers ($\beta_{LowQuality}$ = -3.04, p<.01, Columns 3). We estimate a separate model for males and females to show (via a simple estimation approach) that the penalty for low quality is higher for women relative to men. In this phase female drivers have a higher coefficient on *LowQuality* suggesting that they receive a higher penalty for low quality experience relative to men.

Estimating our full model, we do not find a main effect of *Female* suggesting a lack of a blanket gender bias (i.e. when quality is high). Coupled with the absence of significant *a priori* penalty for female gender status in Phase 1, this suggests negligible support for *H1*. However, we do find significant gender difference ($\beta_{LowQuality*Female}$ = -0.42*, p<.05, Table 2, Column 4) when quality declines. This result indicates the presence of gender bias following a low quality experience, and support for *H2*. In other words, women are penalized to a greater degree than males when quality transgressions occur. The final term (three way interaction between LowQuality, AA, and Female) identifies whether this effect differs for African American women. This coefficient is not significant and suggests a statistically indistinguishable difference in the penalty between Caucasian and African American women. We also assess potential gender bias in the relative change in ratings from Phase 1 to Phase 2. Thus, we revise our dependent variable to be the difference between the rating given to the driver in Phase 1 and Phase 2 (Column 5). We again find consistent results with our main analysis.

Further parsing of our data reveals that Caucasian males (our primary social ingroup) seem to be driving this gender bias in ratings (Columns 6 and 7). Estimating our main model with only Caucasian male raters reveals a larger bias against women if a low quality experience is described ($\beta_{LowQuality*Female}$ = -0.73*, p<.05, Table 2, Columns 6). This suggests that an error of attribution may be occurring because the bias is against an outgroup and accrues only when quality transgressions manifest. This mechanism is corroborated when we focus on Caucasian male raters' perceptions of low quality experiences provided by African American drivers, which reveals some indication of bias against African American males after a low quality experience ($\beta_{LowQuality*AA}$ = -0.57, p<.1, Columns 6).

Next, we analyzed whether these effects would be ameliorated by when the historical quality information was high versus low. In particular, we suspected that Caucasian male raters might present less bias against female drivers if female drivers had a track record of high quality performance on the platform (i.e. where high historical was quality). We find that if a driver had high historical quality and then had a low quality experience, Caucasian male raters still disproportionately punished female drivers with nearly an additional 1.2 star reduction in rating (Column 8). This result suggests that high historical quality is unlikely to ameliorate bias against women emerging from Caucasian male drivers.

Table 2. Gender bias in ratings



	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Phase	Phase 1							
Sample	Full Sample	Male Drivers	Female Drivers	Full Model	Rating Difference	Caucasian Males	Excluding Caucasian Males	High Historical Quality
Low Quality	-1.218**	-2.620**	-3.037**	-2.620**	-2.557**	-2.376**	-2.878**	-1.843**
	(0.114)	(0.141)	(0.137)	(0.141)	(0.158)	(0.214)	(0.194)	(0.315)
African American (AA)	-0.0297	0.0751	-0.0353	0.0751	0.0358	0.119	0.0390	0.407
	(0.0971)	(0.108)	(0.0985)	(0.108)	(0.130)	(0.206)	(0.107)	(0.284)
Female	-0.00785			0.0580	0.145	-0.0231	0.102	0.147
	(0.0937)			(0.108)	(0.140)	(0.209)	(0.106)	(0.321)
Low Quality*AA	0.00398	-0.284	0.0915	-0.284	-0.121	-0.567+	-0.0116	-0.982*
	(0.165)	(0.193)	(0.190)	(0.193)	(0.214)	(0.313)	(0.249)	(0.442)
Low Quality*Female	0.00738			-0.417*	-0.475*	-0.729*	-0.0358	-1.232**
	(0.154)			(0.197)	(0.226)	(0.317)	(0.247)	(0.449)
AA*Female	0.0241			-0.110	-0.126	-0.0259	-0.155	-0.299
	(0.141)			(0.146)	(0.191)	(0.272)	(0.154)	(0.372)
Low Quality*AA*Female	-0.108			0.375	0.328	0.575	0.0615	0.953
	(0.228)			(0.271)	(0.311)	(0.453)	(0.332)	(0.630)
Constant	5.845**	6.559**	6.617**	6.559**	1.261**	6.449**	6.645**	6.308**
	(0.0661)	(0.0816)	(0.0714)	(0.0816)	(0.0989)	(0.152)	(0.0832)	(0.247)
Observations	919	436	475	911	911	400	511	200
R-squared	0.344	0.651	0.688	0.671	0.585	0.628	0.719	0.628

Robust standard errors in parentheses ** p<0.01, * p<0.05, + p<0.1

To assess support for our final hypothesis, the gendered nature of tasks, we evaluate the role of highly gendered tasks in the observed bias against women (Table 3). As a note, the dependent variable in Table 3 is still the overall rating provided to participants. We start by parsing our data by drivers that provide high versus low quality experiences and find consistent results with our prior analysis; the coefficient on *female* is only significant when the experience quality is low (see Columns 1 and 2). Thus, we focus on low quality drivers when evaluating the effect of gendered tasks on this bias. In particular, we evaluate the strength of gender bias when the negative features of the experiences are highly gendered (viz. cleanliness, driving style, and navigation) versus when they are not (viz. efficiency of the pickup and helping with luggage). We find that low quality experiences along highly gendered dimensions of the experience are associated with penalties for women (Columns 3-5). We note that sample size differs between columns because only a subset of the dimensions of the ride experience were negative in the low quality condition. We opted for this approach in order to avoid scenarios that were so negative that they would not be credible. In contrast, when the low quality experiences are along dimensions that are not highly gendered, gender bias disappears (Column 6 and 7). Utilizing a continuous measure ranging from 1, where only one of the dimensions of low quality is highly gendered, to 3, where all three negative dimensions are highly gendered (*Gendered*) supports this finding. Specifically, we find a significant and negative interaction between *Female* and *Gendered* (Column 8). Overall, our results support *H3* and suggest that gender bias emerges when women perform poorly on highly gendered dimensions of the service.

We also consider a series of robustness checks and extensions of our analysis. We find that our results are robust to accounting various features of our rater, including their education levels, age, familiarity with ride sharing. We also find consistent results when we estimate reduced form models that identify the effect of women on average (as opposed to separating the effects for Caucasian and African American women). These analyses are excluded due to space constraints but can be provided upon request.

Table 3. Effect of gendered tasks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Low	Car	Driving				
Sample	High Quality	Quality	Condition	Style	Route	Pickup	Luggage	Gendered
Female	0.0580	-0.359*	-0.540**	-0.525*	-0.542**	-0.006	0.213	0.209
	(0.108)	(0.164)	(0.205)	(0.214)	(0.205)	(0.209)	(0.209)	(0.316)
African American	0.0751	-0.209	-0.300	-0.266	-0.293	-0.093	0.146	-0.195
	(0.108)	(0.160)	(0.204)	(0.193)	(0.192)	(0.228)	(0.224)	(0.160)
Female*AA	-0.110	0.265	0.375	0.345	0.519+	0.0271	-0.468	0.200
	(0.146)	(0.228)	(0.289)	(0.290)	(0.279)	(0.301)	(0.304)	(0.222)
Gendered								-0.211+
								(0.116)
Female*Gendered								-0.285+
								(0.151)
Constant	6.559**	3.939**	4.060**	3.612**	4.076**	3.826**	3.600**	4.331**
	(0.0816)	(0.115)	(0.144)	(0.145)	(0.139)	(0.159)	(0.160)	(0.248)
Observations	462	449	296	235	316	213	205	449
R-squared	0.001	0.012	0.026	0.029	0.021	0.001	0.013	0.059

Robust standard errors in parentheses ** p<0.01, * p<0.05, + p<0.1

6. Discussion and Conclusion

Results from a novel experiment indicate several important findings. Conditional upon inferior service being rendered, women are penalized to a far greater degree than men, particularly by male raters. This penalty accrues notably for highly "gendered" tasks, such as the cleanliness of the vehicle, while men are penalized more uniformly for imperfect service. Further, Caucasian males disproportionately penalize outgroup providers, conditional upon imperfect service. Surprisingly, prior to having an experience with the driver, no bias exists when historical quality information is available. However, when the same raters are presented with a salient experience, bias emerges, but only in low quality situations, suggesting errors of attribution may be key in explaining the observed biases on these platforms.

Notable contributions stem from this observation. Theoretically, we contribute to a rich, but emerging, literature discussing the biases in perceptions of platform based work. We extend extant research in supervisor bias as well. To the degree that many aspects of bias in the manager-subordinate relationship have been investigated, including: gender bias [7], race [9], ingroup biases [15], managerial beliefs [21], and even beliefs about gender roles [13]; it is notable that each of these investigations has occurred in contexts where a traditional manager is evaluating a subordinate. The context of ridesharing and upends this relationship, because the evaluation of the worker (i.e. the driver) is distributed over many evaluators, as opposed to a single person. Thus, it is incumbent upon researchers to consider the biases that these relationships may be subject to, not as a function of micro-foundational interpersonal dynamics, but instead as a function of macro level biases.

This research also has implications for design science work in the form of algorithmic debiasing. Inasmuch as this work has demonstrated proof of concept in many contexts, including medicine [45] and digital design [46], our work highlights a new direction this work should be taken, i.e., the gig-economy. Further, we underscores the importance of researchers moving their findings out of the academic space, and into real time environments.

Finally, this work contributes to the emerging stream of literature discussing the welfare implications of platforms and the digital economy. While such literature has highlighted both positive and negative social outcomes, we advance it by considering how bias may be affecting those who work on these platforms, and what steps can be taken to limit it.

These findings also yield important practical implications. First, following the arguments of Becker [31], the firm puts itself at a strategic disadvantage if it systematically undervalues talent from outgroups (e.g., women). Insofar as ridesharing firms are known to aggressively cull drivers from their ranks, it is possible that competitors may be able to use this indifference towards systemic bias in ratings in order to grow higher quality labor pools at equal or lower costs. Second, despite the fact that the bias we observe originates from a non-employee of the firm, and is directed to a non-employee of the firm, the firm may place itself in a tenuous legal position if it does not intervene to limit the effect of such bias.

In conclusion, despite the overwhelming evidence that online reviews are useful to consumers and can contribute to sales, there is a dark side to rating systems. Where prior research has shown that ingroup members will *attribute* lower quality to ascriptive characteristics of the outgroup, our work goes one step further and empirically demonstrates that prejudiced raters not only attribute poor quality to the minority class to which the driver belongs, but they subsequently penalize the driver by rating them lower after having a salient experience. Further, we find that these penalties are likely to manifest to a greater degree when female drivers are performing highly gendered tasks, suggesting that perceptions of gender roles do exist in these markets.

7. References

[1] Eisenmann, T., G.G. Parker, and M.W. Van Alstyne, *Platform Envelopment.* Strategic Management Journal, 2011. **32**(12): p. 1270-1285.

[2] Forman, C., A. Ghose, and W. B., *Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets.* Information Systems Research, 2008. **19**(3): p. 291-313.

[3] Gao, G., et al., *Vocal Minority and Silent Majority: How Do Online Ratings Reflect Population Perceptions of Quality?* MIS Quarterly, 2015. **39**(3): p. 565-589.

[4] Akerlof, G.A., *The Market for "Lemons": Quality Uncertainty and the Market Mechanism.* Quarterly Journal of Economics, 1970. **84**(3): p. 488-500.

[5] Duan, W., B. Gu, and A.B. Whinston, *Do Online Reviews Matter? An empirical investigation of panel data.* Decision Support Systems, 2008. **45**(4): p. 1007-1016.

[6] Godes, D. and J.C. Silva, *Sequential and Temporal Dynamics of Online Opinion.* Marketing Science, 2012. **31**(3): p. 448-473.

[7] Bielby, W.T. and J.N. Baron, *Men and Women at Work: Sex segregation and statistical discrimination.* American Journal of Sociology, 1986. **91**(4): p. 759-799.

[8] Reskin, B.F., D.B. McBrier, and J.A. Kmec, *The Determinants and Consequences of Workplace Sex and Race Composition.* Annual Review of Sociology, 1999. **25**(1): p. 335-361.

[9] Elliott, J.R. and R.A. Smith, *Race, Gender, and Workplace Power.* American Sociological Review, 2004. **69**(3): p. 365-386.

[10] Edelman, B.G., M. Luca, and D. Svirsky, *Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment.* American Economic Journal: Applied Economics, 2017. **9**(2): p. 1-22.

[11] Younkin, P. and V. Kuppuswamy, *The Colorblind Crowd? Founder race and performance in crowdfunding.* Management Science, 2017. **Articles in Advance**: p. 1-19.

[12] Eagly, A.H., *Sex Differences in Social Behavior: A Social-role interpretation.* 2013: Psychology Press.

[13] Eagly, A.H. and S.J. Karau, *Role Congruity Theory of Prejudice Toward Female Leaders.* Psychological Review, 2002. **109**(3): p. 573-598.

[14] Park, S.H. and J.D. Westphal, *Social Discrimination in the Corporate Elite: How Status Affects the Propensity for Minority CEOs to Receive Blame for Low Firm Performance.* Administrative Science Quarterly, 2013. **58**(4): p. 542-586.

[15] Allport, G.W., *The Nature of Prejudice*. 25th Anniversary ed. 1979, New York: Basic Books.

[16] Hawksworth, J. and R. Vaughan, *The Sharing Economy - Sizing the revenue opportunity*. 2014, PricewaterhouseCoopers.

[17] Dellarocas, C. and R. Narayan, *A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth*. Statistical Science, 2006. **21**(2): p. 277-285.

[18] Ge, Y., et al., *Racial and Gender Discrimination in Transportation Network Companies*. 2016, NBER Working Paper No. 22776

[19] Acquisti, A. and C.M. Fong, *An Experiment in Hiring Discrimination via Online Social Networks.* SSRN 2031979, 2015: p. 1-36.

[20] Bertrand, M. and S. Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A field experiment on labor market discrimination.* American Economic Review, 2004. **94**(4): p. 991-1013.

[21] Carnahan, S. and B.N. Greenwood, *Managers' Political Beliefs and Gender Inequality Among Subordinates: Does His Ideology Matter More than Hers?* Administrative Science Quarterly, 2018. **63**(2): p. 287-322.

[22] Davison, H.K. and M.J. Burke, *Sex discrimination in simulated employment contexts: A meta-analytic investigation.* Journal of Vocational Behavior, 2000. **56**(2): p. 225-248.

[23] Koch, A., S.D. D'Mello, and P.R. Sackett, *A Meta-Analysis of Gender Stereotypes and Bias in Experimental Simulations of Employment Decision Making.* Journal of Applied Psychology, 2015. **100**(1): p. 128-161.

[24] Ayres, I. and P. Siegelman, *Race and Gender Discrimination in Bargaining for a New Car.* American Economic Review, 1995. **85**(3): p. 304-321.

[25] Swim, J., et al., *Joan McKay versus John McKay: Do gender stereotypes bias evaluations?* Psychological Bulletin, 1989. **105**(3): p. 409-429.

[26] Wallston, B.S. and V.E. O'Leary, *Sex Makes a Difference: Differential perceptions of women and men*, in *Review of Personality and Social Psychology*, L. Wheeler, Editor. 1981, Sage Publications: Newbury Park, CA. p. 9-41.

[27] Unger, R.K., *Male is Greater than Female: The socialization of status inequality*. The Counseling Psychologist, 1976. **6**(2): p. 2-9.

[28] Nieva, V.F. and B.A. Gutek, *Sex Effects on Evaluation.* Academy of Management Review, 1980. **5**(2): p. 267-276.

[29] Cohen, L.E., J.P. Broschak, and H.A. Haveman, *And then there were more? The effect of organizational sex composition on the hiring and promotion of managers.* American Sociological Review, 1998. **63**(5): p. 711-727.

[30] Hall, J.V. and A.B. Krueger, *An Analysis of the Labor Market for Uber's Driver-Partners in the United States*. 2015, Princeton University. Industrial Relations Section: Princeton, NJ.

[31] Becker, G.S., *The Economics of Discrimination*. 2nd ed. 1971, Chicago, IL: University of Chicago Press.

[32] Bolzendahl, C. and D.J. Myers, *Feminist Attitudes and Support for Gender Equality: Opinion change in women and men, 1974-1998.* Social Forces, 2004. **83**(2): p. 759-790.

[33] Britton, D.M., *The epistemology of the gendered organization.* Gender & society, 2000. **14**(3): p. 418-434.

[34] Gorman, E.H., *Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms.* American Sociological Review, 2005. **70**(4): p. 702-728.

[35] Li, G., et al., *Are female drivers safer? An application of the decomposition method.* Epidemiology, 1998: p. 379-384.

[36] Westphal, J.D. and P. Khanna, *Keeping Directors in Line: Social distancing as a control mechanism in the corporate elite.* Administrative Science Quarterly, 2003. **48**(3): p. 361-398.

[37] Castilla, E.J. and S. Benard, *The paradox of meritocracy in organizations.* Administrative Science Quarterly, 2010. **55**(4): p. 543-676.

[38] Cohn, S., *The Process of Occupational Sex-Typing*. 1985, Philadelphia, PA: Temple University Press.

[39] Casserly, M. *Playing House In The Office: The Cookie Conundrum*. 2012 [cited 2017 November 21]; Available from: https://tinyurl.com/forbescasserly.

[40] Uhlmann, E.L. and G.L. Cohen, *Constructed Criteria: Redefining merit to justify discrimination.* Psychological Science, 2005. **16**(6): p. 474-480.

[41] Buhrmester, M., T. Kwang, and S.D. Gosling, *Amazon's Mechanical Turk a New Source of Inexpensive, yet High-Quality, Data?* Perspectives on Psychological Science, 2011. **6**(1): p. 3-5.

[42] Todd, A.R., K.C. Thiem, and R. Neel, *Does Seeing Faces of Young Black Boys Facilitate the Identification of Threatening Stimuli?* Psychological Science, 2016. **27**(3): p. 384-393.

[43] Galinsky, A.D., E.V. Hall, and A.J. Cuddy, *Gendered Races: Implications for interracial marriage, leadership selection, and athletic participation.* Psychological Science, 2013. **24**(4): p. 498-506.

[44] Todorov, A., et al., *Inferences of Competence from Faces Predict Election Outcomes.* Science, 2005. **308**(5728): p. 1623-1626.

[45] O'Malley, A.J., et al., *Case-Mix Adjustment of the CAHPS® Hospital Survey.* Health Services Research, 2005. **40**(6): p. 2162-2181.

[46] Schneider, C., M. Weinmann, and J. vom Brocke, *Choice Architecture: Using Fixation Patterns to Analyze the Effects of Form Design on Cognitive Biases*, in *Information Systems and Neuroscience*. 2015, Springer. p. 91-97.