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

Brad N Greenwood
Carlson School of Management
University of Minnesota

Idris Adjerid **Corey M Angst**
Mendoza College of Business
University of Notre Dame

Abstract

The advent of the Internet and the digitization of commerce have provided both a mechanism by which goods and services can be exchanged, as well as an efficient way for consumers to voice their opinions about retailers, i.e. online rating systems. Yet, recent work has begun to uncover significant biases that manifest during the review process. In particular, it has suggested that the gig-economy's elimination of arm's-length transactions may further 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 extant research to consider not simply willingness to transact, but post transaction perceptions of quality. Further, we 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, markedly worse ratings accrue for females. These penalties are exacerbated when female drivers are performing tasks which are perceived to be highly gendered.

Key Words: *Online Rating Systems, Electronic WOM, Gender, Experiment, Bias*

Introduction

The advent of the Internet and the digitization of commerce have provided both a more efficient mechanism by which goods and services are exchanged (Eisenmann et al. 2011; Parker and Van Alstyne 2005; Parker et al. 2016), as well as an improved way for consumers to voice their opinions about retailers and service providers (Brynjolfsson et al. 2003; Clemons et al. 2006; Forman et al. 2008; Gao et al. 2015; Gerstner et al. 1994; Kuruzovich et al. 2008; Mudambi and Schuff 2010). Indeed, online ratings systems, a key component of matching platforms, have been widely heralded for obviating the *Lemons Market* problem that emerges in markets characterized by a lack of trust and quality uncertainty (Akerlof 1970). Yet, just as evidence is beginning to emerge suggesting that reviews are strongly predictive of sales (Dellarocas et al. 2007; Forman et al. 2008; Zhu and Zhang 2010), increase product salience (Duan et al. 2008; Goes et al. 2014), and are useful to consumers (Chatterjee 2001), research has also revealed that significant bias can emerge during the review process (Duan et al. 2008; Gao et al. 2015; Godes and Silva 2012).

Concomitant with an increased interest in quantifying the extent of bias in online reviews, a change in the nature of digital transactions has occurred. In particular, digital platforms have increasingly made personal information about transacting parties available, thereby reducing the anonymity that has traditionally characterized online transactions (e.g., those conducted at an arm's length on platforms like eBay or Amazon). This phenomenon can be observed on a variety of digital platforms (e.g., Airbnb, Kickstarter, Uber), which provide photos, videos, names, and quality information to participants. Thus, one might expect that this decreased anonymity may introduce additional bias into perceptions of the quality (Bielby and Baron 1986; Devine 1989; Dezso et al. 2013; Elliott and Smith 2004; Lowery et al. 2001; Reskin et al. 1999), an observation which has been made by several teams of researchers (Edelman et al. 2017; Ge et al. 2016; Younkin and Kuppuswamy 2017). Yet, as researchers have delved further into this phenomenon, quantifying the extent of discrimination in online platforms (Edelman et al. 2017) and the mechanisms by which it occurs (Younkin and Kuppuswamy 2017), the majority of research has focused on how factors like race affect the willingness to transact *ex ante*, rather than the actual evaluation of the quality of service. Edelman et al. (2017) and Younkin and Kuppuswamy (2017), for example, examine 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. Ge et al. (2016), in closely related work examining ridesharing services, find that female riders had to wait longer, had their rides cancelled more frequently, and were taken on more expensive routes than riders with male sounding names; but not the ratings female participants receive.

In our work, 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 rendered. Further, we examine how these evaluations are moderated by the ratée's historic quality, the ascriptive characteristics of the rater, and the various facets of the service provided, (e.g. pickup, navigation, etc.). In doing so, we draw upon a rich literature discussing gender roles and bias (Baron et al. 1991; Eagly 2013; Eagly and Karau 2002; Ely 1995). We then develop theory which first posits that because driving is typically a male dominated profession (Eagly 2013), notably on ridesharing platforms (Hall and Krueger 2015), the incongruence with professional roles will cause a significant *a priori* penalty for female drivers. We then argue, because eschewed gender roles result in women being a social outgroup, female drivers will be disproportionately penalized for poorer levels of service, as compared with males. Finally, we decompose these effects and examine which types of service failures are penalized to a greater degree, and which characteristics of women relate to increased penalization. We believe such extensions are critical, because an undue focus on ensuring women have access to these markets does not guarantee they will be fairly evaluated (Lee and Huang 2017), especially when obfuscation tactics (e.g. whitewashing) are unavailable (Younkin and Kuppuswamy 2017).

To empirically validate our theory, we execute a two-phase experiment. With IRB permission, we present a new ridesharing service, Agile Rides, to participants and ask them to review our drivers. Participants were told that their ratings would influence the performance evaluation of the driver. In the first phase, we present a mock mobile application, in which the gender and historic quality data about the driver are manipulated; thereby establishing a baseline assessment of the driver. 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, as well as provide information on gender and historical quality, Phase 2 allows us to mimic the decision point of consumers of the service, 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 the quality of this transaction influences this bias, and if historic quality of the driver, and/or characteristics of the rater moderate these effects. The experimental approach, which draws from seminal research in the psychology domain (Goldberg 1968) offers us significant benefits over other alternatives, such as a qualitative or secondary data (Hekman et al. 2010). Beyond problems of accessing secondary data, there are also significant endogeneity concerns and the inability to rigorously observe and quantify "true quality." Similarly, the qualitative approach, in which researchers immerse themselves into the context, creates concerns of stereotype threat and interpretation biases (Aronson et al. 1999; Steele and Aronson 1995). The experimental approach resolves these issues by allowing us to rigorously control for quality while randomly assigning other factors.

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. This lends credence to the claims of recent researchers that the observation of historic quality may mitigate *a priori* biases against social outgroups (Younkin and Kuppuswamy 2017). Further, we find no evidence of gender bias when the experience is high quality. Yet, as quality deteriorates, the penalty for women is disproportionately larger than it is for Caucasian men, suggesting that errors of attribution may be at play (Park and Westphal 2013). Interestingly, this effect is primarily driven by Caucasian male raters.

We also observe that, for female drivers, the type of quality transgression significantly affects the ratings they receive. For example, women are not disproportionately penalized for issues related to the pickup or handling of luggage (i.e. evaluation criteria with no historic relationship with gender). However, they are penalized significantly more for transgressions related to the cleanliness of their vehicle, their style of driving, and their ability to navigate (i.e. roles which are strongly gendered or subject to gender stereotype). Moreover, when considering the potential for high historical quality to ameliorate this effect, we find that women continue to be penalized for lower quality experiences even when historical quality is high. This result lends support to prior work showing that when women demonstrate success in gendered-roles, it violates gender stereotypes (Unger 1976), resulting in less favorably ratings than equally qualified males (Nieva and Gutek 1980; Wallston and O'Leary 1981).

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 (Gao et al. 2015), our work provides additional insights into mechanisms which drive such biases. In particular, gender bias emerging exclusively for low quality experiences suggests that errors of attribution may be key in driving the observed effects (Allport 1979; Pettigrew 1979). At the same time, the finding that this penalty accrues even when historical quality is high suggests that providing such information is unlikely to ameliorate the problem, even if it does increase initial willingness to transact (Younkin and Kuppuswamy 2017). In particular, we can expect the reputation of high quality women to diminish more rapidly on these platforms, assuming that the frequency of low quality transaction is equally likely across all high quality individuals. These results are critical because, although scholars have highlighted many reasons why the review generation process may yield bias (Gao et al. 2015; Godes and Silva 2012), extant research provides few insights into why characteristics of the rater or ratée may influence reviews themselves. Our research augments the work of scholars who have documented disparities that exist in the digital economy (Edelman and Luca 2014; Edelman et al. 2016; Rhue and Clark 2016) by providing insights into why such effects occur, and how they affect evaluations after transaction.

Second, our work begins to push the boundary of bias in management research beyond a traditional workplace setting. Digital platforms, where buyers and sellers can rate each other, are estimated to contribute almost $335B to the world's economy by 2025 (Hawksworth and Vaughan 2014), and these new organizational forms create intriguing interpersonal dynamics that warrant the attention of scholars. In particular, the differential nature of evaluation that occurs on these platforms, as compared with traditional employment, may be problematic. In traditional research and practice, the supervisor has the unique ability to affect a subordinate's career, in the form of a direct evaluation of performance. Inherent within this relationship is the possibility that biases will surface. This stands in contrast to online ratings, where many users have the ability to exert a small amount of power over *many* service providers. As a result, expressions of bias or discrimination is costless for the rater (Becker 1971; 1976; Guryan and Charles 2013), because no avenues for legal reciprocity currently exist in the platform-economy. Considering that these platforms provide access to gainful employment for the un- or under-employed (Burtch et al. 2016; Hall and Krueger

2015), it is concerning that bias emerges most often in under-represented classes.

Finally, from a practical standpoint, these results highlight the need for digital platforms to begin to investigate debiasing procedures (Lee et al. 2015). To the degree that researchers in medicine (O'Malley et al. 2005) and digital design (Schneider et al. 2015) have provided proof that this can be done, it underscores the need to push the boundaries of debiasing practices outside of the academic space (Lee et al. 2015). Moreover, emerging firms, such those participating in the gig-economy, could find themselves in a legally actionable position by failing to take steps to reduce inequalities for protected classes, despite the fact that the prejudicial act (i.e., the biased rating) was against a non-employee by a non-employee (Aloisi 2015; Flake 2016).

In what follows, we provide a review of related literature and develop our hypotheses. This is followed by our experimental overview and design. We then describe our measures, estimation approach, and results. We verify our results by performing several robustness checks and empirical extensions. Finally, we conclude with a discussion of our results and offer guidance for future studies.

Related Literature

Since the emergence of the internet and electronic commerce, information systems researchers have embraced the topic of user generated content and ratings (Chatterjee 2001; Chen and Xie 2008; Chevalier and Mayzlin 2006; Clemons et al. 2006; Dellarocas 2003; Dellarocas et al. 2010; Dellarocas and Narayan 2006; Duan et al. 2008; Forman et al. 2008; Godes and Mayzlin 2004; Harrison-Walker 2001; Hennig-Thurau et al. 2012; Trusov et al. 2009; Zhu and Zhang 2010). Traditionally, research in this domain has followed four related, but conceptually distinct, paths: i) the process by which ratings are generated, ii) rating dynamics over time, iii) the effect of ratings on performance, and iv) biases that exist within the ratings' systems themselves. We do not exhaustively review this literature (for a recent review see Gao et al. (2015)), but instead provide a targeted overview that focuses on aspects which relate to bias in ratings. Specifically, falling under the fourth category where bias has been observed, we address two distinct streams of work. The first argues that aspects of the ratings' process might contribute to bias (Dellarocas and Narayan 2006; Li and Hitt 2008; Richins 1983). The second investigates the impact of rater and ratée characteristics on willingness to transact (Acquisti and Fong 2015; Edelman et al. 2016; Ge et al. 2016; Rhue and Clark 2016).

In the first stream, researchers argue that there are selection issues associated with rating a product online (Li and Hitt 2008). If a consumer's experience is not particularly notable, i.e., does not violate expectations in an overtly positive or negative way, then the rater may not feel compelled to inform others of their experience, thus limiting the number of reviews (Dellarocas and Narayan 2006; Richins 1983). In addition, consumers may be positively inclined towards a product *ex ante*, thereby creating a selection bias in terms of who has the opportunity to rate. For example, fans of a popular book or film series (e.g., Harry Potter) may be more likely to purchase a sequel than a consumer who has no knowledge of the series. As a result, the quality of the product may be exaggerated, as compared with true quality, because an excess of consumers who are positively predisposed to the product initially rate it (Godes and Silva 2012). Finally, there is often an impulse to exaggerate quality at the end of the quality spectrum (Gao et al. 2015); which pushes a marginally negative review more negative, or vice versa.

Compelling research in the second stream of literature, i.e. bias based on the ascriptive characteristics of the participants, is emerging, and suggests that factors like race and gender may influence the willingness of agents to interact with each other. Edelman et al. (2016), for example, find that African-American renters on the popular homestay network Airbnb are less likely to be accepted by hosts and more likely to be subject to cancellations; a finding also observed in ridesharing (Ge et al. 2016), consumer-to-consumer sales (Doleac and Stein 2013), and job search (Bertrand and Mullainathan 2004). Similarly, Acquisti and Fong (2015) find that Muslim job applicants are less likely to be called back for a job than identically qualified Christian candidates. Distressingly, racial and ethnic biases have also been observed against service providers as well. Rhue and Clark (2016) and Younkin and Kuppuswamy (2017), for example, find that biases exist on the crowdfunding website Kickstarter in the form of discrimination against African-American project founders evident by a decreased willingness to fund such campaigns. In the work closest to our own, Ge et al. (2016) find that women who utilize ridesharing services are taken for longer, more expensive trips, indicating that the service they receive may be exploitive in the absence of guaranteed pricing.

While this research provides critical insights into how ascriptive characteristics influence the willingness of parties to transact *ex ante*, it provides minimal insights into how ratings might be affected by the ascriptive characteristics of platform service providers. Understanding such differences is crucial. While prior research has mostly focused on access to markets, notably for African Americans, limited information is known about the biases subgroups may be subject to once access to the market is gained. Coupled with the fact that extant research strongly rejects the notion that simply allowing sub-groups to access markets will ensure equality (Baron and Bielby 1980; Baron et al. 1991; Bielby and Baron 1986; Carnahan and Greenwood 2017; Eagly 2013; Eagly and Karau 2002; Hekman et al. 2010; Roth 2004), it is incumbent upon researchers to quantify such biases; not simply because they are empirically unknown, but because such information is critical to the design of effective interventions which may ameliorate such biases. Moreover, while received research provides extensive insights into the biases social outgroups might face in the form of hiring, wage allocation, or promotion (Bell 2005; Cardoso and Winter-Ebmer 2010; Castilla and Benard 2010; Davison and Burke 2000; Swim et al. 1989), and also the penalties associated with social outgroups performing what are perceived to be ingroup tasks (Koch et al. 2015; Nieva and Gutek 1980; Olian et al. 1988; Wallston and O'Leary 1981), more work needs to be done in order to address how online markets contribute to (or diminish) these biases. In short, understanding such delineations is critical for effective intervention.

In what follows, we develop theory which explicitly discusses how extant literature may inform our understanding of these gaps, both in terms of expectations of performance (i.e. before the observation of quality), and the evaluation of actual performance (i.e. after quality has been observed). In doing so, we focus specifically upon gender biases, as opposed to biases manifesting based on age, race, or national origin. We do this for two reasons. First, while gender discrimination has been studied extensively in offline contexts (see (Ayres and Siegelman 1995; Davison and Burke 2000; Koch et al. 2015; Olian et al. 1988)) and in peer-to-peer lending (Barasinska and Schäfer 2014; Duarte et al. 2012; Pope and Sydnor 2011; Ravina 2012)), limited work has delved into such biases in the gig-economy; with the notable exception of Ge et al. (2016) who examined the role of gender discrimination in ridesharing albeit not from a ratings perspective. This allows us to broaden the corpus of literature which actively discusses issues of bias in online markets. Second, from a theoretical perspective, deep streams of literature in psychology, sociology, economics, political science, and organizational theory exist examining perceptions of women in the workplace, as well as perceptions of their performance (Ayres and Siegelman 1995; Cohen et al. 1998; Davison and Burke 2000; Koch et al. 2015;

Nieva and Gutek 1980; Olian et al. 1988; Swim et al. 1989; Unger 1976; Wallston and O'Leary 1981). As a result, we are able to glean deep 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.

Hypothesis Development

Performance Expectations

Why might women be subject to biased expectations of performance in digital platforms? Intuitively, it is plausible that women might be able to capitalize on these markets to a far greater extent than men. To the degree that women are more likely to shoulder domestic responsibilities, such as raising children or maintaining the home (Bolzendahl and Myers 2004; Davis and Greenstein 2009), it is plausible that women would benefit to a greater degree from the flexible working arrangements afforded by digital platforms. Received research supports such a claim, finding that women are more likely to favor flexible working schedules when engaging in contract or on-demand work (Ellingson et al. 1998), notable when they are balancing work and family roles or working around the demands of being a primary caregiver (Frone et al. 1992; Loscocco 1997). Yet, despite such potential benefits, empirical work surrounding the gig-economy indicates that the majority of both riders and drivers (viz. raters and ratées) who participate on ridesharing platforms are men (Hall and Krueger 2015). This creates two potential problems for female drivers.

First, it has the potential to cast women as a social outgroup, which opens them up to issues of taste based discrimination (Becker 1971). Taste based discrimination is premised on the notion that a manager, individual, or customer 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 because it would be more costly for the manager to hire a member of the discriminated class (Charles and Guryan 2008). And, despite criticisms that this irrationality should equilibrate in the long run because markets are competitive (Arrow 1972), research in the space of workplace discrimination has uncovered many places where bias perpetuates. Eagly and Karau (2002), for example, argue that beliefs about gender roles and stereotypes still exist in the modern workforce; and that many people still believe women lack the mindset for certain roles (such as leadership (Bolzendahl and Myers 2004; Ridgeway 1997)). Moreover, there may be significant ingroup and homophily preferences, where individuals favor those who look and act like them (Allport 1979; Reskin et al. 1999; Tajfel and Turner 1979).

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 (Hall and Krueger 2015), could be seen as violating traditional gender roles (Bielby and Baron 1986; Eagly 2013). To date, many scholars in sociology and organizational theory have argued that social perceptions often cast occupations in terms of "men's work" and "women's work" within the organization (Bielby and Baron 1986; Britton 2000), or within society as a whole. While this is often seen as an attempt by men to ensure their status within an occupation, i.e. inflate occupational security, it can also be a result of the occupation being male dominated (Britton 2000), and thereby reliant on masculine tendencies in order to be done proficiently. 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 (Cohen et al. 1998) or an embedded belief in gender based behavioral qualities which are needed to succeed in an occupation (Ely 1995; Gorman 2005). And, as a result of perceived lack of fit with the position, i.e. driving, it is plausible that women will be expected to perform at a lower rate (Eagly 2013; Eagly and Diekman 2012; Heilman and Eagly 2008).

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 (Li et al. 1998). The popular press further attests to such stereotypes, with most of the population firmly believing that men are superior drivers, despite their greater willingness to speed and drive under the influence of alcohol (Hartocollis 2010; Sunderland 2017). Thus, to the degree that women may be perceived to lack the fit and temperament to be a skilled driver, and to the extent that a driving as a profession eschews traditional gender roles, it is plausible that women may be anticipated to perform worse than men, all else equal.

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

Evaluation of Performance

As discussed in H1, traditional economic and sociological thought suggests that prejudicial bias emerges because of information asymmetries between transacting parties (Altonji and Pierret 2001; Arrow 1998), such as a perceived fitness for a role. Thus, inasmuch as ridesharing passengers possess the ability directly observe and evaluate 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 and organizational theory would challenge such a clean economic view of bias in perceptions of quality. For example, scholars have argued that outgroup biases may manifest in numerous ways, including, but not limited to: employment decisions (Davison and Burke 2000; Koch et al. 2015), performance appraisals (Park and Westphal 2013), compensation (Westphal and Khanna 2003), and ratings of quality (Goldberg 1968; Swim et al. 1989). In addition, researchers have suggested that although members of an ingroup typically do not penalize members of an outgroup for exceptional or acceptable service (Wallston and O'Leary 1981), they are likely to penalize members of the outgroup more severely, as compared with members of the ingroup, for deficiencies in service (Davison and Burke 2000; Nieva and Gutek 1980).

What does this mean in the context of online reviews when quality can be observed? Potentially, this suggests that absent anything notable or 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 out of the ordinary about the product or service, from a random stroke of luck or misfortune 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) (Davison and Burke 2000; Nieva and Gutek 1980; Wallston and O'Leary 1981). Take, for example, the case of Uber, where ratings overwhelmingly follow a J-shaped distribution (Hall and Krueger 2015) with few 1-star ratings and many 5-star ratings. It is plausible, in these contexts, that each driver receives a 5 out of 5-star rating on the overwhelming majority of trips (conditional upon the trip not being notable). Yet, if a woman driver provides poor quality service (whether it was in her control or not), she may be penalized more than male drivers because ridesharing firms (i.e. service firms such as Uber or Lyft) lack unambiguous performance standards, i.e. ratings are ultimately subjective (Crandall and Eshleman 2003). Therefore, we propose:

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

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 (Castilla and Benard 2010), our final hypothesis relates to conditions under which men and women are likely to be penalized equally, or disproportionately, for performing specific gender-stereotype tasks (Davison and Burke 2000). Our arguments integrate insights from extant literature on gender roles as well as literature discussing outgroup biases (Baron et al. 1991; Britton 2000; Davison and Burke 2000; Eagly 2013; Heilman and Eagly 2008). We then propose that biases in evaluation will most prominently manifest when female drivers underperform in tasks which are highly "gendered" (Britton 2000) or where there is a greater expectation of a woman's ability to succeed (Davison and Burke 2000; Oliver 1977).

As discussed, occupations are often broadly cast as "men's" or "women's" work by broader society (Bielby and Baron 1986; Davison and Burke 2000; Hartnett and Bradley 1986; Heilman 2001). And, intuitively, this notion of the "gendered" occupation can be extended to the task itself. For example, although the notion of the "good-provider" role as male within the family unit has steadily decreased over the past several decades, some tasks remain viewed as more feminine (e.g. cleaning, cooking, grocery shopping) or masculine (e.g. home repair, financial tasks, yard work) (Cohn 1985; Keith and Schafer 1980; Perry-Jenkins and Crouter 1990). Even within 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 (Casserly 2012). 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 (Cohn 1985), and street smarts, a task traditionally associated with masculinity (Uhlmann and Cohen 2005), 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, as compared with their male counterparts. Importantly, we would also expect to see that females will be rated lower on male-gendered tasks as well,

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

because 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 (Bhattacherjee 2001; Oliver 1977). Further, because women often accrue additional penalties for performing traditionally male tasks (Eagly 2013; Eagly and Karau 2002; Heilman and Eagly 2008), notably when they are performed poorly (Lee and Huang 2017) it is likely that disproportionate penalty will accrue for these tasks as well. Importantly, it is unlikely that similar penalties would accrue for men, because of their status as the social ingroup (Park and Westphal 2013). In other words, although it is likely that men would be penalized for shortfalls in performance, it is unlikely to be undue based on gendered nature of the task because the occupation itself is inherently masculine (Eagly and Karau 2002; Heilman and Eagly 2008; Ross 1977), thus protecting their role as a member of the ingroup. Therefore, we propose:

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.

Experiment Overview and Design

As discussed, 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). Prior research has validated AMT samples as at least as representative as other Internet samples, and significantly more representative than student samples (Buhrmester et al. 2011). Not only have central findings in IS and the decision sciences been replicated using AMT samples (Goodman et al. 2013; Steelman et al. 2014), but AMT offers an effective payment and reputation management system. This offers researchers the ability to exclusively sample participants of higher quality, thereby ensuring superior data integrity; see Peer et al. (2014). 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).

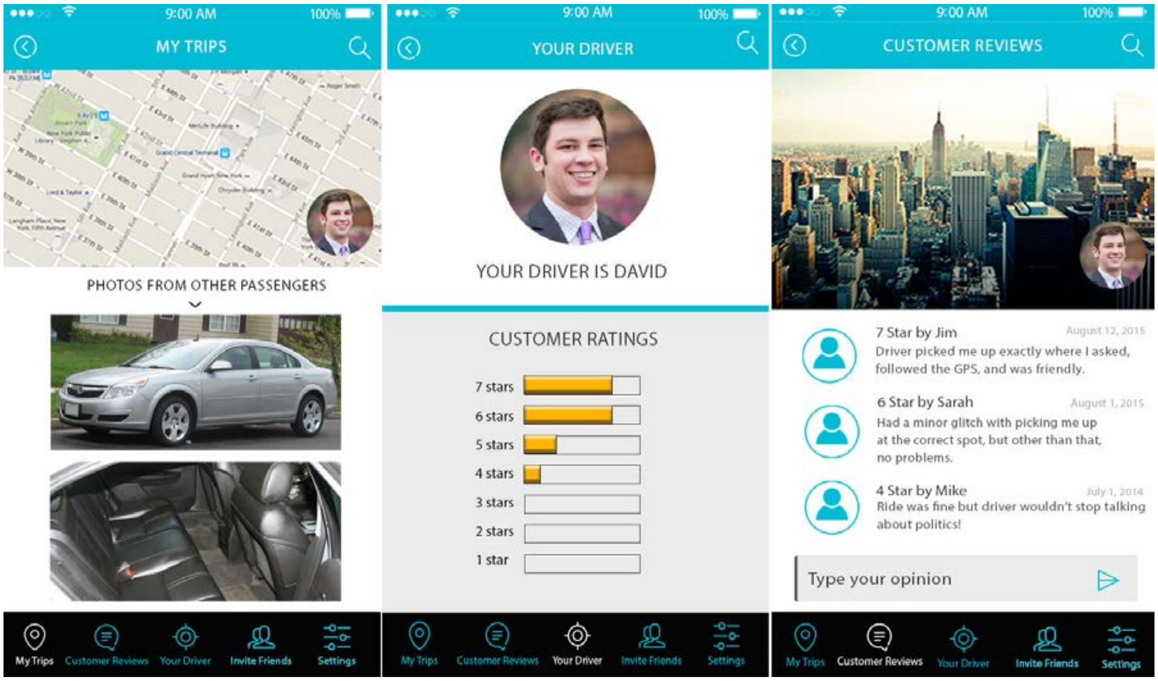
Our experiment employed a 2 (gender) x 2 (race) x 2 (Historical Quality) x 2 (Experience Quality), between-subjects design. Our first two dimensions (gender and race), were manipulated in the study by presenting the subject with driver photographs that varied across gender (Male, Female) and race (Caucasian, African American). We chose to also manipulate race in order to evaluate robustness of gender effects across racial lines. This is consistent with prior literature focused on gender bias (Swim et al. 1989; Unger 1976). We manipulated quality by altering the information subjects were given about the driver. Our experiment had two distinct phases[2] 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 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.

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. Specifically, we required their help in evaluating our drivers' performance. Human Intelligence Tasks (HIT) involving employee evaluations are commonly carried out on AMT (Berinsky et al. 2012)). Presenting ourselves as a real company and having participants engage in ratings

[2] The term "Phase" is used for ease of exposition of the experiment; subjects were not told that they were in Phase 1 or 2 while participating in the study.

that they believed had actual impact for drivers was intended to bring the experimental setting closer to a real world setting. We also created a new account on AMT to run this study so that individuals would not have had any historical interaction with the account on the platform.



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 1, 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: Example of High Historical Quality Driver]

Following this, participants provided general demographic data about themselves 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. This baseline from Phase 1 is particularly useful because it helps us address potential confounds from raters that exhibit bias against particular drivers before even having the ride experience described to them (i.e., Phase 2). The panels were designed to resemble a mobile web application (modeled on popular ride sharing apps) and were intended to mimic screenshots from the application. 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 that takes up less than ¼ of the total screen and the second panel has an image of the driver that is approximately twice the size of the image in the other two panels. 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. A seven point scale was used to allow for broader range of response from participants. Prior works indicates that seven point scales are more suited to electronic distribution of survey instruments (Finstad 2010); and that reliability does not differ significantly between five and seven point scales (Johns 2010; Preston and Colman 2000)). Photos of all drivers are available upon request.

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 could be 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. Prior to running our main experiment, two additional pre-studies were conducted (described below). These were intended to refine and validate the manipulations used in it.

Pre-Study 1: Race and Gender Manipulations

In the first pre-study, we focused on validating the manipulations of race and gender used in the main experiment. Specifically, we sought to confirm that there was agreement about the race and gender of the driver, as to avoid introducing unintended bias into the experiment by presenting drivers with ambiguous ascriptive characteristics (or seemed to be of international descent). 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 (Todorov et al. 2005; Zebrowitz 1996). Taking this approach allowed us to pre-empt experimental design concerns by ensuring consistency in the characteristics of the drivers used in the experiment.

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. 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); all of which is standard practice in absolute zero-acquaintance studies (Naumann et al. 2009). We then created a short survey that was used to validate that the photographs were appropriate for the manipulations we intended.

We recruited 48 participants from AMT and asked them to provide their input on the students based solely on the student's photograph. This type of evaluation of a person based on the presentation of only a photograph is known as a zero-acquaintance study of judgment (Naumann et al. 2009). Its reliability and consistency relative to in-person, face-to-face evaluations has been tested in a variety of contexts (Todorov et al. 2008) and has been shown to be an appropriate substitute (Vazire et al. 2008). We did not use full-body photographs to avoid additional information about the individual being gleaned from factors like full attire, body positioning, posture, and so forth (Ambady and Rosenthal 1992). Finally, the size and proportion of the headshot were identical for all individuals. Names for the individuals were chosen from a 2014 online repository of popular names from Johnson & Johnson, where the name appeared by gender and mother's ethnicity. To reduce the bias associated with names, we found the most popular names for both African Americans and Caucasians; "David" for males and "Kayla" for females. It should be noted that while these names are common across ethnicities, we have no data regarding their correlation with wealth.

Subjects were then asked to provide their opinion on the gender, race, and ethnicity (i.e., whether or not the subject was born in the United States) of the person in the photograph. They were also asked to rate each person on trustworthiness, attractiveness, kindness, and warmth. For these measures, individuals were rated on a Likert scale ranging from 1-Strongly Disagree to 5-Strongly Agree. 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%). Regression analysis confirmed that agreement on these dimensions are not significantly different for both females and males (see Tables 1a and 1b). 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. The only exception was that African American women were rated as slightly less attractive than their Caucasian counterparts. This effect did not show up across race for males. Through this pre-study, we were able to identify individuals that generated broad agreement on race and gender between raters while also exhibiting minimal differences in initial perception of these individuals.

[Table 1a: Differences in Pre-Study for Women]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent	Gender	Race Agree	U.S	Trustworthy	Attractive	Kind	Welcoming
Variable	Agree						
African American	-0.0208	-0.0313+	-0.0746	-0.0216	-0.309*	-0.0934	-0.125
	(0.0147)	(0.0179)	(0.0471)	(0.0875)	(0.122)	(0.0889)	(0.0888)
Constant	--	--	0.917**	3.948**	3.656**	4.125**	4.167**
	--	--	(0.0284)	(0.0652)	(0.0795)	(0.0666)	(0.0605)
Observations	192	192	191	191	191	191	191
R-squared	0.011	0.016	0.013	0.000	0.033	0.006	0.010

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

[Table 1b: Differences in Pre-Study for Men]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent	Gender	Race Agree	U.S	Trustworthy	Attractive	Kind	Welcoming
Variable	Agree						
African American	--	-0.0208	0.0208	-0.0104	0.177	-0.0104	-0.125
	--	(0.0147)	(0.0290)	(0.103)	(0.137)	(0.0856)	(0.0924)
Constant	--	--	0.948**	3.750**	3.437**	3.979**	3.979**
	--	--	(0.0228)	(0.0755)	(0.0991)	(0.0628)	(0.0645)
Observations	192	192	192	192	192	192	192
R-squared		0.011	0.003	0.000	0.009	0.000	0.010

Robust standard errors in parentheses; ** p<0.01, * p<0.05, + p<0.1
Note: There was no disagreement for Gender across Male Subjects

Pre-Study 2: Quality

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. 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 face of the driver in the pre-study was replaced with a gender-neutral silhouette.

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 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 (Table 2).

[Table 2: Pre-Study 2 Results Showing Intended Effect of Quality Manipulations on Ratings]

	(1)	(2)
Phase	Phase 1	Phase 2
High Quality	1.274**	
	(0.176)	
Quality		0.985**
		(0.0515)
Constant	4.370**	6.342**
	(0.116)	(0.152)
Observations	99	137
R-squared	0.353	0.641

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

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 (Imbens and Wooldridge 2007). We estimate this model an OLS with robust standard errors. Our estimated model is described below:

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

OverallRating$_i$ is a continuous measure from 1-7 that captures the overall star rating given to the driver by a rater *i*. *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 (relative to Caucasian male drivers) 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 (relative to Caucasian males). A significant coefficient of β_5 would provide evidence for *H2*, and suggests that women accrue a more severe penalty (relative to Caucasian males) 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.

Sample

919 participants completed the full experiment (sample descriptive statistics are provided in Table 3). Our sample had an average age of 34, was 73% Caucasian, 58% male, and fourteen percent had a college education. Also, our sample had knowledge of and experience with ride sharing. 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. In addition, 64% of our sample had engaged with a ride sharing service; the majority having had experience with Uber. 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 3: Sample Composition and Randomization Check]

		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

Results

We first analyze the impact of race, gender, and quality on the baseline assessments of our drivers in Phase 1 (Table 4, 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. Similar effects are observed for various sub-groups as the raters of interest, and indicate that focusing on social ingroups, such as Caucasian males (who typically dominate managerial positions and the ridesharing market (Cohen and Huffman 2007a; Hall and Krueger 2015)) does not create otherwise unobserved bias in the results. Results are available upon request. 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). Moreover, in this phase female drivers have a higher coefficient on *LowQuality* relative to male drivers 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 4, 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 their male counterparts when quality transgressions occur. Although this coefficient is identified through variation from Caucasian women (i.e. when AA=0), this acts as our baseline estimate of gender bias in our model. As a result, 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. Although we do not find significant differences in ratings in Phase 1, 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. Importantly, we find that women are punished more harshly for a salient low quality experience. Caucasian male drivers received approximately a 2.6 star drop from their initial rating following a negative experience. However, women saw a 14% larger drop over Caucasian males due to the observed bias (approximately 3 stars).

Further parsing of our data reveals that Caucasian males seem to be driving this gender bias in ratings. Again, we focus on Caucasian male rates due as our primary social ingroups because they typically dominate managerial positions and the ridesharing market (Cohen and Huffman 2007a; Hall and Krueger 2015). 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 4, 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). Excluding Caucasian male raters results in only quality significantly driving differences in rating (Table 4, Columns 7), indicting white male raters as the critical group driving biases in our setting. Further subsample analysis, i.e. focusing on women or minorities beyond Table 4 Columns 6 and 7, did not yield meaningful differences and is available from the authors upon request.

Next, we analyzed whether these effects would be ameliorated by high historical quality. Specifically, we evaluated how the bias exhibited by Caucasian male raters differed 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

[Table 4: 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

high historical was quality). Results indicate when female drivers had high historical quality, they were still punished for low quality experiences relative to Caucasian male drivers. In particular, we find that if a driver had high historical quality and then had a low quality experience, Caucasian male raters 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.

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 5). We start by parsing our data by drivers that provide high versus low quality experiences. In doing so, we 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 significant penalties for women (Columns 3-5). In contrast, when the low quality experiences are along dimensions that are not highly gendered, gender bias disappears (Column 6 and 7). Evaluating this effect using 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.

[Table 5: Effect of Gendered Tasks]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample	High Quality	Low Quality	Car Condition	Driving Style	Route	Pickup	Luggage	Gendered
Female	0.0580	-0.359*	-0.540**	-0.525*	-0.542**	-0.00609	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	-0.0928	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

Robustness & Extensions

We also consider a series of robustness checks and extensions of our analysis. Because Caucasian male raters appear to drive the observed effects, we focus our robustness and extensions on that sub-group within our data. First, we consider whether accounting for various features of our rater, including their education levels, age, and familiarity with ride sharing, impacts our results. Although, randomization helps ensure that these features are randomly distributed across conditions, any correlation between these features and our manipulations could influence the results. Results are in Table 6 and indicate that controlling for these factors does not affect our results (Column 1). We also parsed our data by those who have used ride sharing in the past vs. those who have not (Columns 2 and 3, respectively). This helps to out rule out concerns that our effects are driven by individuals without experience on ride sharing platforms, particularly if individuals without experience using ridesharing platforms act in ways that are not realistic or appropriate for the context. Results are strongest when limiting our sample to individuals who have used ride sharing platforms previously. Related to this concern, we also estimate our model excluding those who indicated being either "somewhat unfamiliar" or "very unfamiliar" with ride sharing platforms (Column 4). Again, we find that our results are consistent when excluding these individuals, suggesting that gender biases may be reinforced (rather than ameliorated) by experience on ridesharing platforms. We also estimate a simple model including only indicators for our main manipulated factors and the interaction between *Female* and *LowQuality* (Column 5). This model captures the average difference between the main manipulated groups while also allowing the ratings to vary for women when experience quality is low. In other words, this model captures the average bias against women across both African American and Caucasian women. Similar to prior estimations, we find a significant and negative interaction terms, reinforcing that an average bias against women.

[Table 6: Robustness and Extensions]

	(1)	(2)	(3)	(4)	(5)
		Ride Sharing	Non-Ride	Excluding	
	With Controls	User	Sharing User	Unfamiliar	Simple Model
Low Quality	-2.339**	-2.303**	-2.565**	-2.370**	-2.642**
	(0.208)	(0.279)	(0.283)	(0.220)	(0.157)
African American (AA)	0.165	0.144	0.0275	0.216	-0.0378
	(0.187)	(0.280)	(0.191)	(0.187)	(0.113)
Female	0.0302	0.0486	-0.206	-0.0171	-0.0330
	(0.204)	(0.284)	(0.248)	(0.213)	(0.137)
Low Quality*Female	-0.851**	-0.891*	-0.344	-0.734*	-0.464*
	(0.319)	(0.416)	(0.453)	(0.327)	(0.226)
Low Quality*AA	-0.658*	-0.535	-0.517	-0.632*	
	(0.304)	(0.443)	(0.390)	(0.309)	
AA*Female	-0.112	-0.193	0.376	-0.177	
	(0.259)	(0.372)	(0.298)	(0.260)	
Low Quality*AA*Female	0.807+	0.823	-0.0734	0.724	
	(0.455)	(0.615)	(0.633)	(0.455)	
College	-0.0189				
	(0.150)				
Age	-0.00545				
	(0.00701)				
Rideshare Familiar	-0.132+				
	(0.0788)				
Rideshare User	-0.116				
	(0.124)				
Rideshare Driver	-0.483*				
	(0.218)				
Constant	7.014**	6.389**	6.615**	6.468**	6.523**
	(0.318)	(0.201)	(0.139)	(0.158)	(0.119)
Observations	400	258	142	373	400
R-squared	0.641	0.591	0.710	0.636	0.624

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

Discussion and Conclusion

While the existence of both bias in the review generation process (Gao et al. 2015; Godes and Silva 2012), as well as bias stemming from the ascriptive characteristics of transacting parties on digital platforms (Edelman et al. 2016; Ge et al. 2016), has gained increased attention from both scholars and policy makers, limited work has been devoted to quantifying bias which may persist after a worker gains access to the market. This is particularly problematic, because ascriptive characteristics of workers are well known to introduce bias into perceptions of quality (Bielby and Baron 1986; Dezso et al. 2013; Reskin et al. 1999). Taking an experimental approach, we investigate this gap in a context where many of the proposed solutions to ameliorating bias based on ascriptive characteristics (e.g. whitewashing) (Younkin and Kuppuswamy 2017) are infeasible: gender biases in ridesharing markets. Results from a novel, two phased, experiment indicate three critical findings. First, despite established evidence from sociology that suggests women may be overtly penalized for even participating in ridesharing markets (due to a violation of gender roles (Eagly 1987; Heilman and Eagly 2008)), results indicate that, conditional upon information about historic quality being available (Younkin and Kuppuswamy 2017), there is little evidence *ex ante* of gender bias. However, conditional upon an inferior service being rendered, we find that women are penalized to a far greater degree than men, particularly so by male raters. Finally, we find that this penalty accrues notably for tasks that are highly "gendered," such as the cleanliness of the vehicle, while men are penalized more uniformly across tasks for imperfect service.

Notable contributions to research and practice stem from this observation. From a theoretical standpoint, as alluded to above, we contribute to a rich, but still emerging, literature discussing the biases in perceptions of platform based work. Although future work is needed to corroborate the generalizability of our findings, i.e. under what circumstances women and other social outgroups are more or less likely to be penalized, this initial step is nevertheless important. Furthermore, our work extends extant research in supervisor bias significantly as well. To the degree that many aspects of bias in the manager-subordinate relationship have been investigated, including: gender bias (Bielby and Baron 1986; Cohen and Huffman 2007b), race (Cohen and Huffman 2007a; Elliott and Smith 2004), ingroup biases (Allport 1979; Brewer 1979), political ideology and managerial beliefs (Carnahan and Greenwood 2017), and even beliefs about gender roles (Eagly 1987; Eagly and Karau 2002); it is notable that each of these investigations have occurred in contexts where a traditional manager is evaluating a subordinate. The context of ridesharing and the gig-economy challenges this relationship at a fundamental level, because the evaluation of the worker (i.e. the driver in the case of Uber or the homeowner in the case of AirBnB) is distributed over a wide number of evaluators, as opposed to a single person. Thus, it is incumbent upon the research community 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 societal, i.e. macro, level biases.

This research also has important implications for design science work in the form of algorithmic debiasing. Inasmuch as this work has already demonstrated proof of concept in many contexts, including medicine (O'Malley et al. 2005) and digital design (Schneider et al. 2015), our work highlights a new direction this work should be taken, i.e., towards the gig-economy. Further, this work underscores the importance of researchers moving their findings out of the academic space, and into real time production environments.

Finally, this work contributes to the emerging stream of literature discussing the welfare implications of platforms and the digital economy (Bapna et al. 2016; Chan and Ghose 2014; Chan et al. 2016; Greenwood and Agarwal 2016; Greenwood and Wattal 2017). While such literature has highlighted both positive and negative social outcomes that stem from digital platforms, we advance this work by considering how bias may be affecting the participants who work on these platforms, and what steps must be taken to limit it.

These findings also yield important practical implications. To the degree that understanding the nature of biases which may characterize perceptions of quality in online transactions is essential for the design of interventions, at both the firm and societal level, we believe these results speak directly to policy makers. At the firm level, our work underscores both that managers should aggressively pursue algorithmic options of debiasing for multiple reasons (above and beyond the ethical imperative). First, following the arguments of Becker (1971; 1976), the firm puts itself at a strategic disadvantage if it systematically undervalues talent from outgroups (e.g., women or other social minorities). Insofar as Uber and other 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 are 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 (recall that drivers are almost exclusively independent contractors), the firm may place itself in a tenuous legal position if it does not intervene to limit the effect of ascriptively motivated bias. Although these forms of class action lawsuits are typically difficult to demonstrate in court (Hart 2004), the concern is nevertheless pressing, notably if such lawsuits damage the firm's reputation in the open market.

From a public policy perspective, there are also notable implications. Chief among them is that the majority of workplace anti-discrimination law has not been written to deal with the unique organizational context offered by either electronic commerce or the gig-economy. Because non-employees are rating non-employees as a form of evaluation, it is possible that such behavior may go unchecked legally (despite the implications for the firm's reputation). It is therefore incumbent upon legislators to carefully update existing regulations in order to protect against protected classes being unduly punished.

There are limitations to our study. While the experimental approach we utilize affords us a number of advantages (e.g. allows us to exogenously manipulate quality information), it does have limitations. Most notable is that individuals may behave differently in more realistic settings relative to the setting employed in our experiment. While this is a legitimate concern, we employed a number of strategies to reduce these types of concerns. This includes using some deception to present ourselves as an actual company, having participants engage in a task that they believed impacted outcomes for actual drivers, and even creating a mock website for the company if any participants checked for an online presence. More so, there is significant literature (Ajzen 1985; 1991) to suggest that even if participants perceived the situation as more hypothetical than actual, their behaviors correlate well between the two settings. Further, the identified bias in hypothetical, relative to actual, setting suggests our results would be even stronger in a more realistic choice setting. For example, Ajzen et al. (2004) suggest that bias in hypothetical situations emerges because individuals imagine that they will behave in accordance with social norms or expectations than they actually do. Thus, behaviors that run counter to social norms (e.g. racial bias) would actually be under-estimated in hypothetical relative to actual settings. Finally, we cannot differentiate between situations where there was considerable agency on the part of the driver, and situations where there was not. This offers a rich opportunity to expand current work on attribution errors. As agents (e.g. the CEO (Park and Westphal 2013; Wade et al. 2006)) often possess significant agency, the theoretical implications of degree of agency have received limited attention, and are worthy of study.

In conclusion, despite the overwhelming evidence that online reviews are useful to consumers and can contribute to sales, there is a dark side to ratings' systems in the form of bias which recent work has begun to uncover. We build upon prior work that has identified ascriptive characteristics as a barrier to transacting in online markets, and extend it by considering the further implications for evaluations of quality. Findings indicate that Caucasian male raters disproportionately penalize outgroup providers conditional upon suboptimal experiences. Surprisingly, prior to having a salient experience with the driver, when simply presented with historical quality information, no such bias exists. However, when the same raters are presented with a more 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. 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.

References

Acquisti, A., and Fong, C. M. 2015. "An Experiment in Hiring Discrimination via Online Social Networks," *SSRN 2031979*), pp. 1-36.

Ajzen, I. 1985. "From Intentions to Actions: A Theory of Planned Behavior," in *Action Control: From Cognition to Behavior*, Springer Verlag: New York, pp. 11-39.

Ajzen, I. 1991. "The Theory of Planned Behavior," *Organizational Behavior and Human Decision Processes* (50:2), pp. 179-211.

Ajzen, I., Brown, T. C., and Carvajal, F. 2004. "Explaining the Discrepancy Between Intentions and Actions: The case of hypothetical bias in contingent valuation," *Personality and Social Psychology Bulletin* (30:9), pp. 1108-1121.

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

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

Aloisi, A. 2015. "Commoditized Workers The Rising of On-Demand Work, A Case Study Research on a Set of Online Platforms and Apps," *A Case Study Research on a Set of Online Platforms and Apps (July 2015)*).

Altonji, J. G., and Pierret, C. R. 2001. "Employer Learning and Statistical Discrimination," *Quarterly Journal of Economics* (116:1), pp. 313-350.

Ambady, N., and Rosenthal, R. 1992. "Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A meta-analysis," *Psychological Bulletin* (111:2), pp. 256-274.

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., and Brown, J. 1999. "When White Men Can't Do Math: Necessary and sufficient factors in stereotype threat," *Journal of Experimental Social Psychology* (35:1), pp. 29-46.

Arrow, K. J. 1972. "Models of Job Discrimination," in *Racial Discrimination in Economic Life,* A. H. Pascal (ed.), D.C. Heath: Lexington, Mass, pp. 83-102.

Arrow, K. J. 1998. "What Has Economics to Say About Racial Discrimination?," *Journal of Economic Perspectives* (12:2), pp. 91-100.

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

Bapna, R., Ramaprasad, J., Shmueli, G., and Umyarov, A. 2016. "One-way Mirrors in Online Dating: A randomized field experiment," *Management Science* (62:11), pp. 3100-3122.

Barasinska, N., and Schäfer, D. 2014. "Is Crowdfunding Different? Evidence on the Relation between Gender and Funding Success from a German Peer-to-Peer Lending Platform," *German Economic Review* (15:4), pp. 436-452.

Baron, J. N., and Bielby, W. T. 1980. "Bringing the Firms Back In: Stratification, segmentation, and the organization of work," *American Sociological Review* (45:5), pp. 737-765.

Baron, J. N., Mittman, B. S., and Newman, A. E. 1991. "Targets of Opportunity: Organizational and environmental determinants of gender integration within the California civil service, 1979-1985," *American Journal of Sociology* (96:6), pp. 1362-1401.

Becker, G. S. 1971. *The Economics of Discrimination* (2nd ed.) University of Chicago Press: Chicago, IL.

Becker, G. S. 1976. *The Economic Approach to Human Behavior* (1st ed.) University of Chicago Press: Chicago, IL.

Bell, L. A. 2005. "Women-led firms and the gender gap in top executive jobs,").

Berinsky, A. J., Huber, G. A., and Lenz, G. S. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon. com's Mechanical Turk," *Political Analysis* (20:3), pp. 351-368.

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

Bhattacherjee, A. 2001. "Understanding Information Systems Continuance: An expectation-confirmation model," *MIS Quarterly* (25:3) Sep, pp. 351-370.

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

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

Brewer, M. B. 1979. "In-group Bias in the Minimal Intergroup Situation: A cognitive-motivational analysis," *Psychological Bulletin* (86:2), pp. 307-324.

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

Brynjolfsson, E., Hu, Y., and Smith, M. D. 2003. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science* (49:11), pp. 1580-1596.

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

Burtch, G., Carnahan, S., and Greenwood, B. N. 2016. "Can You Gig it? An Empirical Examination of the Gig-Economy and Entrepreneurial Activity," (available at http://ssrn.com/abstract=2744352.).

Cardoso, A. R., and Winter-Ebmer, R. 2010. "Female-led firms and gender wage policies," *Industrial & Labor Relations Review* (64:1), pp. 143-163.

Carnahan, S., and Greenwood, B. N. 2017. "Managers' Political Beliefs and Gender Inequality Among Subordinates: Does His Ideology Matter More than Hers?," *Administrative Science Quarterly* (Forthcoming).

Casserly, M. 2012. "Playing House In The Office: The Cookie Conundrum," (available at https://www.forbes.com/sites/meghancasserly/2012/02/16/playing-house-in-the-office-the-cookie-conundrum/#f0c7c0d36ad2).

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

Chan, J., and Ghose, A. 2014. "Internet's Dirty Secret: Assessing the Impact of Online Intermediaries on HIV Transmission," *MIS Quarterly* (38:4), pp. 955-976.

Chan, J., Ghose, A., and Seamans, R. 2016. "The Internet and Racial Hate Crime: Offline Spillovers from Online Access," *MIS Quarterly* (40:2), pp. 381-403.

Charles, K. K., and Guryan, J. 2008. "Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination," *Journal of Political Economy* (116:5), pp. 773-809.

Chatterjee, P. 2001. "Online Reviews: Do consumers use them?," *Advances in Consumer Research* (28:1), pp. 129-133.

Chen, Y., and Xie, J. 2008. "Online Consumer Review: Word-of-mouth as a new element of marketing communication mix," *Management Science* (54:3), pp. 477-491.

Chevalier, J. A., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online book reviews," *Journal of Marketing Research* (43:3), pp. 345-354.

Clemons, E. K., Gao, G., and Hitt, L. M. 2006. "When Online Reviews Meet Hyperdifferentiation: A study of the craft beer industry," *Journal of Management Information Systems* (23:2), pp. 149-171.

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

Cohen, P. N., and Huffman, M. L. 2007a. "Black Under-Representation in Management Across US Labor Markets," *Annals of the American Academy of Political and Social Science* (609:1), pp. 181-199.

Cohen, P. N., and Huffman, M. L. 2007b. "Working for the Woman? Female managers and the gender wage gap," *American Sociological Review* (72:5), pp. 681-704.

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

Crandall, C. S., and Eshleman, A. 2003. "A Justification-Suppression Model of the Expression and Experience of Prejudice," *Psychological Bulletin* (129:3), pp. 414-446.

Davis, S. N., and Greenstein, T. N. 2009. "Gender ideology: Components, predictors, and consequences," *Annual Review of Sociology* (35), pp. 87-105.

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

Dellarocas, C. 2003. "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science* (49:10), pp. 1407-1424.

Dellarocas, C., Gao, G., and Narayan, R. 2010. "Are Consumers More Likely to Contribute Online Reviews for Hit or Niche Products?," *Journal of Management Information Systems* (27:2), pp. 127-158.

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

Dellarocas, C., Zhang, X. M., and Awad, N. F. 2007. "Exploring the Value of Online Product Reviews in Forecasting Sales: The case of motion pictures," *Journal of Interactive Marketing* (21:4), pp. 23-45.

Devine, P. G. 1989. "Stereotypes and Prejudice: Their automatic and controlled components," *Journal of Personality and Social Psychology* (56:1), pp. 5-18.

Dezso, C., Ross, D. G., and Uribe, J. 2013. "Why are there so few women top managers? A large-sample empirical study of the antecedents of female participation in top management," *Social Science Research Network* (11:1).

Doleac, J. L., and Stein, L. C. D. 2013. "The Visible Hand: Race and Online Market Outcomes," *Economic Journal* (123:572), pp. 469-492.

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

Duarte, J., Siegel, S., and Young, L. 2012. "Trust and Credit: The role of appearance in peer-to-peer lending," *Review of Financial Studies* (25:8), pp. 2455-2484.

Eagly, A. H. 1987. *Sex Differences in Social Behavior: A Social-role interpretation* (1st ed.) Lawrence Erlbaum: Hillsdale, New Jersey.

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

Eagly, A. H., and Diekman, A. B. 2012. "Prejudice in context departs from attitudes toward groups," *Behavioral and Brain Sciences* (35:6), pp. 431-432.

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

Edelman, B., and Luca, M. 2014. "Digital Discrimination: The Case of Airbnb. com," *Harvard Business School NOM Unit Working Paper*:14-054).

Edelman, B. G., Luca, M., and Svirsky, D. 2016. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment," *American Economic Journal: Applied Economics* (Forthcoming), pp. 1-36.

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

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

Ellingson, J. E., Gruys, M. L., and Sackett, P. R. 1998. "Factors Related to the Satisfaction and Performance of Temporary Employees," *Journal of Applied Psychology* (83:6), p. 913.

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

Ely, R. J. 1995. "The power in demography: Women's social constructions of gender identity at work," *Academy of Management journal* (38:3), pp. 589-634.

Finstad, K. 2010. "Response interpolation and scale sensitivity: Evidence against 5-point scales," *Journal of Usability Studies* (5:3), pp. 104-110.

Flake, D. F. 2016. "Employer Liability for Nonemployee Discrimination," *Available at SSRN 2780677*).

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

Frone, M. R., Russell, M., and Cooper, M. L. 1992. "Antecedents and Outcomes of Work-Family Conflict: Testing a Model of the Work-Family Interface," *Journal of Applied Psychology* (77:1), pp. 65-78.

Gao, G., Greenwood, B. N., McCullough, J. S., and Agarwal, R. 2015. "Vocal Minority and Silent Majority: How Do Online Ratings Reflect Population Perceptions of Quality?," *MIS Quarterly* (39:3), pp. 565-589.

Ge, Y., Knittel, C., MacKenzie, D., and Zoepf, S. 2016. "Racial and Gender Discrimination in Transportation Network Companies," NBER Working Paper No. 22776 (available at http://www.nber.org/papers/w22776)

Gerstner, E., Hess, J. D., and Holthausen, D. M. 1994. "Price Discrimination Through a Distribution Channel: Theory and evidence," *American Economic Review* (84:5), pp. 1437-1445.

Godes, D., and Mayzlin, D. 2004. "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science* (23:4), pp. 545-560.

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

Goes, P. B., Lin, M., and Au Yeung, C.-m. 2014. "“Popularity Effect” in User-Generated Content: Evidence from online product reviews," *Information Systems Research* (25:2), pp. 222-238.

Goldberg, P. 1968. "Are women prejudiced against women?," *Society* (5:5), pp. 28-30.

Goodman, J. K., Cryder, C. E., and Cheema, A. 2013. "Data Dollection in a Flat World: The strengths and weaknesses of Mechanical Turk samples," *Journal of Behavioral Decision Making* (26:3), pp. 213-224.

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

Greenwood, B. N., and Agarwal, R. 2016. "Matching Platforms and HIV Incidence: An Empirical Investigation of Race, Gender, and Socioeconomic Status," *Management Science* (62:8), pp. 2281-2303.

Greenwood, B. N., and Wattal, S. 2017. "Show Me The Way To Go Home: An Empirical Investigation of Ridesharing and Motor Vehicle Fatalities," *MIS Quarterly* (Forthcoming).

Guryan, J., and Charles, K. K. 2013. "Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots," *Economic Journal* (123:572), pp. F417-F432.

Hall, J. V., and Krueger, A. B. 2015. "An Analysis of the Labor Market for Uber’s Driver-Partners in the United States," Working Paper 587, Princeton University. Industrial Relations Section. (available at http://arks.princeton.edu/ark:/88435/dsp010z708z67d)

Harrison-Walker, L. J. 2001. "The Measurement of Word-of-Mouth Communication and an Investigation of Service Quality and Customer Commitment as Potential Antecedents," *Journal of Service Research* (4:1), pp. 60-75.

Hart, M. 2004. "Will employment discrimination class actions survive," *Akron L. Rev.* (37), p. 813.

Hartnett, O., and Bradley, J. 1986. "Sex Roles and Work," in *The Psychology of Sex Roles,* D. J. Hargreaves and A. M. Colley (eds.), Harper & Row: London, pp. 215-232.

Hartocollis, A. 2010. "For Women Who Drive, the Stereotypes Die Hard."

Hawksworth, J., and Vaughan, R. 2014. "The Sharing Economy - Sizing the revenue opportunity," PricewaterhouseCoopers. (available at http://www.pwc.co.uk/issues/megatrends/collisions/sharingeconomy/the-sharing-economy-sizing-the-revenue-opportunity.html).

Heilman, M. E. 2001. "Description and Prescription: How gender stereotypes prevent women's ascent up the organizational ladder," *Journal of Social Issues* (57:4), pp. 657-674.

Heilman, M. E., and Eagly, A. H. 2008. "Gender stereotypes are alive, well, and busy producing workplace discrimination," *Industrial and Organizational Psychology* (1:4), pp. 393-398.

Hekman, D. R., Aquino, K., Owens, B. P., Mitchell, T. R., Schilpzand, P., and Leavitt, K. 2010. "An Examination of Whether and How Racial and Gender Biases Influence Customer Satisfaction," *Academy of Management Journal* (53:2), pp. 238-264.

Hennig-Thurau, T., Wiertz, C., and Feldhaus, F. 2012. "Exploring the 'Twitter Effect:' An Investigation of the Impact of Microblogging Word of Mouth on Consumers' Early Adoption of New Products." (available at

Imbens, G., and Wooldridge, J. 2007. "Difference-in-Differences Estimation," Summer Institute, National Bureau of Economics Research. (available at http://www.nber.org/WNE/lect_10_diffindiffs.pdf), pp. 1-19.

Johns, R. 2010. "Likert items and scales," *Survey Question Bank: Methods Fact Sheet* (1), pp. 1-11.

Keith, P. M., and Schafer, R. B. 1980. "Role strain and depression in two-job families," *Family Relations*), pp. 483-488.

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

Kuruzovich, J., Viswanathan, S., Agarwal, R., Gosain, S., and Weitzman, S. 2008. "Marketspace or Marketplace? Online information search and channel outcomes in auto retailing," *Information Systems Research* (19:2), pp. 182-201.

Lee, M., and Huang, L. 2017. "Gender bias, social impact framing, and evaluation of entrepreneurial ventures," *Organization Science* (Forthcoming).

Lee, Y.-J., Hosanagar, K., and Tan, Y. 2015. "Do I Follow My Friends or the Crowd? Information cascades in online movie ratings," *Management Science* (61:9), pp. 2241-2258.

Li, G., Baker, S. P., Langlois, J. A., and Kelen, G. D. 1998. "Are female drivers safer? An application of the decomposition method," *Epidemiology*), pp. 379-384.

Li, X., and Hitt, L. M. 2008. "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research* (19:4), pp. 456-474.

Loscocco, K. A. 1997. "Work-Family Linkages among Self-Employed Women and Men," *Journal of Vocational Behavior* (50:2), pp. 204-226.

Lowery, B. S., Hardin, C. D., and Sinclair, S. 2001. "Social Influence Effects on Automatic Racial Prejudice," *Journal of Personality and Social Psychology* (81:5), pp. 842-855.

Mudambi, S. M., and Schuff, D. 2010. "What Makes a Helpful Review? A study of customer reviews on Amazon. com," *MIS Quarterly* (34:1), pp. 185-200.

Naumann, L. P., Vazire, S., Rentfrow, P. J., and Gosling, S. D. 2009. "Personality Judgments Based on Physical Appearance," *Personality and Social Psychology Bulletin* (35:12) December, pp. 1661-1671.

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

O'Malley, A. J., Zaslavsky, A. M., Elliott, M. N., Zaborski, L., and Cleary, P. D. 2005. "Case-Mix Adjustment of the CAHPS® Hospital Survey," *Health Services Research* (40:6), pp. 2162-2181.

Olian, J. D., Schwab, D. P., and Haberfeld, Y. 1988. "The Impact of Applicant Gender Compared to Qualifications on Hiring Recommendations: A meta-analysis of experimental studies," *Organizational Behavior and Human Decision Processes* (41:2), pp. 180-195.

Oliver, R. L. 1977. "Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation," *Journal of applied psychology* (62:4), p. 480.

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

Parker, G. G., and Van Alstyne, M. W. 2005. "Two-Sided Network Effects: A theory of information product design," *Management Science* (51:10), pp. 1494-1504.

Parker, G. G., Van Alstyne, M. W., and Choudary, S. P. 2016. *Platform Revolution: How networked markets are transforming the economy and how to make them work for you* (1st ed.) W.W. Norton & Company: New York.

Peer, E., Vosgerau, J., and Acquisti, A. 2014. "Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk," *Behavior Research Methods* (46:4), pp. 1023-1031.

Perry-Jenkins, M., and Crouter, A. C. 1990. "Men's provider-role attitudes: Implications for household work and marital satisfaction," *Journal of family Issues* (11:2), pp. 136-156.

Pettigrew, T. F. 1979. "The Ultimate Attribution Error: Extending Allport's cognitive analysis of prejudice," *Personality and Social Psychology Bulletin* (5:4), pp. 461-476.

Pope, D. G., and Sydnor, J. R. 2011. "What's in a Picture? Evidence of Discrimination from Prosper. com," *Journal of Human Resources* (46:1), pp. 53-92.

Preston, C. C., and Colman, A. M. 2000. "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences," *Acta psychologica* (104:1), pp. 1-15.

Ravina, E. 2012. "Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets." (available at https://ssrn.com/abstract=1107307 or http://dx.doi.org/10.2139/ssrn.1107307), pp. 1-79.

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

Rhue, L., and Clark, J. 2016. "Who Gets Started on Kickstarter? Racial Disparities in Crowdfunding Success," SSRN. (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2837042), pp. 1-38.

Richins, M. L. 1983. "Negative Word-of-Mouth by Dissatisfied Consumers: A pilot study," *Journal of Marketing* (47:1), pp. 68-78.

Ridgeway, C. L. 1997. "Interaction and the Conservation of Gender Inequality: Considering employment," *American Sociological Review* (62:2), pp. 218-235.

Ross, L. 1977. "The Intuitive Psychologist and His Shortcomings: Distortions in the attribution process," *Advances in Experimental Social Psychology* (10), pp. 173-220.

Roth, L. M. 2004. "The social psychology of tokenism: Status and homophily processes on Wall Street," *Sociological Perspectives* (47:2), pp. 189-214.

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

Steele, C. M., and Aronson, J. 1995. "Stereotype Threat and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology* (69:5), pp. 797-811.

Steelman, Z. R., Hammer, B. I., and Limayem, M. 2014. "Data Collection in the Digital Age: Innovative Alterantives to Student Samples," *MIS Quarterly* (38:2), pp. 355-378.

Sunderland, M. 2017. "People Think Women Are Worse Drivers Than Men—Statistics Say Otherwise."

Swim, J., Borgida, E., Maruyama, G., and Myers, D. G. 1989. "Joan McKay versus John McKay: Do gender stereotypes bias evaluations?," *Psychological Bulletin* (105:3), pp. 409-429.

Tajfel, H., and Turner, J. C. 1979. "An Integrative Theory of Intergroup Conflict," in *The Social Psychology of Intergroup Relations,* W. G. Austin and S. Worchel (eds.), Brooks/Cole: Monterey, CA, pp. 33-47.

Todorov, A., Baron, S. G., and Oosterhof, N. N. 2008. "Evaluating Face Trustworthiness: A model based approach," *Social Cognitive and Affective Neuroscience* (3:2) June 1, 2008, pp. 119-127.

Todorov, A., Mandisodza, A. N., Goren, A., and Hall, C. C. 2005. "Inferences of Competence from Faces Predict Election Outcomes," *Science* (308:5728), pp. 1623-1626.

Trusov, M., Bucklin, R. E., and Pauwels, K. 2009. "Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site," *Journal of Marketing* (73:5), pp. 90-102.

Uhlmann, E. L., and Cohen, G. L. 2005. "Constructed Criteria: Redefining merit to justify discrimination," *Psychological Science* (16:6) Jun, pp. 474-480.

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

Vazire, S., Naumann, L. P., Rentfrow, P. J., and Gosling, S. D. 2008. "Portrait of a Narcissist: Manifestations of narcissism in physical appearance," *Journal of Research in Personality* (42:6), pp. 1439-1447.

Wade, J. B., Porac, J. F., Pollock, T. G., and Graffin, S. D. 2006. "The Burden of Celebrity: The impact of CEO certification contests on CEO pay and performance," *Academy of Management Journal* (49:4), pp. 643-660.

Wallston, B. S., and O'Leary, V. E. 1981. "Sex Makes a Difference: Differential perceptions of women and men," in *Review of Personality and Social Psychology,* L. Wheeler (ed.), Sage Publications: Newbury Park, CA, pp. 9-41.

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

Younkin, P., and Kuppuswamy, V. 2017. "The Colorblind Crowd? Founder Race and Performance in Crowdfunding," *Management Science*).

Zebrowitz, L. A. 1996. "Physical Appearance as a Basis of Stereotyping," in *Stereotypes and Stereotyping,* C. N. Macrae, C. Stangor and M. Hewstone (eds.), The Guilford Press: New York, pp. 79-120.

Zhu, F., and Zhang, X. 2010. "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," *Journal of Marketing* (74:2), pp. 133-148.