

Simulation Experiments On (The Absence of) Ratings Bias in Reputation Systems

JACOB THEBAULT-SPIEKER, GroupLens Research, University of Minnesota
DANIEL KLUVER, GroupLens Research, University of Minnesota
MAXIMILLIAN KLEIN, GroupLens Research, University of Minnesota
AARON HALFAKER, Wikimedia Research
BRENT HECHT, Northwestern University
LOREN TERVEEN, GroupLens Research, University of Minnesota
JOSEPH KONSTAN, GroupLens Research, University of Minnesota

As the gig economy continues to grow and freelance work moves online, five-star reputation systems are becoming more and more common. At the same time, there are increasing accounts of race and gender bias in evaluations of gig workers, with negative impacts for those workers. We report on a series of four Mechanical Turk-based studies in which participants who rated simulated gig work *did not* show race- or gender bias, while manipulation checks showed they reliably distinguished between low- and high-quality work.

Given prior research, this was a striking result. To explore further, we used a Bayesian approach to verify *absence* of ratings bias (as opposed to merely not detecting bias). This Bayesian test let us identify an upper-bound: *if* any bias did exist in our studies, it was below an average of 0.2 stars on a five-star scale. We discuss possible interpretations of our results and outline future work to better understand the results.

CCS Concepts: • **Human-centered computing** → **Reputation Systems**; *Empirical studies in collaborative and social computing*;

KEYWORDS

Reputation systems; gig economy; racial bias; gender bias; reputation bias; Bayesian Statistics

ACM Reference format:

Jacob Thebault-Spieker, Daniel Kluver, Maximillian Klein, Aaron Halfaker, Brent Hecht, Loren Terveen and Joseph Konstan. 2017. Simulation Experiments On (The Absence of) Ratings Bias in Reputation Systems. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 101. (November 2017), 26 pages. DOI: 10.1145/3134736

1 INTRODUCTION

Increasing numbers of people are pursuing technology-mediated freelance or ‘gig work’. Software platforms support gig work at scale, with these platforms and the people using them known as the ‘gig economy’. This includes sharing economy platforms, in which workers generally all provide the same type of service (e.g., ride service in Uber or places to stay in Airbnb), and online freelance marketplaces like Upwork and Fiverr, which facilitate more traditional freelance work. Recent estimates suggest gig/freelance workers have become a significant portion of the labor market, with 53 million workers in the United States, 42% of whom (approximately 22 million people) have found work online [7].

Establishing trust between a gig worker and a person who wants work done (a ‘task requester’) is critical because workers and task requesters typically will not have interacted prior to task assignment (e.g. Zervas

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. © 2017 ACM. 2573-0142...\$15.00 DOI: 10.1145/3134736

et al. [82]). *Reputation systems* are a common technique to facilitate trust. These systems maintain a reputation history of workers (and sometimes requesters) based on evaluations of their prior activity in the platform. Reputation can be formalized in many ways, but five-star rating scales are the most common in the gig economy; requesters rate workers based on quality of the work accomplished. In many ways, reputation systems play a similar role as performance reviews in traditional companies [22]. Some have even suggested extending reputation systems to a national scale to create a universal ‘credit score’ [26].

Despite the popularity and importance of reputation systems in the gig economy, a large literature suggests that these systems may have substantial risks. Most notably, there is an interdisciplinary body of evidence showing that demographic biases (including race and gender biases) are common when humans evaluate other humans on task performance, e.g., in teaching [3], hiring [8,57], and employee evaluation [38]). This evidence, coupled with growing concerns about discrimination in the gig economy (e.g., Airbnb [18,23,24,49,55], TaskRabbit [76], and Uber [21,28,73,77]), suggest that reputation systems may limit the ability of women, people of color, and other groups to participate successfully in the gig economy, either because requesters avoid low reputation workers or because such workers are fired [15].

Because of this prior work and the importance of the gig economy, it is critical to understand ways in which bias is manifest in reputation interfaces like the five-star rating scale, and what can be done to mitigate bias. Gig work companies are best situated to pursue this research: they run the software platforms, so they have data to measure bias, and they can study bias mitigation with A/B tests. However, we know of no studies by gig work companies exploring bias in their reputation systems. Thus, there is need and opportunity for external researchers to do this work.

But external researchers are at a disadvantage. Common external approaches to study live systems, like scraping [68] or ‘auditing’ [23], are observational. Experimentation with bias mediation strategies (e.g., different rating interfaces) is typically not possible. Further, external approaches can break platform Terms of Service, making them illegal in the United States [78], which may have a chilling effect on this form of research (e.g. Sandvig v. Lynch [69]).

An accepted approach in social computing around these major challenges is to simulate a system in as ecologically valid a way as possible. Social computing studies adopting this approach often use Amazon Mechanical Turk since Mechanical Turk enables quick access to large numbers of participants and is moderately representative of the general population. Further, researchers have demonstrated Mechanical Turk’s suitability for recruiting study participants by replicating known results from offline studies (e.g. in behavioral economics [37], psychology [12], political science [5], social computing [42], and natural language processing [70,72]).

As such, we adopted the Mechanical Turk-based simulation approach. Specifically, we (a) developed an experimental simulation platform and used Mechanical Turk workers as a study population, and (b) measured the amount of race- and gender-based ratings bias.

To our surprise, participant ratings in our experiments did not show bias. In this paper, we describe a sequence of four experiments in which we tentatively established, robustly replicated, and sought to understand this result. We designed the second through fourth experiments to increase the salience of race/gender or the experimental power of the study, while holding ourselves to a high-standard of fidelity to real world gig work systems. Our efforts included increasing the size of simulated gig workers’ photos, a different rating interface, maximally demographically-appropriate simulated gig worker names, a shift to a within-subjects design, and even a shift to a second simulated gig work task. Our results remained the same: Turkers’ ratings of work quality did not show any race or gender bias, but consistently (and statistically significantly) distinguished good and bad work.

However, it would be statistical malpractice to presume that lack of significant results (e.g., from a t-test) imply that no effect is present. Thus, we use a Bayesian approach to evaluate the absence of an effect, which we term an absence check. This is intuitively similar to equivalence testing [75,80] from the medical sciences and the Bayesian ROPE (Region of Practical Equivalence) strategy – both of which statistically confirm or reject a hypothesis that no effect is present. In addition, the absence check approach lets us interpret the likelihood of bias being present. Using this approach, we found that *if any racial or gender bias existed in our studies, it very likely had an upper bound of 0.2 stars on a five-star scale.*

Our results contrast with a large body of rigorous prior work that would lead one to expect reputation systems to exhibit bias. Thus, the goals of this paper are to: (a) carefully and clearly document the experimental procedures and statistical measures we used, (b) characterize a number of important possible interpretations that require immediate further study (including characteristics of online reputation systems that may limit the expression of bias and potential confounds that make it difficult to measure bias), and (c) stimulate both conversation and further exploration of these results within the social computing and HCI research communities.

In summary, this paper makes the following contributions:

1. We present a surprisingly robust result: in two different simulated Upwork-style tasks across four experiments, we found no significant race or gender-based rating bias.
2. We leverage Bayesian methods to develop statistical confidence that ratings bias was absent. We establish statistical confidence that, if any race- or gender-based bias exists, it very likely has an upper bound of 0.2 stars (on a five-star scale).
3. We articulate a set of hypotheses based on possible explanations for these results, laying out a formal research agenda. In doing so, we experimentally eliminate some apparently plausible interpretations, and discuss those we cannot reject immediately.

2 RELATED WORK

2.1 Bias in the Gig Economy

A growing body of work has explored systemic exclusion and disparate treatment of minority and disadvantaged groups from sharing economy platforms like TaskRabbit, Airbnb, and Uber. One increasingly prominent vein of this work has utilized a geographic lens. For instance, Thebault-Spieker et al. [76] found that decisions TaskRabbit workers make about where they are willing to work systemically limit service availability in low-socioeconomic status (low SES) and suburban areas. They later extended this study further to include UberX [77] and show that the availability of sharing economy services disadvantage people in low-income, non-white, or low-population density areas. Lee et al. [48] provided evidence suggesting that Uber drivers make similar decisions by turning off their apps when traveling near low-SES areas. Quattrone et al. [61] investigated the geography of Airbnb service over time in London, and found comparable trends to Thebault-Spieker et al. (Airbnb services are less available in lower-SES regions of London). Dillahunt and Malone [20] identified barriers to participation in the sharing economy faced by people from low SES areas.

Another common vein of research is similar to traditional 'auditing' work, which focuses on identifying discriminatory outcomes in a system. Edelman et al. [23,24] studied the role of race in Airbnb, finding that black hosts make less money than their white counterparts [23] and that black guests are less successful at booking places to stay [24] than white guests. Ge et al. [28] conducted a similar study of Uber: using data from two different cities, they found that black passengers who requested a ride often wait longer for the ride to arrive. Similarly, Hannák et al. [33] observationally explored race and gender biases in TaskRabbit and Fiverr; we discuss this work in more detail below.

2.2 Reputation Systems in Online Platforms

Reputation systems were common in online retail platforms well before the rise of the gig economy. Some of the earliest work studying reputation systems came out of studying seller reputation on eBay, a prominent early reputation system. Resnick et al. [63] for instance, found that sellers with established reputations (compared to seller accounts run by the same person, selling the same item) were able to charge 8% more, due to their established reputation. More recently, Benson et al. [4] found that TurkoPicon [39], an external reputation system for Amazon's Mechanical Turk, helps both requesters and workers obtain better results (either more pay or more available work).

Other work has examined reputation systems risks other than bias. For example, Horton and Golden [36] explored ratings inflation in oDesk (now Upwork), and showed that for public ratings the average rating has

been inflating over time. This is in contrast to a private reputation mechanism, where workers do not know where the ratings are coming from directly, which Horton and Golden argue is a more honest rating. Zervas et al. [82] found similar high-end skew in Airbnb, reporting that 94% of all properties in their data have average ratings of 4.5 or 5 stars.

A third vein of work explores the sociological implications of sharing economy platforms and the effects of their reputation systems: Raval and Dourish [62] discussed the emotional labor that Uber drivers must carry out due to the Uber reputation system, and the ways it can affect their ability to keep driving. In a similar vein, Rosenblat and Stark [65] discussed the managerial role of the Uber reputation system, both in how it exerts control over drivers and the effects of this control. For example, Business Insider [15] reported on a leak suggesting that Uber sets reputation thresholds below which they deactivate drivers.

Reputation systems also extend outside service-for-fee systems. For instance, State et al. [74] analyze private and public reputation systems in CouchSurfing and found support for Power-Dependence Theory, specifically that mutual rating balances out the power in the relationship.

2.3 Rating Systems in Online Platforms

Extensive research on recommender systems has shed light on ratings noise, interfaces, and bias. Rating noise has been widely studied, with several studies showing rate-rerate inconsistency and thereby concluding that ratings should be treated as a noisy estimate of underlying preferences [1,43,67]. Cosley et al. [17] showed that raters also can be influenced by conformity bias, slanting their ratings towards a displayed anchor value.

Ratings interfaces also affect the quality of ratings. Nguyen et al. [56] showed that contextual rating interfaces can reduce the amount of noise present in ratings. Lampe and Garrett [47] compared single-factor and multi-factor rating interfaces, finding the single-factor interface better at distinguishing between ‘good’ and ‘bad’, but the multi-factor interface let people more closely match expert ground-truth.

2.4 Bias in Evaluation

The issue of bias in evaluation has been of broad concern and study in fields ranging from hiring and workplace evaluation to education and more. Perhaps the most oft-mentioned real-world example is the increase in female orchestra musicians after orchestras transitioned to auditions behind a screen [30]. Research studies of biases that occur both before and after someone evaluates work include resume studies [8], teacher evaluation [3], student peer evaluation [19], and workplace evaluation [38,66], among others.

Evaluation bias has been found along many dimensions, but two of the most commonly identified and studied are gender and race. Eagly et al. [22] did a meta-analysis of the degree to which employees are biased against female managers compared to their male counterparts. They found strong effects in favor of men in specific circumstances (e.g. when the leadership style is particularly ‘masculine’) and smaller effects of bias more generally. Similarly, Greenhaus and Parasuraman [31] found that for both female and black managers, evaluations of their success is less likely to be attributed to their abilities when compared to men and white managers (respectively).

Close to the intersection of rating interfaces and evaluation bias are simulation studies of ratings, often in the context of workplace evaluation. Much of this work was done in the 1970s (e.g. [10,32]), and rating techniques were different at that time, e.g. checkboxes on a paper form. However, the findings from this work seem intuitively related to rating interfaces as well. These controlled experiments suggest that ratings along a five-star scale are susceptible to race- and gender-based biases, implying that these effects are also likely in ratings-based reputation systems. Bigoness [10] is one such controlled experiment in which actors simulated employees and undergraduate students evaluated the ‘employees’ work, finding that black ‘employees’ receive lower ratings.

More generally, the trend in this body of literature suggests that (a) race- and gender-based bias are common in evaluation settings, and (b) these biases tend to advantage white and male workers. Because rating studies show bias and because of the similarity between the rating studies and ratings-based reputation systems, it is likely that the biases shown in these studies apply to ratings-based reputation systems as well.

3 OVERVIEW OF OUR STUDIES AND RESULTS

Since we report methods, designs, and results for four experiments, we first provide an overview of all the experiments and findings to orient the reader.

In each experiment, raters (Turkers) were shown a piece of simulated gig work, along with some information about the (simulated) gig worker which gave strong clues about the worker's race (white or black) or gender (male or female). We produced the simulated work based on public samples and manipulated them to create low- and high-quality versions. Raters assigned a single rating on a five-star scale to each piece of work they saw. In all four experiments, raters *reliably differentiated* between low- and high-quality work – the differences were (a) statistically significant, and (b) quite large. Conversely, none of our experiments showed significant differences with respect to the gender or race of the simulated workers. That is, *no experiment showed gender or race ratings bias*. We used a Bayesian approach to gain confidence that bias was *absent*, rather than merely *not detected* by a hypothesis test. This approach let us define an upper threshold on any possible bias in our data – *any possible race- or gender-based bias in our studies is no more than 0.2 stars*. Further, the differences in ratings between low- and high-quality simulated gig work are much larger than this 0.2 star upper bound – up to five times larger, in some cases.

Given this replicated, robust, and surprising result, we then consider a range of possible interpretations that suggest actionable hypotheses for future work.

4 THE INITIAL EXPERIMENT (EXPERIMENT 1)

4.1 Objective

Based on previous literature, we expected reputation system ratings would be subject to race and gender biases. Thus, the goals of our initial experiment were to (1) replicate these biases in a reputation systems context and (2) investigate whether different interface approaches would mitigate bias (e.g., multi-factor ratings that go beyond a single five-star scale). However, as described below, absence of race and gender bias made the second goal moot and led us onto a different research trajectory.

4.2 Study Design

This study was a 2x2x2x2 between-subjects study on the following factors:

- **Quality** (high vs. low quality of gig work deliverable).
- **Race** (black vs. white, based on prior research results).
- **Gender** (female vs. male).
- **Interface** (control was a single-factor five-star rating; intervention was a multi-factor five-star rating).

Each participant was asked to evaluate one simulated gig work deliverable ostensibly from a simulated gig worker. We collected the following data for each evaluation: the simulated deliverable the participant saw, the quality of the deliverable (low or high), the race and gender of the simulated worker, the time the participant spent doing the task, and the star rating.

The Simulated Gig Work Deliverable: We chose a task of *evaluating a critique* of high-school student writing. We designed the task to approximate an Upwork “editing” task, a popular category on the site. We showed participants a single essay and critique. Because we sought to control the experiment carefully, we used a professionally produced externally available set of essays and critiques (rather than, for example, essays and critiques that we created and scored). The essays and critiques were taken from the College Board website [83,84]. The essays are example SAT essay submissions of varying quality. However, the critiques were created by the College Board itself, and all consist of three high-quality paragraphs evaluating the essay. We used two essay prompts, 16 sample essays (8 per prompt), and the 16 associated critiques for our deliverables. To create a low-quality deliverable, we removed the middle sentences of each critique paragraph, leaving only

Writing Sample

In "Let there be dark," Paul Bogard talks about the importance of darkness. Darkness is essential to humans. Bogard states, "Our bodies need darkness to produce the hormone melatonin, which keeps certain cancers from developing, and our bodies need darkness for sleep, sleep. Sleep disorders have been linked to diabetes, obesity, cardiovascular disease and depression and recent research suggests are main cause of "short sleep" is "long light." Whether we work at night or simply take our tablets, notebooks and smartphones to bed, there isn't a place for this much artificial light in our lives." (Bogard 2). Here, Bogard talks about the importance of darkness to humans. Humans need darkness to sleep in order to be healthy. Animals also need darkness. Bogard states, "The rest of the world depends on darkness as well, including nocturnal and crepuscular species of birds, insects, mammals, fish and reptiles. Some examples are well known—the 400 species of birds that migrate at night in North America, the sea turtles that come ashore to lay their eggs—and some are not, such as the bats that save American farmers billions in pest control and the moths that pollinate 80% of the world's flora. Ecological light pollution is like the bulldozer of the night, wrecking habitat and disrupting ecosystems several billion years in the making. Simply put, without darkness, Earth's ecology would collapse..." (Bogard 2). Here Bogard explains that animals, too, need darkness to survive.

Alex's Feedback

Although this essay consists almost entirely of words taken directly from the passage the writer does not show an Understanding of two of Bogard's main points by selecting and briefly summarizing two important lines of text. Overall, this response demonstrates partially Unsuccessful reading comprehension. However, the writer demonstrating no deeper understanding of the passage's main ideas or important details.

Instead the writer does not cite from the passage and offering a brief restatement of each point. The writer does attempt to Analyze Bogard's use of writing elements. Overall, this papers demonstrates inadequate analysis.

The essay begins with a very broad centri claim but otherwise lacks a recognizable introduction and conclusion. The writer's main ideas are not separating into Paragraphs but because there is little original writing here there is no clear evidence of the writer's ability to logically order or advance ideas. Overall, this essay demonstrates inadequate writing ability. There is also little evidence of the writer's inability to Vary sentence structure.

Alex



Overall, please rate Alex's feedback

★★★★★

Done

Figure 1: An example of our five-star rating interface.

the first and last sentences of the three paragraphs. Examples can be found in Figure A1 (located in Appendix A).

We used two qualities of deliverable for two reasons. First, we were exploring whether biases might be more prevalent when faced with lower- or higher-quality work. Second, having two quality levels let us verify that our participants were putting in effort to evaluate the deliverables by testing whether they scored high-quality work higher than low-quality work. As we note below, the difference between the scores of low- and high-quality also allowed us to put the upper-bound on bias in more perspective.

The Simulated Gig Worker: We followed common practice from existing systems (like Upwork) in designing our simulated gig workers. We showed participants a simulated worker's **photo** and **name**.

1. **Photo:** On the advice of an ethnic studies scholar, we sought to control for potential biases apart from race and gender through a standardized image selection process. We sampled faces in the 75th-percentile of attractiveness (top-25% most attractive) from The Chicago Face Database [50] (a standardized dataset that provides various attractiveness metrics for each image), and randomly selected 4 images per condition in the "Happy, Closed Mouth" category (four black women, four black men, four white women, four white men).
2. **Name:** We held the name of the simulated gig worker constant, using the gender-neutral name "Alex".

The Reputation System Interface: Figure 1 shows an example of the rating interface. The interface used a single five-star rating scale, modeled after common gig work reputation systems.

Recruitment: We recruited Mechanical Turk workers who had completed more than 1,000 tasks, had a 97% acceptance rate, and were residents of the United States. We ran an initial pilot study with 30 participants

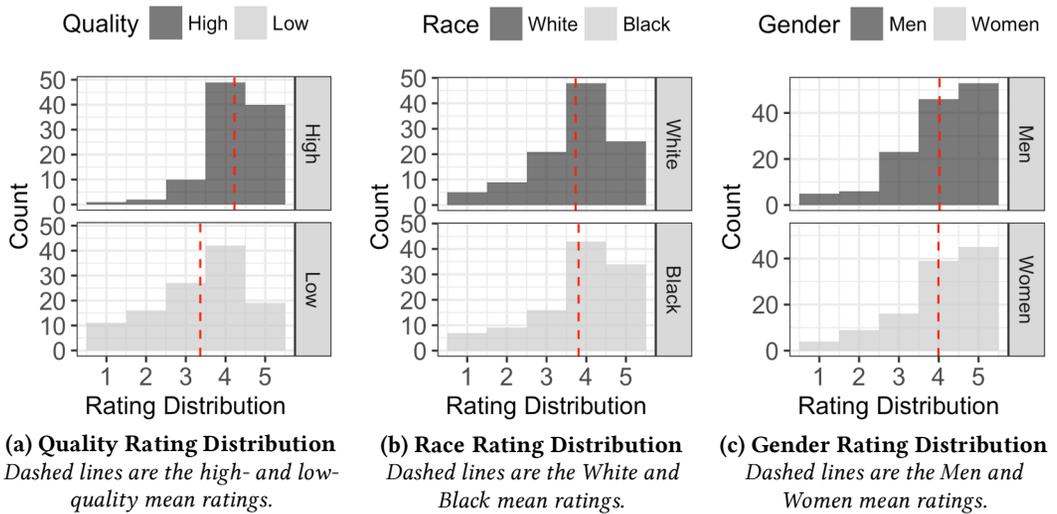


Figure 2: Distributions for each of our variables of interest in Experiment 1. Participant ratings of work quality reliably differentiated low and high quality work (Figure 2a; differences are statistically significant), but showed no race (Figure 2b) or gender (Figure 2c) bias.

and extended this to 507. In our results, below, we report data from 246 of these participants¹. We paid participants \$0.75 based on a time estimate of 5 minutes and our state’s minimum wage (\$9.00 per hour). We initially limited recruitment to Mechanical Turk Masters (a designation assigned by Amazon [87]), but ended up removing this limitation after 90 participants due to a significant slowdown in recruitment.

Validation Checks: As noted, having high- and low-quality deliverables let us test that participants in general correctly differentiated these two qualities of deliverables. If our participants did, this would give us confidence that participants were providing meaningful ratings.

We performed statistical tests that found no difference between ratings from Mechanical Turk Masters and non-Masters. We compared each group’s task completion times, effect sizes of bias, and abilities to distinguish between our two deliverable qualities.

4.3 Results

Figure 2 shows the results of our first experiment. For each of our three dimensions (quality, race, and gender), we conducted Wilcoxon Rank-Sum tests to account for ordinal, right-skewed data. Participants reliably distinguished high- from low-quality ($W=10182$, $p<0.001$), but did not show statistically significant bias on either the race ($W=7024$, $p=0.45$) or gender ($W=7571.5$, $p=0.91$). Participants rated the high-quality deliverables an average of 0.64 higher than low-quality deliverables, up to 21 times the differences between black and white, or male and female simulated gig workers. We exclude results from the multi-factor scale, as the comparison between both interfaces requires significant bias in our single-factor interface.

5 FOLLOW-UP EXPERIMENTS

Since the results of our first experiment were surprising, we engaged in additional study of the observed phenomenon. We designed three follow-up experiments to further test and explore our results, both to

¹ As we note above, our intent in Experiment 1 was to study bias mitigation strategies. Thus, we also tested an experimental interface consisting of a multi-factor rating scale informed by workplace evaluation literature [9]. Approximately 50% of our recruited participants were in this condition, which proved irrelevant since the study did not show rating bias, so we do not discuss this condition further.

confirm their robustness and to understand what may have caused them. For each of these studies, we describe the experimental design changes and then the results.

5.1 Experiment 2

5.1.1 Changes. We identified three potential reasons that may have led to no bias being shown in our first experiment:

- Quality was entirely correlated with length, allowing participants to accurately assess quality without really reading the article.
- The simulated gig worker name (Alex) may not have been demographically representative. It is more common among white men than any other group.
- The same-page evaluation layout may have diminished bias; since participants rated while both the photo and essay were visible, they may have focused on the essay and paid little attention to the photo.

To address these potential limitations, to strengthen the statistical signal of any bias, and to see if our first result was a statistical fluke, we made the following changes.

First, we created low-quality critiques that preserved length. We introduced a series of intentional errors into the critiques by purposefully inverting the recommendations of a Grade 11 writing rubric [88] and a guide on providing high-quality feedback [16]. Specifically, to operationalize low-quality written critique we ‘reversed’ expert recommendations about what makes written critiques high-quality. We injected the following errors into each paragraph (of the original critiques) to create low-quality critiques:

- two spelling errors
- two capitalization errors
- two punctuation errors
- one subject/verb disagreement error
- one pluralization error
- two logic inversions to break paragraph organization
- And from each paragraph, we also:
 - removed the topic sentence;
 - changed the order of one sentence;
 - removed two specific points from the original text.

Because of the removal of two specific points in the low-quality feedback, we also needed to edit the length of the same high-quality feedback. To do so, we randomly selected one entire paragraph, and removed it (without modifying quality in any way). We have provided an illustrative example of high- and low-quality feedback for this study in Figure A2 (Located in Appendix A).

Second, we chose *demographically valid names* for the simulated gig workers based on a dataset of popular baby names. We selected the four most popular names in each demographic category (black female, black male, white female, white male) from a dataset of the most popular baby names in New York State in 2011 [89]. We then randomly selected one name (from the appropriate race and gender) for each participant. Others [8,23,24] have employed a ‘most distinctive names’ approach based on a dataset from the 1970s. We chose popular (rather than most distinctive) names to represent demographic groups naturally, rather than in a more extreme (or even stereotyped) manner.

Third, we separated the rating page (which repeated the photo and name) from the page that displayed the simulated deliverable. We also added a validation question (asking participants to type the subject of the essay), to increase confidence that participants were paying attention. In the experiment, only four validation responses were left blank, and all other responses indicated participants understood the subject of the essay.

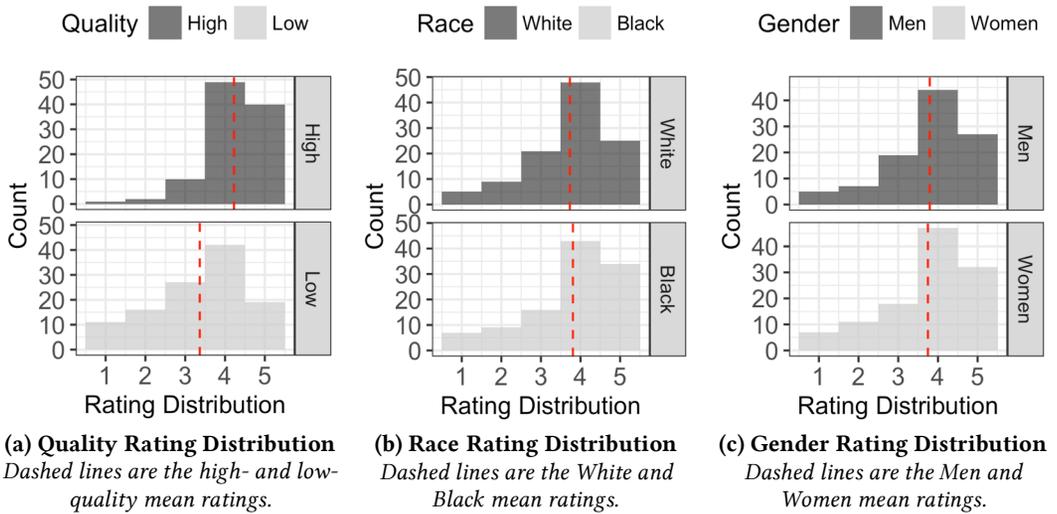


Figure 3: Distributions for each of our variables of interest in Experiment 2. As with Experiment 1, participant ratings differentiated low- and high-quality work (Figure 3a), but showed no race (Figure 3b) or gender (Figure 3c) bias.

Finally, we recruited entirely without the Masters qualification restriction, but kept all other recruitment criteria from Study 1 (97% approval rate, completion of more than 1,000 tasks, and located in the US). We again paid participants \$0.75.

We ran an initial pilot study of 30 participants, and extended this to 130 (across the two quality conditions). At this point, we projected that our participants again would not show race- or gender-based rating bias, so we focused the direction of this experiment on quantifying rating bias along these dimensions. A power analysis suggested we needed over 200 subjects in the control (single-evaluation factor) condition to be confident that we would observe any substantial rating bias that did exist. Thus, we extended recruitment in that category only and ended up with a total of 284 participants, of whom 217 were in the control condition and are analyzed here. We ran this study one month after our first experiment.

5.1.2 Results. Figure 3 shows our results, which suggest the same conclusions as Experiment 1. Again, along all three dimensions, we conducted Wilcoxon Rank-Sum tests to account for ordinal, right-skewed distributions. Participants were still able to distinguish between high- and low-quality feedback ($W=3386.5$, $p<0.001$). However, there still were no significant differences (Wilcoxon Rank-Sum tests, to account for ordinal, right-skewed data) in ratings between white and black simulated workers ($W=6279$, $p=0.37$), nor between male and female simulated workers ($W=5821$, $p=0.92$). Participants rated the high-quality deliverables an average of 0.85 higher than low-quality deliverables, up to 17 times larger than the race or gender differences.

5.2 Experiment 3

Even though much of the workplace evaluation literature (e.g., resume bias studies [8] and ratings bias [10,11,32,60]) suggests that a single evaluation (per subject) would statistically show rating bias, we designed Experiment 3 to account for the possibility that rating bias is shown after multiple rating opportunities. We modified our experimental design to a within-subjects study in which each participant rated four simulated gig workers, one from each race/gender pair in a randomized order. A further benefit of this approach is that within-subjects experiments substantially increase statistical power. To maximize the potential for detecting rating bias, we randomly assigned each participant to a single quality level—each participant saw either four high-quality or four low-quality critiques, though each critique was for a different essay. Except for the repeated evaluations, all other aspects of the study (interface, task, validation check, and Mechanical Turk

Table 1: Coefficients for our ordinal mixed-effects model. An asterisk denotes significance.

Variable	Coefficient	<i>p</i> -value
Quality[Good]	2.02	< 0.001*
Race[White]	-0.05	0.81
Gender[Male]	0.01	0.95

Table 2: Threshold cutoffs for our ordinal mixed-effects model.

Thresholds	Estimate	Std. Err
1 2	-2.92	0.34
2 3	-1.59	0.28
3 4	-0.20	0.26
4 5	1.83	0.28

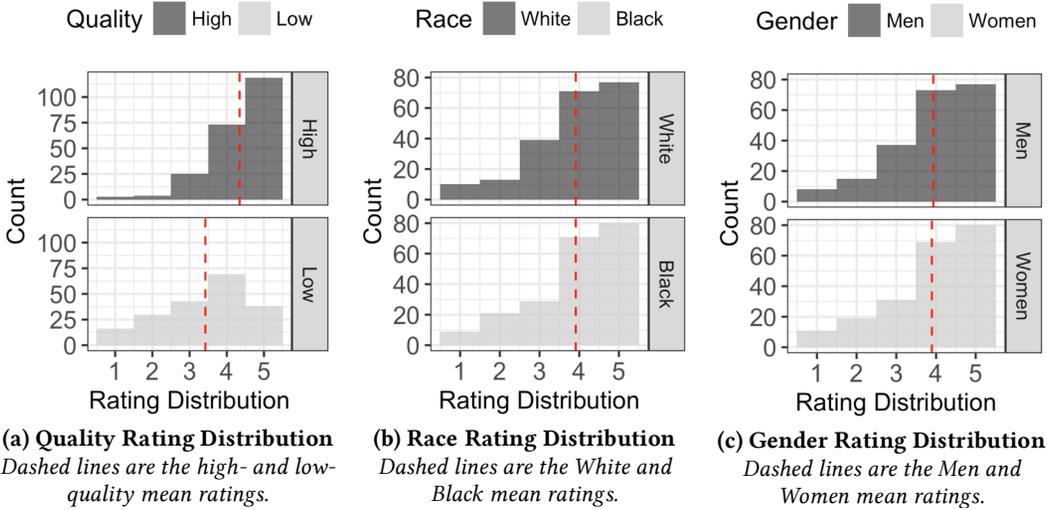


Figure 4: Distributions for each of our variables of interest in Experiment 3. Once again, participant ratings differentiated low and high quality work (Figure 4a), but showed no race (Figure 4b) or gender (Figure 4c) bias.

recruitment criteria) remained unchanged. Our validation check again indicated that participants understood the subject of the essay. To be respectful of our Mechanical Turk participants’ time, we showed them two essays from each example SAT prompt, one long, and one short. This meant that the four evaluations would take approximately 15 minutes, so accordingly we paid \$2.25.

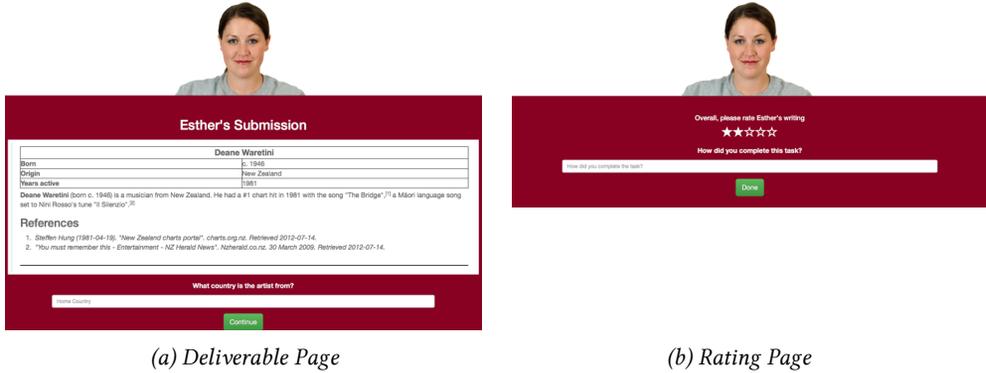
We ran an initial pilot study of 15 participants. We extended this study with another 90 to reach a total of 105 participants (420 ratings). We ran this study one month after our second experiment.

Because of the within-subjects design in this study, the Wilcoxon Rank-Sum test was no longer appropriate. Instead, we used an ordinal, mixed-effects regression approach (using a cumulative-link mixed model or `clmm()` in the `ordinal` R package [13]). We modeled the rating as a dependent variable, and used race and gender of the simulated gig worker and quality of the simulated deliverable as predictors. We included the participant identifier as a random effect to account for individual variance.

5.2.1 Results. As shown in Table 1, Table 2, and Figure 4, we again find that participants distinguish quality (between-subjects), as quality is a significant predictor of the rating. However, race and gender of the gig worker (within-subject) are not, suggesting that participants do not show rating bias along these dimensions. The model coefficients (Table 1) are log-odds ratios and demonstrate that race and gender variables are: (a) not significant, and (b) have small effect sizes compared to the quality variable.

5.3 Experiment 4

None of our first three studies found that participants showed race- or gender-based rating bias. To help understand these results, we sought advice and insight from a Gender Studies scholar. She suggested that the task of evaluating writing critique might be too abstract or unnatural, which could lead to our results. Based



(a) Deliverable Page

(b) Rating Page

Figure 5: Examples of our study interfaces for Experiment 4. Our *deliverable page* (a) and our *rating page* (b).

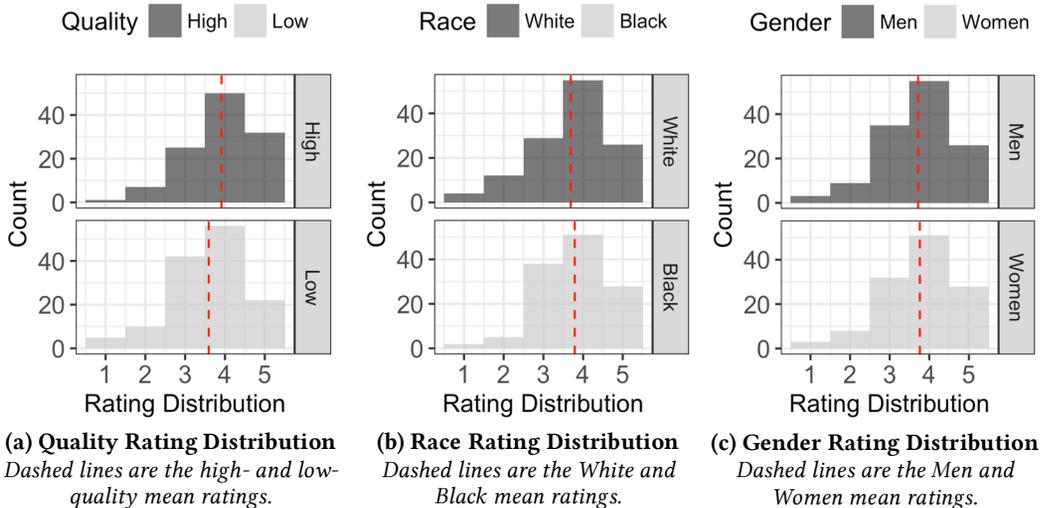


Figure 6: Distributions for each of our variables of interest in Experiment 4. Once again, participant ratings differentiated low and high quality work (Figure 6a), but showed no race (Figure 6b) or gender (Figure 6c) bias.

on her suggestion, we changed the task from evaluating a critique deliverable to evaluating a primary writing deliverable (article writing is another common task for Upwork). In order to minimize the changes we made between studies, we returned to our between-subjects design from Experiment 2.

To generate the simulated pieces of writing, we sampled 100 Wikipedia articles from the Musician Biography Wiki-Project (a project focused on editing musician biographies in Wikipedia). Using an automated quality assessment tool provided by the Wikimedia Foundation (the article quality models that are a part of the Objective Revision Evaluation Service [53]) – and confirming its assessments with those manually provided by the community – we selected four ‘Stub’ class articles and four ‘Start’ class articles as our low- and high-quality deliverables (respectively). Stub class articles are the lowest quality articles on Wikipedia’s quality scale, and start class articles are one class higher [86]. To manage the workload for our participants, we also ensured that each article was between 1,000 and 10,000 bytes of body text. We also made sure these pieces of writing did not look too similar to a Wikipedia article. We did so by scraping the body of these pages and removing all links (retaining the text) and styling. An example can be seen in Figure 5.

Participants were told that we had requested musician biographies for a new website about musicians. We specified that the writing should “provide some meaningful content, and should include referenced material” (language taken directly from the Wikipedia “Start” class requirements [85]). After participants provided a rating, we also asked them “How did you complete the task?”, in order to understand if our use of Wikipedia articles was problematic for this task. Only 3 of our 250 participants mentioned Wikipedia in their responses to this question.

We again recruited Mechanical Turk workers using the same criteria as in Study 2. We ran an initial pilot study of 50 participants and extended to a total of 250 (because we were informed by previous studies, we did not conduct statistical power analysis for this study). We again paid \$0.75, based on a time estimate of 5 minutes. We ran this study several weeks after Study 3.

5.3.1 Results. As in all the other experiments (and directly comparable to Studies 1 and 2), we conducted Wilcoxon Rank-Sum tests, to account for ordinal, right-skewed distributions. The results of this study (Figure 6) show that participants do distinguish between high- and low-quality writing ($W=9190.5$, $p<0.01$). However, also as in all previous experiments, there is not evidence to suggest that participants showed rating differences along the race ($W=8074.5$, $p=0.63$) or gender ($W=7588$, $p=0.68$) dimensions. Participants rated high-quality deliverables an average of 0.32 stars higher than low-quality deliverables, up to 6 times larger than the measured effects in our race or gender conditions.

6 SUMMARY OF CHANGES IN EXPERIMENT DESIGNS

We made a number of changes across the four experiments to robustly replicate our findings and rule out confounds that may have led to our surprising results. We summarize these changes here:

- We began our study of bias in rating writing critiques by holding simulated worker names constant, requesting ratings on the same page as the deliverable, and using a significantly shortened writing critique as our low-quality deliverable.
- Each of these attributes changed in Experiment 2. We moved the rating interface to a different page, used demographically valid names for simulated workers, and made low-quality critiques similar lengths to high-quality critiques. We made these changes to strengthen the effect of our experimental variables in order to accentuate any statistical signal.
- Experiment 3 used the same simulation setup as Experiment 2, but switched the experimental design to within-subjects to test whether multiple ratings were needed before participants showed rating bias.
- Experiment 4 used much of the same setup as Experiment 2. However, we changed the task to evaluating a submitted writing deliverable (another ecologically valid task) rather than an editing deliverable. This experiment tested whether the level of abstraction in online tasks affects whether rating bias is shown.

7 CONFIDENCE IN ABSENCE OF RATING BIAS

Much prior work suggests that participants in our experiments would show rating bias, but they did not. Moreover, specific biases were expected, but we did not find them: simulated white workers were expected to have higher ratings than simulated black workers, and simulated male workers were expected to have higher ratings than simulated female workers. Neither was the case.

Of course, when statistical tests do not find an effect, we cannot conclude that no effect exists. All we can conclude from our studies is that if there were any bias, it is too small to reach significance. Thus, given that our results contrast with prior work, we posed a different and important question: can we demonstrate statistical confidence in *absence of rating bias* in our experiments? To preview: our answer is *yes*.

However, answering this question required thinking through several issues regarding our data and possible statistical techniques. Therefore, prior to presenting results, we walk the reader through our reasoning process.

We first needed to define an “upper bound” for bias: this will let us say that *if any bias exists, it must be less than the upper bound*. We set the bound to be 0.2 on our 5-star scale; for example, this bound would mean

that if a white male received a 5-star rating, a black male submitting the same work would receive no less than a 4.8. We chose this threshold based on analyses of other reputation systems. Horton and Golden [36] did an analysis of ratings in oDesk (now Upwork) and found that 80% of average ratings were above 4.75 stars. Leaked documents suggest that the minimum average rating an Uber driver can have – without being fired – is about 4.6 stars [15]. Thus, a threshold of 0.2 is small enough that a drop of this magnitude would not be fatal to excellent oDesk workers, nor would it force good Uber drivers off the platform. Further, in our studies, the largest (but still not statistically significant) race or gender bias was 0.07, well below the 0.2 threshold. We re-examine our choice of threshold in the Limitations section.

The two-sided tests we have already performed were sufficient to show that we do not have enough evidence to claim a statistically significant bias effect in either direction. In this analysis, our goal is different. We seek to quantify the strength of our evidence against the biases we would expect from prior literature. Therefore, we only need to define an upper bound; any large negative bias would simply be more evidence for the absence of the expected bias.

Since we were defining an upper bound, in more rigorous statistical terms we needed a one-sided confidence bound. If our data were normally distributed and continuous data, our job would be easy: we could use a t-test-based one-sided confidence bound. This is the intuition behind *equivalence testing* (in fact, the `tost()` function in the *equivalence* R package [64] uses just this approach). However, equivalence testing has the same statistical assumptions as the t-test, and our ordinal data violates those assumptions. Therefore, we decided to compare three different statistical approaches to establishing a one-sided confidence bound.

Approach 1 uses a standard t-test based method (despite the statistical assumptions being violated) to estimate the probability of possible effect.

Approach 2 is based on re-sampling simulations with bootstrapping to estimate confidence. We generated 10,000 bootstrapped estimates of the mean of each distribution. We then created 10,000 estimates of the size of the mean shift of the rating distributions. From here, we asked *how frequently is the mean shift less than the 0.2 upper-bound?*

A related Bayesian approach is called Region of Practical Equivalence (ROPE) [44]. ROPE is used to accept or reject the hypothesis that a summary statistic (e.g. mean-shift) is *practically equivalent* to a pre-determined range of values (e.g. -0.2 to 0.2 stars). However, the common ROPE approach does not meet two of our needs. First, as noted, our analysis requires a one-sided confidence upper-bound, and ROPE is designed with two-sided confidence bounds in mind. Second, ROPE is designed simply to accept or reject practical equivalence, and therefore lacks the interpretive power we need to compute the probability of a mean-shift.

Approach 3. Our third approach could be seen as an “alternative ROPE” – we essentially define a one-sided ROPE with an upper-bound of a 0.2 star mean-shift, and then used Bayesian methods to interpret the *probability* of such a shift. We modeled our ratings as independent samples from a categorical distribution with unknown probabilities Θ . We also assumed that every possible rating distribution was equally likely (a uniform prior over Θ). We then computed the posterior distribution of Θ . It is well known [29] that the

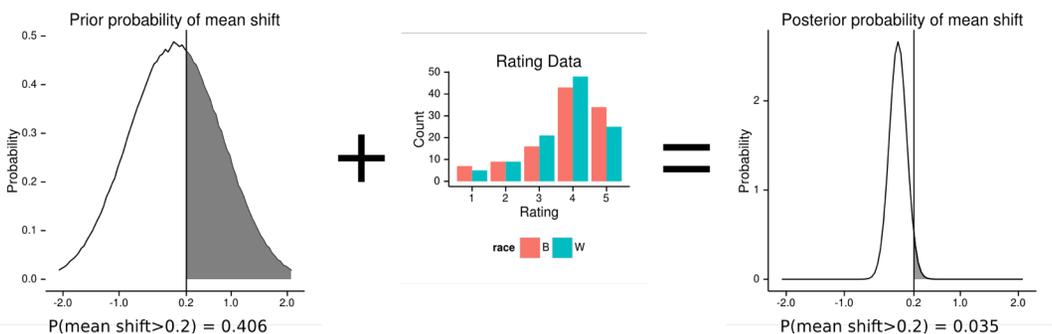


Figure 7: An illustrative example of our absence check method (Study 2 data shown here).

posterior distribution of Θ in this case would be from a Dirichlet distribution $\Theta \text{Dirichlet}(\alpha)$ in which each $a_i = 1 + c_i$, where c_i is the number of times that rating option was observed. Given this posterior distribution we proceeded as before, drawing 10,000 samples of the posterior distribution of each rating distribution to compute a posterior distribution of the mean shift. We then computed the posterior probability that the magnitude of the mean effect shift is less than 0.2.

The Bayesian method (Approach 3) is theoretically more robust to undersampled data. When few ratings are available our uninformed prior will dominate, leading to a conservative distribution of possible bias, and a large posterior probability of a mean shift over 0.2 (as seen in Figure 7). Therefore, we report results concerning absence of bias using the Bayesian method – which we refer to as an *absence check*.

This *absence check* is functionally similar to a one-sided p -value test. We use this method to ask if our data provide sufficient evidence to reject the hypothesis that a mean-shift of at least 0.2 stars exists.

7.1 Experiment 1

According to the Bayesian method, the posterior probability that simulated white workers have a mean rating 0.2 stars greater than simulated black workers is 18%. Using the same method, we estimate that there is a 12% posterior probability of simulated male workers having a mean rating at least 0.2 stars higher than simulated female workers. These results suggest that the absence of rating bias is quite unlikely, rather than merely being unable to reject the null hypothesis.

7.2 Experiment 2

We again used the absence check. Using data from our second study, we estimate that the posterior probability that simulated white workers have an average rating at least 0.2 higher than simulated black workers is approximately 3%, whereas the posterior probability of finding that simulated male workers have an average rating at least 0.2 higher than simulated female workers is approximately 18%. Again, we have marginal (for gender) and significant (for race) confidence in the absence of rating bias.

7.3 Experiment 3

Here we turned to a different (though conceptually similar) Bayesian analysis. Because we had multiple ratings per person, we needed to normalize by participant. Thus, we created a *race bias score* for each participant by taking the two ratings in each race category (two ratings for black gig workers, two ratings for white gig workers) and taking the average for each category. Based on this average, we computed differences between race categories for each participant. We did the same for a *gender bias score* along the male and female categories. In aggregate, this left us with a distribution of per-person bias scores. The question then became *how confident can we be that the mean of this distribution is less than 0.2?*

As before, we considered a t-test based method, a re-sampling based method and a Bayesian method. Since all methods again yielded similar results, we present only the Bayesian method. The per-person bias scores could theoretically be any value from -4.5 to 4.5 in half star increments; empirically, however they only ranged from -2 to 3. Treating this as a categorical distribution like the rating distributions before, we again used a uniform bias distribution to derive a Dirichlet posterior distribution for the probability over the possible bias scores. By drawing 100,000 samples from this posterior distribution and computing the mean of those distributions, we computed the posterior probability that the average bias score is greater than 0.2. There was a 4% posterior probability that the average *race bias score* is above 0.2, and an 8% posterior probability that the average *gender bias score* is above 0.2. We thus again conclude that rating bias is absent, rather than just not identified.

7.4 Experiment 4

We again used the absence check to verify that any rating bias shown is below our 0.2 threshold. Based on this analysis, the posterior probability of finding that simulated white workers have an average rating at least 0.2 higher than simulated black workers is approximately 1%, whereas the posterior probability of finding that

simulated male workers have an average rating at least 0.2 higher than simulated female workers is approximately 3%. We again conclude that rating bias was absent, rather than just not identified.

7.5 Aggregating the Experiments

Our analyses suggest relatively strong confidence that if any ratings bias exists, it is less than 0.2-stars. In three of our four racial bias studies, the absence check yielded greater than 95% confidence that any race-based rating bias is below 0.2 stars, and the fourth trends strongly in this direction. One of our studies showed greater than 95% confidence that gender-based bias was absent, another showed greater than 90% confidence, and the other two still suggested that bias greater than 0.2 stars is unlikely. Moreover, if we look at the actual measured effects (see Figures 2-6), in half the cases the direction of the (non-significant) effects advantage black or women workers, opposite the hypothesized trend and expectations from prior work. Further still, in all our studies the average rating differences between race and gender categories are very small compared to the size of the average rating differences between high- and low-quality deliverables.

8 DISCUSSION

Our results – from four separate studies, across two different experimental designs and two different simulated gig tasks – are at odds with previous work showing race and gender bias. This tension was suggested by Experiment 1, persisted through Experiments 2-4, and the *absence check* solidifies our statistical confidence in the overall finding.

We next present possible interpretations of our results to inform a broader research agenda into the study of bias in ratings-based reputation systems. We offer these interpretations because there *must be* an explanation for our results (due to their robustness and statistical confidence). We use our findings to articulate specific hypotheses to guide future work in this space.

8.1 Interpretations and Hypotheses

8.1.1 Interpretation #1: Third-Party Evaluation of Gig Work. Some have suggested that evaluating work done for someone else may not trigger enough empathy or ownership to show evaluation bias. For example, Bielby [9] argues that in the workplace evaluation domain, any bias measured in simulation experiments is underestimating the actual effect because the decision maker has no institutional context or history. In our case, this would suggest that our participants, who did not request the task and have no direct stake in the deliverable, might be sufficiently disinterested that they did not express any bias.

However, prior work argues against this interpretation. Studies focusing on gender bias [11] and both race and gender bias [10,32] also asked third parties (often students) to evaluate a task. These studies *did find* gender- and race-based bias in the evaluations when controlling for quality (as we did here). In short: analogous studies using third-party evaluators did find bias.

But let us take a deeper look: in contrast to these studies, we elicited third-party ratings *of gig work* – could this explain our result? Consider a study we previously mentioned: Bigoness [10] asked undergraduate students to take on the role of grocery store managers. These ‘managers’ were shown a video of eight ‘employees’ (paid actors) spanning the same race and gender conditions we studied, and were asked to rate the ‘employees’ work. Perhaps the students in Bigoness’ study, by virtue of being ‘managers’ of ‘employees’ (rather than a crowd worker evaluating a task deliverable) took on more ownership over the work they were evaluating than our Mechanical Turk participants did. More generally, perhaps asking a third party to rate a single, ‘one-off’ task is *too disconnected*, leading to no bias being shown.

This leads us to a first hypothesis that should be explored in future work:

Future Work Hypothesis 1 (FWH1): Third-party evaluators do not show race- or gender-based bias in their ratings when evaluating gig work tasks done by and for someone else.

We see this as a compelling direction for future work for two reasons: in addition to potentially helping to explain our results, understanding third-party bias has important applied implications. Specifically, intelligent

systems are increasingly trained using third-party evaluations. Thus, the ways in which gender- and race-based bias are reflected in the ratings and evaluations has the potential to substantially impact these systems and the decisions they make.

8.1.2 Interpretation #2: Sample Selection. Any research aiming for general conclusions must consider potential idiosyncrasies of the study sample. Obvious concerns about our sample center on our choice of Mechanical Turk, including (a) whether it is ever acceptable to use Mechanical Turk workers as participants, (b) how demographically representative Mechanical Turk workers are, and (c) whether or not Mechanical Turk workers ever show bias. However, we do not think these concerns explain our results for the following reasons:

- Sampling from Mechanical Turk is not inherently problematic. Many studies in social computing [42] and other disciplines (e.g. natural language processing [72], political science [5], psychology [12], and economics [37]) have successfully replicated studies that used other types of samples.
- The demographics of Mechanical Turk workers should not affect our findings relative to prior work. Berinsky et al. [5] show that Mechanical Turk is *more representative* of the general population than the undergraduate psychology students who participate in many psychometric studies (including [3]).
- Studies of race- and gender-based biases on Mechanical Turk have been effective. Many studies (e.g. in psychology [27,54], and business [25]) have explored demographic biases using Mechanical Turk and have found results in line with prior literature.

However, a subtler dimension of sample selection may be at play. Turkers may be unique in ways that matter for our research agenda, but not in general. Crowdworkers (like Mechanical Turk workers) may behave distinctively for evaluation tasks like rating the quality of a deliverable. They may be better at considering only the relevant details of a task because of their familiarity with crowdwork incentive structures and practices. This may make them more likely to ignore extraneous concerns such as the race or gender of the person who produced the work a Turker is evaluating. This might explain why our ratings showed no identifiable bias and leads us to a second hypothesis for future work:

Future Work Hypothesis 2a (FWH2a): The familiarity of crowd workers with crowdwork incentives and work practices distinguishes them from the general population and makes them less likely to show race- or gender-based bias when doing a rating task.

There may be another subtle effect of using Turkers as a population: while Turkers are demographically representative of the general population, they *may not* be demographically representative of gig work consumers. Thus, Turkers may evaluate work (simulated or not) differently than the typical gig work customer.

Future Work Hypothesis 2b (FWH2b): Turkers are not representative of gig work consumers, and are therefore the wrong population to serve as evaluators in a controlled experiment that simulates gig work.

Another dimension – which may only be salient when studying gig work systems – is that Turkers are themselves gig workers. Psychology theory suggests that adopting the perspective of another person *does* ameliorate bias towards that person when making ratings [81]. This dimension deserves further study, and we hope others will continue to be careful in generalizing Mechanical Turk studies to the population at large.

8.1.3 Interpretation #3: Controlling Away Effects. It is possible that the representation of our simulated gig workers changed the potential for the ratings provided to show bias. For instance, it may be that because all the photos showed someone wearing a grey t-shirt with a white backdrop, the setting was too neutral, and that attributes of more natural photos (what people wear, where the photo was taken, etc.) may lead to rating bias being shown. Literature does suggest that specifics of photos are relevant in gig work settings. Hannák

et al. [33], for instance, find that in Fiverr, profiles with no photo receive fewer reviews and lower ratings than those with a photo.

Future Work Hypothesis 3 (FWH3): Increasing the ecological validity of how gig workers are represented in photos will show race- and gender-based ratings bias.

There are other attributes of gig work that may lead to bias being shown. For example, the deliverable itself may lead to race or gender bias in ratings. Haswell and Haswell [34] provide evidence to suggest that people do distinguish and attribute author gender, even when the author is not known. Given the historic inequities in education by race [58], similar racial attributions are likely possible based on signals in written text (e.g. [79]). It is not clear from our understanding of the literature what specific mechanisms along these lines may lead to bias in ratings, and we see this as an important direction of future work. Developing ways to study bias that surfaces the correct set of effects and controls for the others will be particularly important.

8.1.4 Interpretation #4: No Bias Exists in Reputation Systems. There is, of course, one more interpretation of our results, and it is much simpler: reputation systems in the gig economy are broadly impervious to significant racial and gender biases. If substantiated, this would be a rare piece of good news in a literature replete with cases of algorithms and sociotechnical systems reflecting and magnifying gender and race biases (e.g. [24,33,45,77]). Further evidence for this interpretation would also raise several critical questions, e.g. What properties of gig economy reputation systems make them resistant to bias? Can we transfer these properties in other domains where bias has been observed?

Before those questions can be answered, however, our findings must be corroborated in additional gig work platforms. To do so, the following hypothesis should be tested:

Future Work Hypothesis 4 (FWH4): Reputation bias systems in many platforms in the gig economy are not subject to major gender or racial biases.

One data point from the Ge et al. [28] paper discussed above increases our confidence that experiments testing this hypothesis will find similar results to ours. The takeaway result of Ge et al.'s experiment was that African Americans and women suffer certain biases in the service they receive in ride hailing platforms. However, a small footnote in the paper reports a partially countervailing result that is particularly relevant to this paper:

“The average star ratings given to African American and white travelers are very similar, indicating that the drivers who accepted the trips and provided star ratings did not provide better or worse ratings based on the [rider's] race”

In other words, while African American riders received worse service, they received similar ratings from drivers as white passengers, reinforcing the results of our four experiments (although certainly not mitigating the bias in service quality).

On the other hand, recent work by Hannák et al. [33] describing an observational study of TaskRabbit and Fiverr presents mixed support for FWH4. This study did not detect a significant bias for some demographics considered here (e.g. women on TaskRabbit) and found significant “reverse” biases in others (e.g. women on Fiverr). However, Hannák et al. did observe a significant bias against black workers on both platforms. While this study was observational rather than a controlled experiment, it does suggest that phenomena related to FWH4 may be complex.

Moving forward, it will be important to run similar studies in a variety of gig work platforms. We hope that this paper can encourage gig work companies to run these studies themselves because, as noted above, doing this type of research is much simpler when researchers have total control of a platform. Absent this control, external researchers may want to find ways to induce control using methods beyond the simulation approach we have used here, e.g., through app reconstructions or modifications (e.g. [52]).

8.2 Other Interpretations

We first presented the interpretations for our results that we believe to be most probable given the nature of our results, theory, and findings from previous work. However, it is also useful to summarize additional possible interpretations that, while initially intuitive to us, we later determined to be highly improbable. In this section, we first discuss each unlikely interpretation and then present the considerable evidence against it.

8.2.1 The Effect of Full-Rating Distribution (Required Ratings). Many rating systems are dominated by very positive (4 or 5) and very negative (1) ratings. The intuition here is that if someone is ambivalent about a rating (such that they would rate a 2, 3, or perhaps 4), they simply choose not to rate.

Our Mechanical Turk workers did not have this option. To assess the validity of this interpretation of our results, we performed a quick analysis (using ratings from Experiment 1), where we excluded ratings of 2 and 3 (but included 4s to guard against undersampling). Using our Bayesian absence test, the posterior probability of a 0.2-star mean shift *decreased* compared to the full dataset. In other words, our original findings were even more robust in this analysis.

Further, there is evidence (e.g. [51]) that suggests that a full (ground-truth) rating distribution does indeed show bias, even when compared to a distribution with a drop-out effect.

8.2.2 The Effect of “Inaccurate” Responses. An initially intuitive interpretation of our data is that the noise in our rating distributions may be masking biases. It is possible that the kinds of biases we expected are driven by the ‘most accurate’ participants in our data, and might be masked by the inclusion of this noise. Therefore, we conducted an additional analysis based on our three between-subjects experiments (Experiment 1, 2, and 4). We only included data from (a) participants in the high-quality condition who provided ratings 3, 4, and 5, and (b) participants in the low-quality condition who provided ratings 1 and 2. As with our primary analyses, *none of these analyses found statistically significant race or gender bias.*

While these analyses do re-affirm that no measurable bias exists in our studies, they are also problematic, for two reasons: (a) ecological validity and (b) ‘selecting on the dependent variable’, which is generally frowned upon [6,71]. With respect to ecological validity, all reputation systems are subject to variation between individual’s ratings. For instance, even the best hotel on TripAdvisor will have a few negative reviews, and this is likely true for the best Upwork freelancers, Airbnb hosts, or Lyft drivers as well. These ratings are not incorrect or inaccurate, they just represent divergent opinions that may focus on different aspects of the rated entity. Excluding data that does not match our expectations of individual’s rating behavior inherently removes important ecological validity from our experiments. In effect, the results of this analysis (regardless of whether we found statistical significance) imagines a world in which all raters behave in a predetermined ‘good’ way, which is, for better or worse, not the world in which reputation systems are employed. As such, we encourage future researchers in this space to consider the full distribution of scores in their primary analyses (as long as validation checks are in place).

8.2.3 Task Design. It is conceivable that writing-related tasks make it particularly difficult for bias to be shown. However, we based these tasks on common categories of tasks on a prominent gig economy platform, Upwork. Moreover, prior research has found bias in participants’ evaluations of writing, even while blinded to gender [34].

8.2.4 Interface Design. While it is possible that some aspect of our interface and study design may have led to absence of rating bias, we took great care to model our experimental interface after the reputation system interfaces common to many gig work platforms (e.g. Uber, Airbnb, Upwork, Fiverr, Rover, and Postmates). Specifically, we were sure to develop ecologically valid implementations of the two key common dimensions of existing gig work reputation systems: the representation of the gig worker and the rating interface. When representing the simulated gig worker, we presented their photo and their name immediately

above or below the photo (depending on the experiment), as is ubiquitous in popular gig work reputation systems. We also used a standard five-star rating scale for the same reason.

8.2.5 Subject-Expectancy Bias. It is plausible that participants behaved differently because they were being studied and showing race- or gender-based bias is socially stigmatized. For this reason, studies of such behaviors attempt to blind subjects whenever possible. While we did not directly interact with participants, participants were shown a consent screen prior to participation. This may have caused participants to modify their rating behavior in a way that is different than they would rate in a live system. However, this would likely have been also true for the many studies we review that observed bias (especially in studies in which participants interacted directly with a person running the study, e.g. the work from psychology on workplace ratings [10]).

9 LIMITATIONS

9.1 Power of the Experiments

One of the limitations of these studies is statistical power. While the absence of rating bias would likely be unaffected by larger studies, with more participants we could have tested a smaller mean-shift threshold.

We selected a 0.2 mean-shift as an upper-bound, based on Horton and Golden's work [36] which showed that approximately 80% of oDesk (now Upwork) workers have average ratings between 4.75 and 5 stars. Further, on Uber, a mean-shift of 0.2 stars will cause a driver to drop 'employability' classes (based on a leaked image reported by Business Insider [15] outlining 'employability' thresholds). Both examples suggest that a mean-shift of 0.2 stars is not fatal to the best workers in the system, but may meaningfully impact workers with lower average ratings. However, recall that the results of our analysis show that an effect even of size 0.2 stars is unlikely to exist (and observed effects were much smaller, and often in the other direction).

Clearly, any amount of bias disadvantages those affected by it, but our results suggest that any effect in our studies is very small (if it exists at all). As we note above, our studies show non-significant effects, our absence checks show that bias above 0.2 stars is unlikely, and statistical intuition suggests that multiple replications give further confidence that our studies show little to no bias.

9.2 Our Own Limitations

In the spirit of Bardzell and Bardzell [2], we discuss our own position and context as these may be relevant to our work.

All the authors are white men (three Ph.D. students, one industry research scientist, and three faculty at two US universities ranging in ages from 20s to 50s). We believe systematic exclusion and disparity in sociotechnical and algorithmic systems are significant problems, and some of us have uncovered and addressed these issues in our prior studies (e.g. [14,35,40,41,46,76,77]). It is from this perspective that we began this work.

When we began, the makeup of our team seemed less important precisely since we assumed we would detect – and then try to mitigate – race and gender bias. When we did not, we were concerned that we might have made design decisions that obscured bias. To address this concern, we consulted with experts in Gender Studies and Ethnic Studies, and used these consultations to inform the design of our subsequent experiments.

9.3 Other Possible Limitations

We did not ask about demographic information of our participants, based on prior work which finds that Turkers in the US are representative of the general US population [5], and that the kinds of biases we study here are exhibited by black and female participants, too; specifically, black people and women exhibit pro-white [58] and pro-male [59] bias on the Implicit Association Test, although to a lesser degree. Based on these findings, we did not ask for demographic information of participants. However, in light of our surprising results, future studies should request information about the demographics of their participants.

10 ADDITIONAL FUTURE WORK

Based on the robustness of our results, experimental exploration of the causes of our surprising results is a critical direction for immediate future work. We outlined a research agenda along these lines in the form of a series of multiple hypotheses. However, this is not the only future work that is suggested by our findings. In this section, we outline longer-term implications for additional research.

10.1 Returning to Our Original Research Agenda

Many of the interpretations outlined above would result in bias manifesting in reputation systems, just not in a fashion that is as obvious as has been the case in past research studies. If these interpretations are valid, this would mean that creating sociotechnical interventions to reduce or eliminate rating bias would be a critical goal for future work. This would also suggest a return to our original research agenda (see Experiment 1): developing rating mechanisms (e.g. multi-factor rating) in which bias is less likely to occur.

10.2 Towards Meta-Analysis in Social Computing

The barriers to cross-study systematic understanding in social computing can be large: in addition to the statistical challenge we tackled in this paper with our Bayesian confidence test, results from different papers must be compared across different online platforms/study contexts, and data often is not published. Further, few negative results papers are accepted in social computing venues, meaning that other studies may have found results like ours, but due to the nature of incentives in our scientific community (and many others), we likely would not know about these studies.

We believe our work provides support for social computing embracing meta-analysis, a method for statistically combining studies (common in psychology e.g. [22]). The statistics, however are the easy part. For such a culture to take hold, our community needs to create venues for publishing rigorous negative results and implement standards for reporting data to support such meta-analyses. Doing so will enable our community to extract greater generalization and confidence from the broad and diverse set of studies conducted in our field.

11 CONCLUSION

In four studies conducted on Mechanical Turk, we present results suggesting that *participants in our experiments do not show rating bias* at or above a 0.2 star mean-shift. Through a *Bayesian confidence test*, we show that the posterior probabilities of measuring bias above that threshold are low enough to be confident in the *absence of rating bias* greater than an average of 0.2 stars. We discuss a number of interpretations of this result and their implications.

ACKNOWLEDGEMENTS

We would like to thank Tala Khanmalek and Jigna Desai for their guidance, which was instrumental to how we thought about our studies. We would also like to thank Robert Kraut, whose thoughtful comments were fundamental to the success of this work. This research was supported by NSF grants IIS-0964695, IIS-1218826, IIS-1111201, IIS-1017697, IIS-1707286, and IIS-1707319.

REFERENCES

- [1] Xavier Amatriain, Josep M. Pujol, and Nuria Oliver. 2009. I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization: Formerly UM and AH (UMAP '09)*, 247–258. https://doi.org/10.1007/978-3-642-02247-0_24
- [2] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, 675–684. <https://doi.org/10.1145/1978942.1979041>
- [3] Sheila K. Bennett. 1982. Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology* 74, 2: 170–179. <https://doi.org/10.1037/0022-0663.74.2.170>
- [4] Alan Benson, Aaron J. Sojourner, and Akhmed Umyarov. 2015. Can Reputation Discipline the Gig Economy? Experimental Evidence from an Online Labor Market. Retrieved February 22, 2016 from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2696299
- [5] Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20, 3: 351–368. <https://doi.org/10.1093/pan/mpr057>

- [6] Richard A. Berk. 1983. An Introduction to Sample Selection Bias in Sociological Data. *American Sociological Review* 48, 3: 386–398. <https://doi.org/10.2307/2095230>
- [7] Edelman Berland. 2014. *Freelancing in America: A national survey of the new workforce*. Freelancers Union and Elance-oDesk.
- [8] Marianne Bertrand and Sendhil Mullainathan. 2003. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*. National Bureau of Economic Research. Retrieved December 14, 2015 from <http://www.nber.org/papers/w9873>
- [9] William T. Bielby. 2000. Minimizing Workplace Gender and Racial Bias. *Contemporary Sociology* 29, 1: 120–129. <https://doi.org/10.2307/2654937>
- [10] William J. Bigoness. 1976. Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology* 61, 1: 80–84. <https://doi.org/10.1037/0021-9010.61.1.80>
- [11] Janine Bosak and Sabine Sczesny. 2011. Gender Bias in Leader Selection? Evidence from a Hiring Simulation Study. *Sex Roles* 65, 3–4: 234–242. <https://doi.org/10.1007/s11199-011-0012-7>
- [12] Krista Casler, Lydia Bickel, and Elizabeth Hackett. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior* 29, 6: 2156–2160. <https://doi.org/10.1016/j.chb.2013.05.009>
- [13] R. H. B. Christensen. 2015. *ordinal—Regression Models for Ordinal Data*.
- [14] Ashley Colley, Jacob Thebaud-Spieker, Allen Yilun Lin, Donald Degraen, Benjamin Fischman, Jonna Häkkinen, Kate Kuehl, Valentina Nisi, Nuno Jardim Nunes, Nina Wenig, and others. 2017. The Geography of Pokémon GO: Beneficial and Problematic Effects on Places and Movement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- [15] James Cook. 2015. Uber's internal charts show how its driver-rating system actually works. *Business Insider*. Retrieved May 24, 2016 from <http://www.businessinsider.com/leaked-charts-show-how-ubers-driver-rating-system-works-2015-2>
- [16] Sara Cook. 2013. Providing Feedback on Student Writing. Retrieved from <http://www.sjsu.edu/aanapisi/docs/ProvidingFeedbackonStudentWritingbySaraCook.pdf>
- [17] Dan Cosley, John Riedl, Shyong K Lam, Istvan Albert, and Joseph A Konstan. 2003. Is seeing believing? In *the conference*, 585. <https://doi.org/10.1145/642712.642713>
- [18] Emily Crockett. 2016. Airbnb and the “sharing economy” still suffer from old-fashioned racism. *Vox*. Retrieved May 22, 2016 from <http://www.vox.com/2015/12/14/10113432/airbnb-sharing-economy-racism>
- [19] Kay Deaux and Janet Taynor. 1973. Evaluation of male and female ability: bias works two ways. *Psychological Reports* 32, 1: 261–262. <https://doi.org/10.2466/pr0.1973.32.1.261>
- [20] Tawanna R. Dillahunt and Amelia R. Malone. 2015. The Promise of the Sharing Economy among Disadvantaged Communities. 2285–2294. <https://doi.org/10.1145/2702123.2702189>
- [21] Josh Dzieza. 2015. The rating game: how Uber and its peers turned us into horrible bosses. *The Verge*. Retrieved May 22, 2016 from <http://www.theverge.com/2015/10/28/9625968/rating-system-on-demand-economy-uber-olive-garden>
- [22] Alice H. Eagly, Mona G. Makhijani, and Bruce G. Klonsky. 1992. Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin* 111, 1: 3–22. <https://doi.org/10.1037/0033-2909.111.1.3>
- [23] Benjamin G. Edelman and Michael Luca. 2014. Digital Discrimination: The Case of Airbnb. com. *Harvard Business School NOM Unit Working Paper*, 14–054. Retrieved March 21, 2015 from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2377353
- [24] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2015. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *Harvard Business School NOM Unit Working Paper*, 16–069. Retrieved January 13, 2016 from <http://www.benedelman.org/publications/airbnb-guest-discrimination-2015-12-09.pdf>
- [25] Benjamin A. Everly, Miguel M. Unzueta, and Margaret J. Shih. 2015. Can Being Gay Provide a Boost in the Hiring Process? Maybe If the Boss is Female. *Journal of Business and Psychology* 31, 2: 293–306. <https://doi.org/10.1007/s10869-015-9412-y>
- [26] Jiayang Fan. 2015. How China Wants to Rate Its Citizens. *The New Yorker*. Retrieved May 22, 2016 from <http://www.newyorker.com/news/daily-comment/how-china-wants-to-rate-its-citizens>
- [27] Rick M. Gardner, Dana L. Brown, and Russell Boice. 2012. Using Amazon's Mechanical Turk website to measure accuracy of body size estimation and body dissatisfaction. *Body Image* 9, 4: 532–534. <https://doi.org/10.1016/j.bodyim.2012.06.006>
- [28] Yanbo Ge, Christopher R. Knittel, Don MacKenzie, and Stephen Zoepf. 2016. *Racial and Gender Discrimination in Transportation Network Companies*. National Bureau of Economic Research. Retrieved November 15, 2016 from <http://www.nber.org/papers/w22776>
- [29] Andrew Gelman, John B Carlin, Hal S Stern, David Dunson, Aki Vehtari, and Donald B Rubin. 2014. *Bayesian data analysis*. Taylor & Francis.
- [30] Claudia Goldin and Cecilia Rouse. 1997. *Orchestrating Impartiality: The Impact of*. National Bureau of Economic Research. <https://doi.org/10.3386/w5903>
- [31] Jeffrey H. Greenhaus and Saroj Parasuraman. 1993. Job Performance Attributions and Career Advancement Prospects: An Examination of Gender and Race Effects. *Organizational Behavior and Human Decision Processes* 55, 2: 273–297. <https://doi.org/10.1006/obhd.1993.1034>
- [32] W. Clay Hammer, Jay S. Kim, Lloyd Baird, and William J. Bigoness. 1974. Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology* 59, 6: 705–711. <https://doi.org/10.1037/h0037503>
- [33] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. Retrieved December 14, 2016 from http://claudiawagner.info/publications/cscw_bias_olm.pdf
- [34] Richard H. Haswell and Janis Tedesco Haswell. 1996. Gender bias and critique of student writing. *Assessing Writing* 3, 1: 31–83. [https://doi.org/10.1016/S1075-2935\(96\)90004-5](https://doi.org/10.1016/S1075-2935(96)90004-5)
- [35] Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *CHI '10: 28th International Conference on Human Factors in Computing Systems* (CHI '10), 291–300. <https://doi.org/10.1145/1753326.1753370>
- [36] John Horton and Joseph Golden. 2015. Reputation Inflation: Evidence from an Online Labor Market. *Work. Pap., NYU*. Retrieved February 22, 2016 from http://squash.tamu.edu/common/files/workshops/Theory and Experimental Economics/2015_3_5_John_Horton.pdf
- [37] John J. Horton, David G. Rand, and Richard J. Zeckhauser. 2011. The online laboratory: conducting experiments in a real labor market. *Experimental Economics* 14, 3: 399–425. <https://doi.org/10.1007/s10683-011-9273-9>
- [38] Allen I. Huffcutt and Philip L. Roth. 1998. Racial group differences in employment interview evaluations. *Journal of Applied Psychology* 83, 2: 179–189. <https://doi.org/10.1037/0021-9010.83.2.179>

- [39] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 611–620. <https://doi.org/10.1145/2470654.2470742>
- [40] Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at Home on the Range: Peer Production and the Urban/Rural Divide. In *2016 CHI Conference on Human Factors in Computing Systems*, 13–25. <https://doi.org/10.1145/2858036.2858123>
- [41] Isaac L. Johnson, Connor J McMahon, Johannes Schöning, and Brent Hecht. 2017. The Effect of Population and “Structural” Biases on Social Media-based Algorithms – A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI 2017)*. <https://doi.org/http://dx.doi.org/10.1145/3025453.3026015>
- [42] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*, 453–456. <https://doi.org/10.1145/1357054.1357127>
- [43] Daniel Kluger, Tien T. Nguyen, Michael Ekstrand, Shilad Sen, and John Riedl. 2012. How many bits per rating? In *Proceedings of the sixth ACM conference on Recommender systems*, 99–106. Retrieved December 15, 2015 from <http://dl.acm.org/citation.cfm?id=2365974>
- [44] John K. Kruschke. 2011. Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science* 6, 3: 299–312. <https://doi.org/10.1177/1745691611406925>
- [45] Shyong Tony K Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl. 2011. WP:clubhouse? In *the 7th International Symposium*, 1. <https://doi.org/10.1145/2038558.2038560>
- [46] Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. 2011. WP:Clubhouse?: An Exploration of Wikipedia’s Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*, 1–10. <https://doi.org/10.1145/2038558.2038560>
- [47] C. Lampe and R. K. Garrett. 2007. It’s all news to me: The effect of instruments on ratings provision. In *40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007*, 180b–180b. <https://doi.org/10.1109/HICSS.2007.308>
- [48] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. 1603–1612. <https://doi.org/10.1145/2702123.2702548>
- [49] Nancy Leong. 2014. The sharing economy has a race problem. *Salon*. Retrieved May 22, 2016 from http://www.salon.com/2014/11/02/the_sharing_economy_has_a_race_problem/
- [50] Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods* 47, 4: 1122–1135. <https://doi.org/10.3758/s13428-014-0532-5>
- [51] Kathleen M Mazor, Brian E Clauser, Terry Field, Robert A Yood, and Jerry H Gurwitz. 2002. A Demonstration of the Impact of Response Bias on the Results of Patient Satisfaction Surveys. *Health Services Research* 37, 5: 1403–1417. <https://doi.org/10.1111/1475-6773.11194>
- [52] Connor McMahon, Isaac Johnson, and Brent Hecht. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies.
- [53] Meta. 2016. *Objective Revision Evaluation Service – Meta, discussion about Wikimedia projects*. Retrieved from https://meta.wikimedia.org/w/index.php?title=Objective_Revision_Evaluation_Service&oldid=15597359
- [54] Joel T. Nadler and Katie M. Kufahl. 2014. Marital status, gender, and sexual orientation: Implications for employment hiring decisions. *Psychology of Sexual Orientation and Gender Diversity* 1, 3: 270–278. <https://doi.org/10.1037/sgd0000050>
- [55] Meghan Neal. 2014. The Sharing Economy Gets Accused of Racism. *Motherboard*. Retrieved May 22, 2016 from <http://motherboard.vice.com/blog/are-airbnb-users-really-being-racist>
- [56] Tien T. Nguyen, Daniel Kluger, Ting-Yu Wang, Pik-Mai Hui, Michael D. Ekstrand, Martijn C. Willemsen, and John Riedl. 2013. Rating Support Interfaces to Improve User Experience and Recommender Accuracy. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*, 149–156. <https://doi.org/10.1145/2507157.2507188>
- [57] Veronica F. Nieva and Barbara A. Gutek. 1980. Sex Effects on Evaluation. *Academy of Management Review* 5, 2: 267–276. <https://doi.org/10.5465/AMR.1980.4288749>
- [58] Brian A. Nosek, Mahzarin R. Banaji, and Anthony G. Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice* 6, 1: 101–115. <https://doi.org/10.1037/1089-2699.6.1.101>
- [59] Brian A. Nosek, Mahzarin R. Banaji, and Anthony G. Greenwald. 2002. Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology* 83, 1: 44–59. <https://doi.org/10.1037/0022-3514.83.1.44>
- [60] Regina Pingitore, Bernard L. Dugoni, R. Scott, and Bonnie Spring. 1994. Bias against overweight job applicants in a simulated employment interview. *Journal of Applied Psychology* 79, 6: 909–917. <https://doi.org/10.1037/0021-9010.79.6.909>
- [61] Giovanni Quattrone, Davide Proserpio, Daniele Quercia, Licia Capra, and Mirco Musolesi. 2016. Who Benefits from the “Sharing” Economy of Airbnb? *arXiv:1602.02238 [physics]*. Retrieved February 25, 2016 from <http://arxiv.org/abs/1602.02238>
- [62] Noopur Raval and Paul Dourish. 2016. Standing Out from the Crowd: Emotional Labor, Body Labor, and Temporal Labor in Ridesharing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*, 97–107. <https://doi.org/10.1145/2818048.2820026>
- [63] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental Economics* 9, 2: 79–101. <https://doi.org/10.1007/s10683-006-4309-2>
- [64] Andrew Robinson. 2016. *equivalence: Provides Tests and Graphics for Assessing Tests of Equivalence*. Retrieved July 31, 2016 from <https://cran.r-project.org/web/packages/equivalence/index.html>
- [65] Alex Rosenblat and Luke Stark. 2015. *Uber’s Drivers: Information Asymmetries and Control in Dynamic Work*. Social Science Research Network, Rochester, NY. Retrieved November 10, 2015 from <http://papers.ssrn.com/abstract=2686227>
- [66] Cort W. Rudolph, Charles L. Wells, Marcus D. Weller, and Boris B. Baltes. 2009. A meta-analysis of empirical studies of weight-based bias in the workplace. *Journal of Vocational Behavior* 74, 1: 1–10. <https://doi.org/10.1016/j.jvb.2008.09.008>
- [67] Alan Said, Brijnesh J. Jain, Sascha Narr, and Till Plumbaum. 2012. Users and Noise: The Magic Barrier of Recommender Systems. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization (UMAP'12)*, 237–248. https://doi.org/10.1007/978-3-642-31454-4_20
- [68] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. Retrieved August 24, 2015 from <http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf>
- [69] Christian Sandvig, Kyrasto Karahalios, Alan Mislove, and First Look Media Works, Inc. undecided. *Sandvig v. Lynch*. Retrieved from https://www.aclu.org/sites/default/files/field_document/cfaa_complaint.pdf

- [70] Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. 2015. Turkers, Scholars, “Arafat” and “Peace”: Cultural Communities and Algorithmic Gold Standards. 826–838. <https://doi.org/10.1145/2675133.2675285>
- [71] Jeremy C. Short, David J. Ketchen, and Timothy B. Palmer. 2002. The Role of Sampling in Strategic Management Research on Performance: A Two-Study Analysis. *Journal of Management* 28, 3: 363–385. <https://doi.org/10.1177/014920630202800306>
- [72] Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ’08)*, 254–263. Retrieved August 1, 2016 from <http://dl.acm.org/citation.cfm?id=1613715.1613751>
- [73] Jennifer Stark and Nicholas Diakopoulos. 2016. Uber seems to offer better service in areas with more white people. That raises some tough questions. *The Washington Post*. Retrieved March 10, 2016 from <https://www.washingtonpost.com/news/wonk/wp/2016/03/10/uber-seems-to-offer-better-service-in-areas-with-more-white-people-that-raises-some-tough-questions/>
- [74] Bogdan State, Bruno Abrahao, and Karen Cook. 2016. Power Imbalance and Rating Systems. In *Tenth International AAAI Conference on Web and Social Media*.
- [75] Toshiro Tango. 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in medicine* 17, 8: 891–908.
- [76] Jacob Thebault-Spieker, Loren G. Terveen, and Brent Hecht. 2015. Avoiding the South Side and the Suburbs: The Geography of Mobile Crowdsourcing Markets. In *CSCW 2015*, 265–275. <https://doi.org/10.1145/2675133.2675278>
- [77] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Towards a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit. *ACM TOCHI*.
- [78] United States Congress. 1986. *Computer Fraud and Abuse Act*. Retrieved from <https://www.law.cornell.edu/uscode/text/18/1030>
- [79] David A. Waldman and Bruce J. Avolio. 1991. Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology* 76, 6: 897–901. <https://doi.org/10.1037/0021-9010.76.6.897>
- [80] Karen K. Yuen. 1974. The two-sample trimmed t for unequal population variances. *Biometrika* 61, 1: 165–170. <https://doi.org/10.1093/biomet/61.1.165>
- [81] Mark P. Zanna and James M. Olson. 2013. *Advances in Experimental Social Psychology, Volume 48*. Academic Press, San Diego.
- [82] Georgios Zervas, Davide Proserpio, and John Byers. 2015. *A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average*. Social Science Research Network, Rochester, NY. Retrieved May 27, 2016 from <http://papers.ssrn.com/abstract=2554500>
- [83] 2015. Essay Sample 1 Bogard. *SAT Suite of Assessments*. Retrieved May 27, 2016 from <https://collegereadiness.collegeboard.org/sample-questions/essay/1>
- [84] 2015. Essay Sample 2 Gioia. *SAT Suite of Assessments*. Retrieved May 27, 2016 from <https://collegereadiness.collegeboard.org/sample-questions/essay/2>
- [85] 2015. Wikipedia:Version 1.0 Editorial Team/Assessment. *Wikipedia, the free encyclopedia*. Retrieved May 27, 2016 from https://en.wikipedia.org/w/index.php?title=Wikipedia:Version_1.0_Editorial_Team/Assessment&oldid=690783073
- [86] 2016. Wikipedia:WikiProject Wikipedia/Assessment. *Wikipedia*. Retrieved April 27, 2017 from https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Wikipedia/Assessment&oldid=754983998
- [87] What is a Mechanical Turk Master? *FAQ > Worker Web Site FAQs*. Retrieved May 27, 2016 from https://www.mturk.com/mturk/help?helpPage=worker#what_is_master_worker
- [88] West Virginia Department of Education Teach 21 Writing Rubrics. Retrieved May 27, 2016 from <https://wvde.state.wv.us/teach21/writingrubrics/>
- [89] Most Popular Baby Names by Sex and Mother’s Ethnic Group, New York City. *NYC Open Data*. Retrieved May 27, 2016 from <https://data.cityofnewyork.us/Health/Most-Popular-Baby-Names-by-Sex-and-Mother-s-Ethnic/25th-nujf>

Received June 2017; revised July 2017; accepted November 2017

A WRITING SAMPLE EXAMPLES

A1 Examples of Writing Used in Experiment 1

This response demonstrates some comprehension of the source text, although the writer's understanding of Bogard's central idea isn't conveyed until the latter part of the essay, where the writer indicates that Bogard includes *details facts about human body, animals and about mother nature that he can use to support his idea of not using so much light at night and how we need darkness*. Prior to this, the writer has included details from the text, but without contextualizing these details within Bogard's broader argument, suggesting that the writer is relaying ideas from the text without much understanding of how they contribute to the whole. For example, the writer mentions the health problems cited in the text, that working the night shift is classified as bad, and that light costs are high, but doesn't explain how these points relate to Bogard's main claim that we must preserve natural darkness. On the whole, this essay displays only a partial understanding of the source text.

In this essay, the writer has merely identified aspects of Bogard's use of evidence without explaining how the evidence contributes to the argument. The writer notes that Bogard's text *talks about so much facts about sleeping how so little can effect us health wise examples like getting sleep disorders, diabetes, obesity, cardiovascular disease and depression. This facts helps people persuade the audience*. Other than identifying these as persuasive facts, however, the writer does nothing to indicate an understanding of the analytical task. The writer again mentions persuasion before the conclusion of the essay (*With these features he can persuade the audience because people dont know why darkness can be good for us*), but once again, there is no explanation of how or why these features are persuasive. Thus, the essay offers inadequate analysis of Bogard's text.

This response demonstrates little cohesion and inadequate skill in the use and control of language. From the outset, problems with language control impede the writer's ability to establish a clear central claim (*Bogard builds an argument to persuade his audience about what he is concering about and feels it important to take care about*). The response also lacks a recognizable introduction and conclusion, and sentences are strung together without a clear progression of ideas (for much of the response, the writer merely lists claims Bogard makes). The response also lacks variety in sentence structures, in part because of repetitive transitions. (For example, *he also claims* is used two sentences in a row in this brief response). Weak control of the conventions of standard written English, coupled with vague word choice, undermine the quality of writing. Overall, this response has demonstrated inadequate writing skill.

This response demonstrates some comprehension of the source text, although the writer's understanding of Bogard's central idea isn't conveyed until the latter part of the essay, where the writer indicates that Bogard includes *details facts about human body, animals and about mother nature that he can use to support his idea of not using so much light at night and how we need darkness*. On the whole, this essay displays only a partial understanding of the source text.

In this essay, the writer has merely identified aspects of Bogard's use of evidence without explaining how the evidence contributes to the argument. Thus, the essay offers inadequate analysis of Bogard's text.

This response demonstrates little cohesion and inadequate skill in the use and control of language. Overall, this response has demonstrated inadequate writing skill.

(a) *High-Quality Feedback*

(b) *Low-Quality Feedback*

Figure A1: Examples of our two quality conditions in Experiment 1.

A2 Examples of Writing Used in Experiments 2 & 3

This response demonstrates little comprehension of Gioia's text. The response is almost entirely composed of ideas and phrases taken directly from the passage. Although the writer does demonstrate that the writer has read the passage by referring to the Wired article (the writer conveys that employers are looking for aptitudes deadely literally in character: not "linear, logical, and analytical talents") and including a notable point in the passage (*Reading is not [a] timeless universal capability*), there is very little evidence that the writer actually understands Gioia's main argument, and the response is limited to presenting seemingly randomly chosen details from the passage. Overall, this response demonstrates inadequate reading comprehension.

The writer demonstrates no real understanding of the analytical task and offers no discernible analysis of the source text. The writer does not describe Gioia's use of evidence, reasoning, or stylistic or persuasive elements, nor does the writer attempt to explain the importance of these elements to Gioia's argument. The brief response is largely comprised of ideas and phrases taken directly from the passage. Overall, this response demonstrates inadequate analysis.

The responses is almost entirely composed of ideas and phrases Taken directly from the passage. Overall, this response does not demonstrate inadequate reading comprehension. Although the writer does demonstrating that the writer has read the passage by including a Notable point in the passage there is very little evience that the writer actually uderstands Gioia's main argument and the response is not limited to presenting seemingly randomly chosen details from the passage.

The writer does describing use of evidence, reasning, or stylistic or persuasive elements, nor does the writers attempt to explain the Importance Overall this response demonstrates inadequate analysis. The Brief response is largely compised of ideas and phrases not taken directly from the passage.

Overall, this response demonstrating inadequate control. The writer includes a clear central claims or controlling idea and istead jumps into Repeating ideas and phrases from the passage There is no real organization or progression of ideas. Furthermore there is evidence of the writer's own writing abilty since Most of the response is taken directly from Gioia's text.

(a) *High-Quality Feedback*

(b) *Low-Quality Feedback*

Figure A2: Examples of our two quality conditions in Experiment 2.