

MIT Open Access Articles

An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Jahanbakhsh, Farnaz, Cranshaw, Justin, Counts, Scott, Lasecki, Walter and Inkpen, Kori. 2020. "An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender."

As Published: <https://doi.org/10.1145/3313831.3376860>

Publisher: ACM|CHI Conference on Human Factors in Computing Systems USB

Persistent URL: <https://hdl.handle.net/1721.1/145659>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender

Farnaz Jahanbakhsh^{¶†}, Justin Cranshaw[†], Scott Counts[†], Walter S. Lasecki[§], Kori Inkpen[†]
 Massachusetts Institute of Technology[¶], Microsoft Research[†], University of Michigan - Ann Arbor[§]
 farnazj@mit.edu, justin@cranshaw.me, [counts, kori]@microsoft.com, wlasecki@umich.edu

ABSTRACT

We study how the ratings people receive on online labor platforms are influenced by their performance, gender, their rater's gender, and displayed ratings from other raters. We conducted a deception study in which participants collaborated on a task with a pair of simulated workers, who varied in gender and performance level, and then rated their performance. When the performance of paired workers was similar, low-performing females were rated lower than their male counterparts. Where there was a clear performance difference between paired workers, low-performing females were preferred over a similarly-performing male peer. Furthermore, displaying an average rating from other raters made ratings more extreme, resulting in high performing workers receiving significantly higher ratings and low performers lower ratings compared to when average ratings were absent. This work contributes an empirical understanding of when biases in ratings manifest, and offers recommendations for how online work platforms can counter these biases.

Author Keywords

Digital Ratings, Gender Discrimination, Social Mimicry, Bias in Ratings, Bias in Gig Platforms

CCS Concepts

•Human-centered computing → User studies; Empirical studies in HCI;

INTRODUCTION

Many people are turning to online platforms such as Uber, Upwork, and Mechanical Turk as their primary source of employment. These platforms often solicit performance ratings from employers based on the quality of service they provide so that future service requesters can make informed decisions on who to hire or how much to tip [8, 35] and so that platform operators can algorithmically manage their workforce, for example by banning workers below a certain rating or ranking search results by rating [31]. Thus, ratings are paramount to the success and livelihoods of workers on these platforms. Since ratings are by nature subjective, they could embed societal biases that

disproportionately harm or favor certain demographics. Such demographic biases may manifest differently at different performance levels. For instance, raters might be influenced by their biases only when their employed worker is not performing well, rating such workers lower if they belong to certain demographic groups than others. Furthermore, biases in these systems may be perpetuated by social influence and herding behavior [6, 60, 50, 52], with ratings given by prior, possibly biased users reinforcing the appeal (or lack thereof) of certain groups of workers, potentially even influencing the ratings of future unbiased raters.

In this work, we study how online workers' ratings are impacted by their gender and the gender of their raters, and how potential rating biases vary across workers' different performance levels. In addition, we examine the effects of displaying aggregated ratings of prior work on workers' future ratings.

Demographics biases have been extensively studied in the prior literature [53, 43, 9, 10, 21]. With the rise of online labor platforms, recent research has investigated such biases in this newer context [66, 33, 22, 56]. Most such studies, however, have been observational. While these offer invaluable contributions grounded in real-world systems, they may lack a more nuanced insight into the circumstances where biases surface that controlled experiments and simulation studies could provide. For instance, addressing the questions of our study using data from existing labor platforms is difficult because we cannot disentangle a worker's true performance from their socially-assigned rating. In addition, controlled settings allow us to experiment with mechanisms to counteract the biases we uncover. To rigorously study the interaction between performance and demographic biases, we developed an experimental framework where we can control and manipulate the true performance of workers in a simulated collaborative task. In our experiments, the simulated workers are overseen in real-time by study participants, who, operating in a managerial context, are asked at the conclusion of the task to rate their performance. In each trial, the simulated workers are drawn from one of 4 experimental profiles (2 genders \times 2 performance levels), with the gender dimension mapping on to the worker profile name and photo, and the performance dimension controlling the quality of their simulated task execution.

The designs of our experimental framework and the collaborative task were influenced by prior related findings. In a recent simulation experiment designed to uncover demographic biases in ratings of online work, Thebault-Spieker et al. asked Mechanical Turk workers to rate essays with different

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.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-6708-0/20/04 ...\$15.00.

<http://dx.doi.org/10.1145/3313831.3376860>

qualities attributed to different randomized demographic profiles [63]. While they were not able to find such biases, they hypothesized that “evaluating work done for someone else may not trigger enough empathy or ownership to show evaluation bias.” Guided by this hypothesis, we designed our tasks so that raters would not only believe they were evaluating real people, but also be personally impacted by the performance of the workers. We conducted a deception study, in which study participants were assigned two simulated workers as teammates with whom they performed a real-time collaborative image labeling task. We designed the task such that simulated workers’ delay in performing the task and the quality of their generated labels would impact the performance of the whole team. After completing the task, participants assessed each simulated worker. We expected that the collaborative nature of the task would increase participants’ stake in the outcome and provide them with context and history, the lack of which has been suggested to underestimate workplace evaluation biases in simulated experiments [34].

Our experimental design was also inspired by the importance of relative judgments for ratings calibration [37]. We pair each participant up with two workers who may have equal or different performance, so that participants can assess worker quality relative to one another. We hypothesized that if all participants only ever view one worker in isolation, they may not discern their performance level as we had intended it. By implicitly inducing this comparison we hoped to enable participants to calibrate their ratings and form a relative judgment on the workers. The relative judgment component of this experiment could make it also applicable to those scenarios where applicants of a limited resource are compared against one another, for instance, when applying for jobs or loans [11, 56].

We explore this setting across three experimental conditions that help us understand the impacts of social signals on ratings of digital laborers as well as their interactions with potential gender biases: *No Ratings*, in which participants could not see ratings of worker profiles given by other raters, *True Ratings*, in which the true aggregated worker ratings were presented, and *Reversed Ratings*, in which we flipped the performance of workers so that it would not match their rating history.

Our results paint a picture of ratings bias in online work that is complex, nuanced, and at times surprising. We found that the gender of workers indeed impacts the ratings they receive. However, this gender bias only manifests in ratings of low performers, and its precise nature is dependent on whether the low performing worker was paired with a fellow low performing coworker, or whether their coworker was a high performer. When two low performers were paired together, low performing male workers received higher ratings than their low performing female counterparts ($d = 0.16 - 0.22$). Drilling into the sources of this bias revealed that this difference was in fact driven by female raters giving low performing female workers lower ratings than low performing males ($d = 0.21$). This observation aligns with findings of previous studies showing in-group biases by women found in other contexts [23, 51, 28]. However, surprisingly, we find a mirror image in trials where a low performer was paired with a high performer. In

such cases, low performing females fared better than low performing males ($d = 0.15$), and this favoritism was driven by male raters ($d = 0.17 - 0.23$). In addition, we observed that the presence of prior ratings negatively affected the ratings of low performers ($d = 0.24$) and positively impacted the ratings of high performers ($d = 0.40$), when workers had similar performance levels, suggesting mimicry of prior ratings also exists on digital labor platforms.

The main contribution of this work is an empirical understanding of circumstances in which biases manifest in ratings of digital laborers. We study representation of prior ratings, gender of workers, and gender of raters as sources of biases. In addition, this work contributes recommendations for mitigating these biases on online labor platforms.

RELATED WORK

We situate our work in the context of prior work on digital labor markets, ratings of online laborers, and gender bias.

Background on Digital Labor Markets

Online labor markets provide services spanning from networked accommodations to short-term contracts with software designers or documents translators. The appealing opportunities for consumers of these markets to find on-demand labor, however, are complicated by the precarious relationships and one-sided power dynamics between platform workers and operators, and what this might portend about the future of work.

A body of work has been studying the dynamics of these markets and their complex issues. Alkhatib et al. examine crowdwork as a resurgence of piecework to understand the limits of on-demand labor and laborers’ relationship to their work [4]. Glöss et al. investigate how digital ride-sharing has added new demands to the labor that is expected from drivers [27]. Researchers have also studied the uneven benefits that these markets offer to users with different socioeconomic status [64, 20]. Another body of work attempts to empower online laborers and address the issues affecting them. To help mitigate the power asymmetry that existed between employers and workers on Mechanical Turk, Irani and Silberman developed Turkopticon, allowing workers to post reviews of employers [36]. Salehi et al. introduce Dynamo to help Mechanical Turk workers assemble towards a collective action [59].

In this work, we examine how ratings of digital workers can be impacted by the implicit biases that work requesters may harbor. Understanding these biases is the first step towards designing platform that can correct them.

Ratings in Digital Labor Markets

Labor platforms provide service requesters with a large pool of online laborers as well as signals of their performance to help differentiate between them. These signals have strong impacts on workers’ livelihood, for instance affecting their probability of being hired [8, 35] and their future earnings [54].

Because these signals play such an important role in the livelihood of online workers, the impact of any factor beyond performance on them needs to be studied as a potential source of bias, for instance whether ratings are public or private [25] or

power imbalance resulting from resource ownership [3]. Another factor is how signals of past performance are presented to future evaluators. While studies have examined effects of exposing such signals on ratings of online merchandise or content [6, 60, 50, 52], less is known about how presenting previous ratings of online workers affects their future ratings.

Our work aims to bridge this gap by soliciting ratings in both the absence and the presence of previous ratings of workers as well as when their previous ratings do not match their current performance. The latter case may occur when workers with previously poor performance who have boosted their quality of work, for instance through gaining experience, are trying to overcome their initial ratings or when those who have made an initial good impression fail to maintain high quality work.

Gender Discrimination

Gender discrimination in traditional markets is well studied [53, 5, 38, 61, 21]. Gender biases have been found in many contexts including organizational compensations [10], availability of business loans to men and women [24], car dealership prices quoted to male buyers versus females [7], and even science faculty's ratings of application materials that were randomly assigned either a female or a male name [51].

With the rise of digital markets, another strand of research has been investigating gender biases in the online world. The unique characteristics of online labor platforms including the relatively short period of employment and potential social influence of other raters can impact and reshape the nature of biases on such platforms and therefore, it is important to investigate how biases transfer to this context. In a case study on two freelancing marketplaces, Hannak et al. report that worker's race and gender impact the social feedback that they receive, although the impact is different on each platform [33]. Differential treatment of genders has also been studied in the context of e-commerce websites [41], online loans [56], code contributions on Github [62], and recruitment platforms [11].

While gender bias has been studied extensively, the circumstances under which it surfaces on online labor platforms are less known. In a simulation study to uncover demographics biases and their relation to the quality of work, Thebault-Spieker et al. asked turkers to rate essays with different qualities attributed to different demographic profiles. However, they did not find any demographics biases. One reason they cite for this result is that such biases are unlikely to manifest when evaluating gig work tasks done by and for someone else [63].

We extend this body of work to better understand when gender biases on online markets appear or are stifled. We conduct a controlled experiment in which we study the effect of both worker and rater gender on ratings of online workers, controlling for workers' true performance. We also expand current knowledge by investigating how displaying, withholding, or manipulating prior ratings reshape these biases. We attempt to recreate elements of real labor platforms that may evoke implicit biases. We designed the task such that the subjects believed their utility would be impacted by the performance of the workers. Therefore, we are able to elicit ratings that are not hypothetical, but rather represent actual ratings in the

wild. Furthermore, by presenting workers side by side during a collaborative task, we attempt to draw relative judgments from our participants that we expect to be swayed by their implicit biases. The relative judgment component can also provide insight into bias in settings where applicants of a limited resource are compared against each other on an online platform. Examples of such scenarios include online peer-to-peer loans [29], hiring gig workers [11], and crowdfunding [49].

RESEARCH QUESTIONS

Given that we assume that a worker's performance will impact the ratings they receive, our work seeks to explore the following research questions:

- **RQ1:** In a group task, does comparative performance of other workers impact the ratings that a worker receives?
- **RQ2:** Does revealing previous ratings impact the future ratings an online worker receives?
- **RQ3:** Does an online worker's gender affect their rating?
- **RQ4:** Does a rater's gender impact the ratings they give?

METHOD

To investigate these research questions, we designed and developed a platform where our participants could interact with and then evaluate a set of simulated workers. Participants collaborated with simulated workers on an image labeling task. The purpose of the task was first, to reveal the performance of workers in a natural setting and expose participants to the impact of their collaborators' high or low performance; and second, to serve as a pretext for soliciting ratings afterwards.

Our study was approved by our Institutional Review Board. Because the study involved a deception component, we presented the project as a collaborative image labeling task to study user interaction with our technology. After the study was over, we debriefed participants via a message explaining that the two workers they had collaborated with had been in fact simulated and we had varied their performance and gender. We also described that the goal of the study was to understand bias in ratings based on the systematic variations and that any publication of the research would report only aggregate summaries of results, with no identifiable information.

Task

Study participants interacted with simulated workers on a "Collaborative Image Labeling" task. After consenting to the study, each participant was assigned two simulated workers who were introduced as people from our image project. We expected the presence of two workers in the task would elicit a relative judgment on the workers from participants. Together, participants and their assigned partners provided tags for a series of 5 images. For each displayed image, the two simulated workers would generate 5 tags each (see Figure 1). The participant's role was to select the 3 most descriptive tags from the worker generated lists and submit them as the final tag list.

After completing the labeling phase, participants evaluated the workers three ways: 1) by rating them on a conventional 5 star system; 2) by stating how willing they are to work with each worker in the future (on a 5 point scale); and 3) distributing a bonus of \$1 between the two workers. The evaluation survey

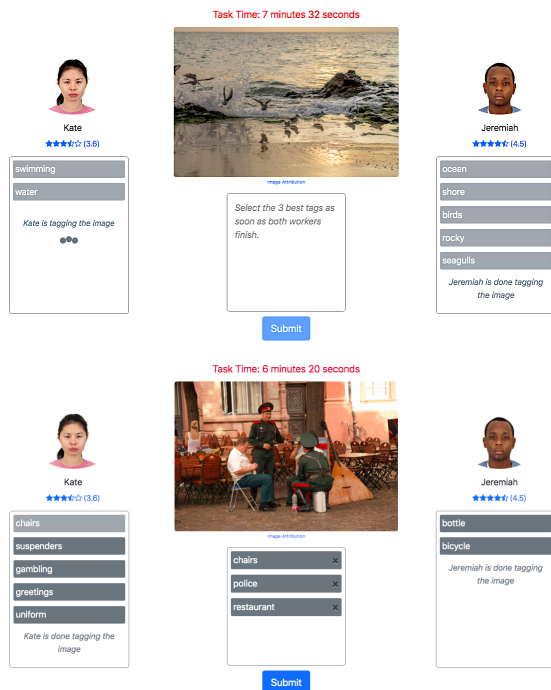


Figure 1: The task interface. For each image, two simulated workers each generate 5 labels which appear sequentially and with different delays. Here, Jeremiah is a high performer and Kate is a low performer. Top: Jeremiah finishes his tags before Kate. The participant cannot select the final tags until both partners are done. Bottom: In another round the participant selects three of Jeremiah's tags to the final list.

also contained two open-ended questions asking how each worker can improve. Following the evaluation, participants completed a demographics survey.

The two workers assigned to a participant varied along some or all dimensions of performance, gender, and race. We differentiated between high and low performing workers using the speed of their response and the quality of their generated tags.

Worker generated tags for images would arrive sequentially and with different delays. The delay for each tag was drawn from a uniform distribution of 1.5-6 and 5-9 seconds for a high- and a low-performer worker respectively. We chose these delay distributions because they felt natural and yet their difference was noticeable. The delay length impacted the efficiency of the whole team because participants could not start constructing the final tag list until both of their worker partners had submitted all of their tags. In addition, to make the delays more perceptible and to also increase a sense of time urgency, we displayed a timer counting up until the tags for the last image were submitted. To ensure that participants would stay focused on the task and notice the difference in delays, we limited the HIT assignment time on MTurk to 15 minutes. As a further check, we measured and logged the time they switched browser tabs while their partners were generating tags. Worker generated tag lists served as another signal of performance. Low performing workers had a minimum of

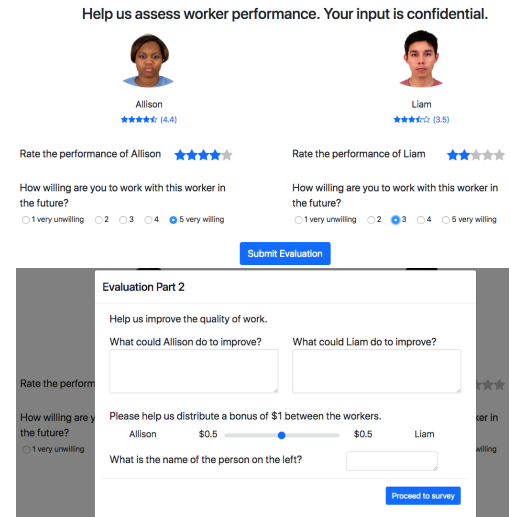


Figure 2: A two phase evaluation interface to rate worker performance, willingness to work with in the future, areas for improvement, and bonus distribution between the two workers.

1 and a maximum of 3 low quality tags mixed in their list, whereas high performing workers' tags were all high quality. We decided it would be unnatural for paired workers to always suggest mutually exclusive lists. Therefore, from the pool of tags for each image, we classified a subset as "obvious". A maximum of 2 tags were drawn from this pool for each image and were submitted by both workers.

Although we distributed a bonus to all participants, task instructions stated that the submitted tags by participants should be of high quality and the task performed in an efficient manner for participants to receive a bonus. With this objective, the quality of worker generated tags and their delays would impact participants' utility. We expected this impact to cause participants to more realistically rate their partners.

Worker Profiles

The experimental factors of our study were gender and performance (2 genders \times 2 performance levels). We created 16 simulated worker profiles that represented the 4 experimental factors and gave each profile a picture and a name. We chose worker profile pictures from the Chicago Face Database [48]. The dataset contains measures along various face characteristics, such as attractiveness and how prominently each face is perceived as belonging to various races. For generalizability of our results and to reduce potential effect of face characteristics other than gender on study participants, we counterbalanced the other dimensions of the profiles. For attractiveness and happiness, we filtered the dataset down to those profiles whose ratings of these measures fell between the 2nd and the 4th quantiles. Another profile dimension was age. Because it has been reported that online laborers are mostly young [58, 40], to choose representative profiles, we further narrowed our pool of faces to those whose age estimate was below the median.

The other dimension of a profile was race which, unlike the other dimensions, was categorical. To counterbalance this

dimension, we chose four races—Asian, Black, Hispanic, and White—for each gender and performance profile. We selected these four races because these were the races present in the database and they are the top four distinct races/ethnicities that are most represented in the US population [1]. For counterbalancing the dimension of race, it was important to choose examples that would appear to participants as belonging to the intended races. Therefore, for each combination of gender and race from our filtered dataset, we selected two faces which were most prominently rated as belonging to their race.

We then randomly assigned each face of a gender and race combination to a performance condition (high or low). We also colored in the uniform grey shirts in the profile pictures to make them look more natural. To choose worker names, we consulted New York City’s dataset of popular baby names that is categorized by babies’ gender and mothers’ ethnicity [2].

Participants

We recruited our study participants from Amazon Mechanical Turk and limited the HIT to US based workers. We compensated them with a base rate of \$1.50 and a bonus of \$1 which we granted to anyone who submitted acceptable answers. With the HIT assignment time set to 15 minutes, the minimum average hourly wage was \$10. A total of 1588 turkers participated in our study (50% female). The medians for age, highest education level achieved, and household annual income were 25-34, Bachelor’s degree, and \$30,500-\$59,999 respectively. 64% of participants were white, 15% Hispanic, 9% Black, and 7% were Asian. With respect to marital status, 42% identified as single, 42% were married, 13% cohabitating, and 4% were divorced. Each user could participate in the study only once.

Image and Tag Selection

Initially, we selected a set of 10 images published under the Creative Commons license on Flickr. To collect tags for these images, we recruited 7 people from our institution to separately generate at least 10 labels per image. Since we were also interested in obtaining low quality tags, we limited the tagging time per image so that the time constraint would pressure people into completing their lists by adding lower quality tags. To determine the quality of produced tags for each image, we first counted the number of people who had cited each tag in their lists, merging tags we considered similar (e.g., children and kids). After counting repeated tags for each image, we selected a set of 10 tags that were mentioned by more than one user and labeled them as high quality. From this set, we classified 4 tags that were cited the most often as “obvious” tags. Obvious tags usually referred to a salient feature of the image and could be suggested by both simulated workers when performing the task. The candidates for low quality tags were those that were cited only once and that the research team considered to have a far but plausible association with the image (e.g., umbrella for an image of a beach where no umbrella was present). Finally, we narrowed our images down to a set of 5 that had the highest user agreement on tags.

The Multiverse

To study bias in ratings in the absence of social signals, we assigned participants to a “*No Ratings*” universe where worker

ratings provided by other participants were not displayed. For each task trial, we assigned a randomly selected pair of workers to the participant. The side each worker was placed on (left, or right) was also determined randomly. A total of 549 participants completed this task and provided 1098 ratings.

Next, to examine how ratings evolve when social signals from others are available, we created 8 universes where participants could view worker ratings both in the course of their collaboration on the task and when providing evaluations (Figures 1 and 2). Each universe contained a replica of the same simulated profiles dwelling in isolation from other universes. Therefore, ratings of a profile in one universe were not impacted by ratings given to its counterparts in other universes. These independent universes would provide a naturalistic setting for evolution of ratings and also allow us to account for random noise caused by the context of a particular universe. The concept of universes in this study is similar to the experimental “worlds” in [60], where the popularity of songs as a quality signal was shown in some worlds but withheld in others.

As participants arrived in the platform, they were randomly assigned to a universe and a pair of simulated workers in that universe who were not engaged in performing the task at that moment. Upon assignment, the two workers would then be unavailable until they were rated by the participant or we determined that the participant had abandoned the task. This worker availability constraint would inhibit concurrent evaluations and help us track the sequential progression of ratings. In addition, since we expected the choice of a workers’ partner would impact their ratings, when assigning partners, we iterated through all combinations of workers until all simulated profiles had been paired with each other approximately once.

We set initial ratings of each of the 16 profiles as the average of the ratings they obtained in the *No Ratings* universe. After each evaluation, we updated the ratings of the evaluated workers by incorporating the new ratings into their averages. The initial rating for each profile was given a weight of 10 in the average. We chose this weight because it was small enough to allow for change in the average given the limited number of subsequent ratings and yet it was large enough for the average ratings to have some stability. Worker ratings in the interface were shown with the precision of one decimal place.

In 4 of the 8 universes where ratings were shown, we flipped the performance of workers at the initialization of the universes so that workers’ current performance would not match their rating history. These “*Reversed Ratings*” universes would help us gain better insight into how raters’ exposure to workers’ prior ratings affects their rating decisions. We refer to the other 4 universes where performance of workers did not differ from before as “*True Ratings*”. Across these 8 universes with ratings, we collected 2370 profile ratings (each participant evaluated two workers) of which we included responses from 955 participants in our analyses. We discarded those ratings from the trials where two workers had been paired with each other before in the same universe. In addition, for updating ratings of profiles, our system automatically flagged and ignored ratings from those users who failed attention checkers in the interface. Our analyses also excludes these responses. After

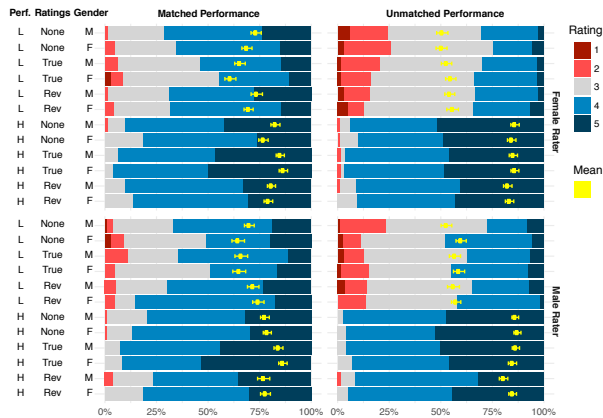


Figure 3: Distribution of workers ratings grouped by comparative performance of paired workers, rating condition, worker performance, worker gender, and rater gender.

the data filtering, the lowest and highest number of participants assigned to a universe were 118 and 120 respectively.

RESULTS

In the sections that follow, we first discuss results related to the star ratings workers received across all conditions, summarized in Figure 3 with the mean worker ratings for each condition in Table 1. We then discuss results related to our secondary outcome measures, willingness to work together and bonus.

When reporting the results of pairwise comparisons, in addition to p values, we have included Cohen’s d as a measure of effect size and 95% confidence intervals. When discussing differences, we refer to effect sizes. Per Cohen’s rules of thumb for interpreting d, we refer to those around 0.2 as small, those in the range of 0.5 as medium, and over 0.8 as large, while noting that these are not firm thresholds [14, 39, 46]. Because the distribution of p values across replications of the same study can be very wide, reliance on a p-value of 0.05 as a binary cut-off can be misleading [18, 17, 65, 39]. Instead, effect sizes convey the magnitude of the effect and confidence intervals provide a continuum that better conveys the probability of the true difference between the two populations being 0. This approach is especially useful for the comparisons across our small stratified subgroups where because of a relatively small sample size, our power to reject the null hypothesis is not high.

As expected, a worker’s performance level had a significant impact on the rating they received. Examining just the *No Ratings* condition where prior ratings were not presented, the results showed that low performing workers were rated significantly lower than high performing workers with a large effect size [$\mu_{\text{low}} = 3.44$, $\mu_{\text{high}} = 4.30$, $t(1003.4) = 17.67$, $CI = [-0.96, -0.77]$, $p < 0.01$, $d = 1.07$]. These results validate the core experimental paradigm in that low and high performing simulated workers were perceived as significantly different from one another in the appropriate direction.

Effects of Matched vs Unmatched Performance (RQ1)

We hypothesized that ratings a worker receives may be impacted by their performance relative to that of their paired co-

worker. We examined the difference in ratings across all conditions between workers in *matched performance* trials where both workers have the same performance level, and workers in *unmatched performance* trials with one low-performing and one high-performing worker.

In unmatched performance trials across all conditions, the stark contrast between the high and low performing workers brought out more differentiation in the ratings compared to matched performance trials. Low quality workers received a lower mean rating of 3.20 [$\sigma = 0.85$] when they were paired with a high quality worker in unmatched trials, compared to being paired with another low quality worker in matched trials [$\mu = 3.74$, $\sigma = 0.83$]. The difference was statistically significant and the effect size was medium [$t(1488.3) = 12.38$, $p < 0.01$, $CI = [-0.62, -0.45]$, $d = 0.64$]. The reverse was true for high quality workers. Their mean rating was higher when they were paired with a low quality worker in unmatched trials [$\mu = 4.39$, $\sigma = 0.62$] compared to being paired with a high quality worker in matched performance trials [$\mu = 4.24$, $\sigma = 0.69$]. The difference was statistically significant and the effect size was small [$t(1449.7) = 4.47$, $CI = [0.08, 0.22]$, $p < 0.01$, $d = 0.23$].

SUMMARY: This finding suggests that participants made relative judgments about the paired workers, exacerbated by unmatched performance: a low performer seems even worse when in direct contrast to a high performer, and vice versa.

Effects of Prior Ratings on Future Ratings (RQ2)

Given the differences observed between matched and unmatched performance trials, we perform the remaining analyses on each of these data partitions separately. To understand how exposure to ratings from other raters affects the rating decisions of participants, we performed several Welch’s t-tests between ratings of workers with the same performance across *No Ratings*, *True Ratings*, and *Reversed Ratings* conditions.

As illustrated in Table 1, our results suggest that participants are taking cues from prior ratings, making judgments that are more extreme in the presence of these social cues than without them. In matched performance trials, low performing workers in the *True Ratings* condition were rated statistically significantly lower compared to when their rating history was absent in the *No Ratings* condition. Similarly, high performers received significantly higher ratings in the presence of their prior *True Ratings* compared to those in the *No Ratings* condition. In the *Reversed Ratings* condition, we observed similar evidence of social influence, though in the opposite directions. High performing workers were rated significantly lower than when their *True Ratings* were shown. Conversely, low performing workers were rated significantly higher than when their *True Ratings* were displayed. These results indicate that raters nudge their scores towards the displayed rating, even when the worker’s performance does not reflect this rating.

In unmatched trials, the absence or presence of prior ratings did not impact future rating, as shown in Table 1, with the exception of limited evidence in one case. This finding suggests the performance differential in unmatched trials is strong enough to overcome the social influence of prior ratings.

Table 1: Welch’s t-tests between ratings of workers in matched and unmatched performance trials across conditions. Bonferroni-adjusted significance threshold is $p \simeq 0.02$. The tests for matched trials show that low performing workers were rated significantly lower and high performers significantly higher when prior ratings were shown compared to when they were withheld. In the reversed ratings condition high performers did not receive ratings as high (or low performers as low) as their counterparts in the true ratings condition. In contrast, for unmatched trials, within each performance stratum there is no statistically significant difference in ratings, regardless of rating condition. The effect size suggests very limited evidence of this bias in unmatched trials.

	Perf.	Ratings Cond.	Means	t test, CIs, & Cohen’s d values
Matched	Low	None vs True	$\mu_N = 3.76, \mu_T = 3.56$	$t(469.6) = 2.59, CI = [0.05, 0.35], p < 0.01, d = 0.24$
		True vs Reversed	$\mu_T = 3.56, \mu_R = 3.88$	$t(440.3) = 4.13, CI = [-0.47, -0.17], p < 0.01, d = 0.39$
	High	None vs True	$\mu_N = 4.16, \mu_T = 4.42$	$t(488.8) = 4.43, CI = [-0.38, -0.15], p < 0.01, d = 0.40$
		True vs Reversed	$\mu_T = 4.42, \mu_R = 4.15$	$t(439.1) = 4.30, CI = [0.15, 0.39], p < 0.01, d = 0.41$
Unmatched	Low	None vs True	$\mu_N = 3.14, \mu_T = 3.24$	$t(535.3) = 1.37, CI = [-0.25, 0.04], p = 0.17, d = 0.12$
		True vs Reversed	$\mu_T = 3.24, \mu_R = 3.23$	$t(507.0) = 0.09, CI = [-0.14, 0.15], p = 0.92, d = 0.01$
	High	None vs True	$\mu_N = 4.43, \mu_T = 4.42$	$t(530.7) = 0.30, CI = [-0.09, 0.12], p = 0.77, d = 0.03$
		True vs Reversed	$\mu_T = 4.42, \mu_R = 4.31$	$t(503.8) = 2.03, CI = [0.004, 0.22], p = 0.04, d = 0.18$

SUMMARY: This finding suggests that displaying ratings of prior work, even when misaligned with a worker’s current performance, affects their future ratings. This influence however, mostly surfaces in matched performance trials, i.e., in absence of a stark difference between performances of paired workers.

Effects of Workers’ Gender on Ratings (RQ3)

To test whether the worker’s gender has an effect on ratings, we developed a linear mixed effects model with ratings as the dependent variable and performance and worker gender as independent variables. We considered the universes as a random factor to account for variations in ratings caused by the context of a universe rather than our experimental factors. We used the function “lmer” from the R package “lme4” to define the model. We then fit the model to the data from matched and unmatched performance trials separately.

A Wald Chi-Square test on the fitted model to the matched performance trials indicated that both performance and gender have a significant effect on ratings [$\chi^2(1) = 97.65, p < 0.01$; $\chi^2(1) = 5.66, p = 0.02$]. To analyze how the gender effect varies by performance, we partitioned the data into high- and low performing segments and fit the same model to each partition, removing performance as a variable. A post-hoc Least Square Means (LSmeans) test on the model for low performers revealed that low performing females were on average rated 0.13 points lower than their male counterparts with a small effect size [$\mu_f = 3.67, \mu_m = 3.80, SE = 0.06, p = 0.03, CI = [-0.26, -0.01], d = 0.16$]. For high performing workers, there was negligible difference in means across genders [$\mu_f = 4.23, \mu_m = 4.24, SE = 0.05, CI = [-0.11, 0.09], p = 0.82, d = 0.01$].

To understand whether the gender effect for low performers varied by condition (*No, True, and Reversed Ratings*), we examined the ratings of low performing female and male workers across all conditions. For matched trials, in the *No Ratings* condition, low performing females were rated lower than their male counterparts with a small effect size [$\mu_f = 3.67, \mu_m = 3.85, t(256.8) = 1.79, CI = [-0.39, 0.02], p = 0.07, d = 0.22$]. In the *True Ratings* condition, the effect size of

gender was still small but with a lower magnitude [$\mu_f = 3.49, \mu_m = 3.64, t(219.8) = 1.29, CI = [-0.37, 0.08], p = 0.2, d = 0.17$]. In the *Reversed Ratings* condition, the difference between mean ratings for women and men was negligible [$\mu_f = 3.85, \mu_m = 3.91, t(218.6) = 0.60, CI = [-0.27, 0.14], p = 0.55, d = 0.08$]. This result suggests that displaying ratings may play some mitigating role in counterbalancing the gender bias. For unmatched performance trials, across all conditions, performance had a significant effect on ratings [$\chi^2(1) = 475.44, p < 0.01$] and gender’s impact on ratings was not significant [$\chi^2(1) = 3.13, p = 0.08$].

SUMMARY: These analyses suggest that gender of workers does in fact impact the ratings they receive. This gender influence emerges in matched performance trials when workers cannot be differentiated performance-wise, only targets low performing workers, and works in favor of male workers and against females. The effect size of gender in ratings of low performers is the highest in *No Ratings* condition when ratings from other raters are hidden, and becomes diluted when true ratings are displayed. The difference is negligible in the *Reversed Ratings* condition in which low performers, the target of this gender bias, are displayed with high prior ratings.

Effects of Raters’ Gender (RQ4)

It is conceivable that the gender of the rater affected their ratings of workers [12]. Figure 3 shows the distribution of ratings across matched and unmatched trials grouped by condition, workers’ performance and gender, and the gender of raters.

Differential Impact of Female Participants

Since we observed that among matched low performers, males fared better than females, we examined how low performers of different genders were rated differently by female and male raters by inspecting matched trials across all conditions. We observed that female raters gave lower ratings to low performing female workers than they did to low performing males with worker gender having a small effect size [$\mu_f = 3.66, \mu_m = 3.82, t(367.2) = 2.00, CI = [-0.33, -0.003], p = 0.05, d = 0.21$]. However, the effect size of worker gender in

the ratings that low performing female and male workers received from male raters was negligible [$\mu_f = 3.70$, $\mu_m = 3.78$, $t(305.4) = 0.93$, $CI = [-0.28, 0.1]$, $p = 0.36$, $d = 0.10$].

Differential Impact of Male Participants

In unmatched performance trials, we observed that within each performance stratum, means across groups were consistent (see Figure 3). The only exception to this consistency of ratings was that males seemed to give higher ratings to low performing females compared to low performing males across all conditions. We observed a small effect size of worker gender in the ratings of low performing females and males who were rated by males in unmatched performance trials [$\mu_f = 3.35$, $\mu_m = 3.20$, $t(385.3) = 1.72$, $CI = [-0.02, 0.31]$, $p = 0.09$, $d = 0.17$].

SUMMARY: These findings imply that gender of raters plays a role in the ratings that workers receive. This effect interacts with the gender of workers, emerges in ratings of low performing workers only, and is dependent on the relative performance between paired workers. It appears that matched performance elicits gender biases from female raters and unmatched performance evokes biases in male raters. In matched trials, female raters evaluate low performing males more favorably than they do low performing females. In contrast, in unmatched performance cases, male raters give higher ratings to low performing females compared to their low performing male peers.

Willingness to Work Together and Bonus (RQ3, RQ4)

To determine how workers' performance and gender affected their bonus and participants' willingness to work with the workers, we built two mixed effects models to explain these outcome variables. These models were similar to the one for ratings and were fitted to the data from the matched and unmatched performance trials separately.

A Wald test revealed that as expected, workers' performance had a significant effect on participants' willingness to work with them in the future, in both unmatched and matched performance trials [$\chi^2(1) = 397.62$, $p < 0.01$ for unmatched; $\chi^2(1) = 75.08$, $p < 0.01$ for matched]. The model fitted to unmatched performance trials suggested that the gender of workers had a significant effect on participants' willingness to work with them [$\chi^2(1) = 4.91$, $p = 0.03$]. To examine how the gender effect varies by performance, we partitioned the data by performance and fit the same model to each partition, removing performance as a variable. The post-hoc LSmeans test on the model for low performers did not yield a significant p-value and the effect size was minimal [$\mu_f = 3.35$, $\mu_m = 3.21$, $SE = 0.08$, $CI = [-0.01, 0.29]$, $p = 0.07$, $d = 0.13$]. For high performers, the effect size of gender was negligible [$\mu_f = 4.59$, $\mu_m = 4.56$, $SE = 0.05$, $CI = [-0.06, 0.12]$, $p = 0.48$, $d = 0.05$]. In matched trials, gender of workers did not affect participants' willingness to work with them [$\chi^2(1) = 0.44$, $p = 0.51$].

A similar test on the model with bonus as the dependent variable revealed that as expected, performance had a significant effect on workers' bonus in unmatched trials [$\chi^2(1) = 597.98$, $p < 0.01$], with the high performing workers receiving significantly more bonus [$\mu_{\text{high}} = \$0.65$, $\mu_{\text{low}} = \$0.35$]. The model

also suggested that gender had a significant effect on bonus in the unmatched trials [$\chi^2(1) = 4.42$, $p = 0.04$]. To analyze how the gender effect varies by performance, similar to our approach for the other measures, we partitioned the data by performance and fit the same model to each partition, removing performance as a variable. We then performed a post-hoc LSmeans test on the same model for low performers and calculated the effect sizes. We observed that women were given more bonus compared to their male peers with a small effect size [$\mu_f = 0.36$, $\mu_m = 0.33$, $SE = 0.01$, $CI = [0.002, 0.05]$, $p = 0.04$, $d = 0.15$]. The effect size of gender in apportioned bonus across genders was trivial among high performers [$\mu_f = 0.66$, $\mu_m = 0.65$, $SE = 0.01$, $CI = [-0.01, 0.03]$, $p = 0.38$, $d = 0.06$]. As expected, there was no effect of performance on bonus in matched trials [$\chi^2(1) = 0.24$, $p = 0.62$] as both workers had similar performance and the bonus was typically split evenly. Gender also did not affect bonus in matched trials [$\chi^2(1) = 0.00$, $p = 1.0$].

We performed additional tests to understand if the rater's gender affected the difference in apportioned bonus between low performing female and male workers in unmatched trials. Consistent with what we observed for ratings, male raters gave more bonus to low performing females than to low performing males with a small effect size [$\mu_f = 0.37$, $\mu_m = 0.33$, $t(375.2) = 2.26$, $CI = [0.005, 0.08]$, $p = 0.02$, $d = 0.23$]. The effect size of gender in the bonus that low performing workers of either gender received from female raters was trivial [$\mu_f = 0.35$, $\mu_m = 0.33$, $t(380.2) = 0.71$, $CI = [-0.02, 0.04]$, $p = 0.48$, $d = 0.07$].

SUMMARY: These findings suggest that low performing female workers in unmatched performance trials receive a higher bonus than low performing males. However, there does not seem to be a notable difference in willingness to work with workers of either gender. Women receive the bonus advantage from male raters. These results on bonus allocations combined with those on ratings demonstrate a consistent trend.

DISCUSSION

The results of our experiments contribute empirical understanding of the complex and nuanced circumstances in which biases manifest. In addition, our work points to several important implications for designing rating mechanisms in digital labor markets that can mitigate these biases.

Gender Bias Is Context-Dependent

In our experiments, we observed that the impact of gender is dependent on the comparative performance of paired workers. This gender bias at times works against female workers and at others, works in their favor. In addition, this bias targets low performers only. This finding is important because lower performance workers are often most vulnerable to platform decisions. For instance, Uber and Lyft deactivate drivers whose average rating falls below a certain threshold. With a gender bias in play, these layoffs can impact low performing drivers of different demographics disproportionately.

In the Absence of Performance Difference

In trials where the performances of paired workers were equal (matched performance), low performing females received

lower ratings compared to their male counterparts ($d = 0.16$). This gender bias first manifested in the *No Ratings* condition where we did not reveal prior rating history of workers ($d = 0.22$). The initial ratings in the universes where aggregate ratings were displayed were calculated using not only ratings from the matched trials in the *No Ratings* universe but also those from the unmatched trials which did not incorporate this gender bias. Therefore, in universes with displayed ratings, there was no significant difference in the starting ratings of low performing females and males [$t(545.0) = 0.28$, $p = 0.78$, $\Delta\mu = 0.02$]. Nevertheless, in *True Ratings* condition, we observed that low performing females were rated lower than males ($d = 0.17$), similar to the *No Ratings* condition. While an absolute difference of 0.15 between ratings of males and females may seem negligible, this was enough to be noticeable in our interface, which rounded ratings to the nearest decimal.

While low performers were impacted by gender bias, there was no significant difference in ratings of high performers across genders in matched trials. This result is surprising and disagrees with some prior work on gender bias in traditional settings reporting that competent males are rated higher than equally competent females, while incompetent males are rated lower than equally incompetent females [53, 19]. However, Nieva and Gutek caution against generalizing these results to situations that do not conform to the “hypothetical person” paradigm where the evaluators do not have extensive contact with the focal individual [53]. Our study involved a collaborative setting where evaluators were invested in the outcome of the task and impacted by the performance of the workers. Therefore, this difference in context, as Nieva and Gutek point out, may have contributed to the difference in results.

It is possible our lack of ratings difference among high performers is due to a ceiling effect. It is also possible that bias is more nuanced and does not manifest against high performing women when their competency is perceived. This issue is highly contested with prior research supporting both explanations. In a study on evaluation of paintings, Pheterson et al. found a similar lack of bias against high performing women. By controlling for the gender of painters and status of paintings, they found that accomplished women were evaluated as favorably as men while women who were perceived as less accomplished were judged less favorably [55]. In contrast, other studies report competent women are in fact discriminated against. For example, in a study comparing cooperative and competitive teams, men “only like a competent woman from a distance; when men and women have to work together or against one another, they prefer an incompetent woman as much as a competent one. When deciding which member should be omitted from the group, both men and women showed a tendency to reject the competent woman relative to the competent man [32].”

We found no difference in ratings across genders when low performers started with artificially high ratings ($d = 0.08$, *Reversed Ratings* condition). This finding suggests that one mechanism for countering the bias that targets low performing females is to equalize all workers by setting their initial ratings to a high value. Furthermore, platforms can test whether the

ratings of their workers are biased by artificially inflating their scores and observing if they fall back to their previous value.

Surprisingly, we found that the source of the difference in ratings of low performing workers across genders was in fact female raters rating low performing males higher ($d = 0.21$). In contrast, male raters did not evaluate low performers of either gender differently ($d = 0.10$). These results are in alignment with Goldberg’s study in which college women evaluated essays attributed to males more positively than those attributed to females [28]. This in-group bias by women has also been studied more recently in the context of hiring student candidates and evaluating teachers [51, 23]. Although these studies reported the bias exists to this date, others have failed to corroborate their findings [47, 13]. Our results suggest that the anti-woman bias by women exists in ratings of online workers.

Although ratings in matched performance trials were subject to subtle contextual biases against women, we observed no effect of gender on participants’ willingness to work with workers in the future and the bonuses the participants distributed to them in these trials. One explanation for these results is that our participants may have interpreted the questions as explicitly impacting the livelihood of the workers and our decision to retain the workers for future tasks. This assumption and the empathy for their fellow workers, may have prompted our participants to provide these evaluations with more discretion. Examples of mutual help and support behaviors among crowd workers have been reported in prior work [30, 67].

When Performance Difference is Noticeable

In unmatched performance trials, we observed a difference in ratings across genders when low performing workers were rated by male raters. Across the three conditions, low performing women who were paired with a high performer received higher ratings ($d = 0.17$) and were given more bonus ($d = 0.23$) compared to their low performing male peers. This differential evaluation by male raters caused low performing females to have an overall higher bonus in unmatched trials compared to low performing males.

We believe this discriminatory rating of low performing females by male raters may not be an in-group bias by men towards men, but rather male raters’ benevolence towards women. The reason is that the effect size of gender in ratings that low performing male workers received from female or male raters in unmatched trials is almost negligible [$\mu_{f \text{ raters}} = 3.10$, $\mu_{m \text{ raters}} = 3.20$, $t(388.3) = 1.14$, $CI = [-0.27, 0.07]$, $p = 0.25$, $d = 0.12$]. A possible explanation is that the clear performance difference between these low performing females and their matched partners may have given rise to benevolent sexism in men towards female workers who by simply being female evoke compassion and a desire to protect in men who endorse such attitudes [26].

Mimicry of Prior Ratings

In the matched performance trials in the *True Ratings* condition, high performers received a higher rating when their rating history was displayed compared to when it was hidden ($d = 0.40$). Similarly, low performing workers received lower ratings in the presence of their rating history ($d = 0.24$).

Therefore, it appears that when prior ratings served as a confirmation of a worker's high or low performance, it gave raters more confidence to push their ratings farther towards one end or the other. The impacts of displaying prior ratings has been studied in the context of online product evaluation [50, 52]. Our results indicate that mimicry of previous ratings by future raters also generalizes to the ratings of online workers.

Although it appears that people were influenced by prior ratings even if they did not align with the performance of workers (*Reversed Ratings* condition), displaying these reversed ratings in fact did not result in high performing workers receiving different ratings from their counterparts in the *No Ratings* condition [$t(475.3) = 0.08$, $CI = [-0.12, 0.13]$, $p = 0.94$, $d = 0.01$]. Displaying reversed ratings for low performers however, increased their scores with a small effect size [$t(482.7) = 1.61$, $CI = [-0.26, 0.03]$, $p = 0.11$, $d = 0.15$]. The *Reversed Ratings* condition therefore, seems to have acted as a minor correction and can be used by platforms as an intervention to combat the inflation of scores. However, it is unclear whether it is ethical to randomly lower the ratings of high achievers to see if mimicry has contributed to their high scores.

This bias surfaced consistently in matched trials, although we observed a limited evidence of this bias in one case in unmatched trials (Table 1). Therefore, it appears that people mimic prior ratings in the absence of clear differentiating signals. To prevent these biases from polluting ratings of workers, we recommend that when soliciting worker ratings, platforms do not ask for an overall rating such as how much a user liked a worker. Instead, they can guide the user through a list of well-defined criteria, asking how the worker has performed along each criterion to help them make a distinction that is based on the worker's qualities. Indeed, multi-item rating instruments and questions focusing on specific behaviors have been shown to increase rating accuracy [15, 42].

Other Implications

While some of our observed effect sizes may appear small, we believe that they are nonetheless consequential because they could compound over time. For instance, we observed a herding behavior for ratings that may push ratings of the population subject to biases farther apart from those of their peers in the long run. Testing this hypothesis however, requires collecting larger-scale data and is left to future work. Furthermore, even a small difference in ratings of digital workers can affect how they are ranked in the search results of a labor marketplace. Bias related to presentation order [16] can potentially exacerbate their differences by causing higher ranked workers to be recruited more and earn more as a result.

It is possible that presenting workers side by side in our study may have evoked an explicit comparison of workers and abetted biased ratings. Nevertheless, we believe our results can also generalize to existing online labor platforms where users do not necessarily interact with more than one worker at a time. Ratings of workers in online services are becoming more ubiquitous and people use several online labor services either at once or close together in time (e.g., hiring Uber drivers to and from a destination). Because an element of comparison may

be present when rating workers in such scenarios, these cases resemble the trials of our study.

In addition, the implications of our study could apply to platforms where digital crowd workers collaborate with each other in real-time [57, 45, 44, 39], and may be rated by peers rather than by employers. Such platforms have been growing in recent years and it is imaginable that they will evolve to add social components. By knowing about potential biases that can arise in such scenarios, designers of these platforms can make more informed decisions on how to counteract the biases.

LIMITATIONS AND FUTURE WORK

One direction for future work is to include the race of workers as a factor affecting ratings and study its interaction with gender and performance. The complexity of our results, with interactions between performance, pairings, and demographics of the workers and of the raters inhibits our ability to study the effect of race on ratings. Additionally, it is possible that the race of the worker interacts with the race of rater, similar to the results we observed for gender. The demographics of our raters being primarily white limits our ability to detect interaction effects. Future work can examine the impact of race by collecting ratings from a more diverse population. One potential limitation of our study is that in the pairwise comparison of workers, other confounding factors may have affected ratings for each worker, such as the gender of who the worker was paired with in each trial. Examining the effects of such factors requires collecting more data and is left to future studies. Another direction for future work is to test how the biases we uncovered transfer to sequential ratings. While we presented workers side by side, future work could have participants interact with and rate one worker at a time. Finally, for the completion of certain gig tasks, work requesters may employ a group of workers to collaborate on the task [57]. While it is conceivable that the relative judgment would also be applicable in that scenario, future work could study how our results transfer to these settings where multiple workers are involved in the task and some may have differing roles.

CONCLUSION

We studied the effects of gender and prior ratings and their interaction with performance on ratings of online platform workers. We found that a gender bias targets low performing workers and its nature varies by the comparative performance of paired workers. For instance, among paired workers with similar performance, low performing male workers were preferred over their female counterparts ($d = 0.16 - 0.22$). We also found that when the comparative performance of platform workers was similar, historical performance ratings exacerbated performance evaluation, skewing low performer ratings even lower ($d = 0.24$), and high performer ratings higher ($d = 0.40$). These findings suggest mechanisms for mitigating the effects of such biases, including introducing short term artificial adjustments to rating histories.

ACKNOWLEDGMENTS

We would like to thank Prof. David Karger for his insightful comments on the work and Ezra Karger and Prof. Elchanan Mossel for their feedback regarding the statistical analyses.

REFERENCES

- [1] 2010. Overview of Race and Hispanic Origin: 2010. (2010). <https://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf>
- [2] 2018. Popular Baby Names | NYC Open Data. (2018). <https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf>
- [3] Bruno Abrahao, Karen Cook, and others. 2016. Power Imbalance and Rating Systems. In *Tenth International AAAI Conference on Web and Social Media*.
- [4] Ali Alkhatib, Michael S Bernstein, and Margaret Levi. 2017. Examining crowd work and gig work through the historical lens of piecework. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 4599–4616.
- [5] Joseph G Altonji and Rebecca M Blank. 1999. Race and gender in the labor market. *Handbook of labor economics* 3 (1999), 3143–3259.
- [6] Sinan Aral. 2014. The problem with online ratings. *MIT Sloan Management Review* 55, 2 (2014), 47.
- [7] Ian Ayres and Peter Siegelman. 1995. Race and gender discrimination in bargaining for a new car. *The American Economic Review* (1995), 304–321.
- [8] Rajiv D Banker and Iny Hwang. 2008. Importance of Measures of Past Performance: Empirical Evidence on Quality of e-Service Providers. *Contemporary Accounting Research* 25, 2 (2008), 307–337.
- [9] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.
- [10] Emilio J Castilla. 2008. Gender, race, and meritocracy in organizational careers. *Amer. J. Sociology* 113, 6 (2008), 1479–1526.
- [11] Jason Chan and Jing Wang. 2014. Hiring biases in online labor markets: The case of gender stereotyping. *Proceedings of the International Conference on Information Systems (ICIS)* (2014).
- [12] Liang Chen. 2018. Exploring Gender Effects on Peer Rating in Open Innovation and Crowdsourcing: A Case of Website Evaluation. (2018).
- [13] Mary Ellen Cline, David S Holmes, and Jana C Werner. 1977. Evaluations of the work of men and women as a function of the sex of the judge and type of work. *Journal of Applied Social Psychology* 7, 1 (1977), 89–93.
- [14] Jacob Cohen. 1988. The effect size index: d. *Statistical power analysis for the behavioral sciences* 2 (1988), 284–288.
- [15] Jean M Converse, Converse Jean McDonnell, and Stanley Presser. 1986. *Survey questions: Handcrafting the standardized questionnaire*. Vol. 63. Sage.
- [16] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 87–94.
- [17] Geoff Cumming. 2008. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science* 3, 4 (2008), 286–300.
- [18] Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.
- [19] Kay Deaux and Janet Taynor. 1973. Evaluation of male and female ability: Bias works two ways. *Psychological reports* 32, 1 (1973), 261–262.
- [20] Tawanna R Dillahunt and Amelia R Malone. 2015. The promise of the sharing economy among disadvantaged communities. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2285–2294.
- [21] Alice H Eagly, Mona G Makhijani, and Bruce G Klonsky. 1992. Gender and the evaluation of leaders: A meta-analysis. *Psychological bulletin* 111, 1 (1992), 3.
- [22] Benjamin G Edelman and Michael Luca. 2014. Digital discrimination: The case of Airbnb. com. *Harvard Business School NOM Unit Working Paper* 14-054 (2014).
- [23] Naomi Ellemers, Henriette Van den Heuvel, Dick De Gilder, Anne Maass, and Alessandra Bonvini. 2004. The underrepresentation of women in science: Differential commitment or the queen bee syndrome? *British Journal of Social Psychology* 43, 3 (2004), 315–338.
- [24] Michael Fay and Lesley Williams. 1993. Gender bias and the availability of business loans. *Journal of Business Venturing* 8, 4 (1993), 363–376.
- [25] Apostolos Filippas, John J Horton, and Joseph Golden. 2017. Reputation in the long-run. (2017).
- [26] Peter Glick and Susan T Fiske. 1997. Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of women quarterly* 21, 1 (1997), 119–135.
- [27] Mareike Glöss, Moira McGregor, and Barry Brown. 2016. Designing for labour: uber and the on-demand mobile workforce. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 1632–1643.
- [28] Philip Goldberg. 1968. Are women prejudiced against women? *Trans-action* 5, 5 (1968), 28–30.
- [29] Laura Gonzalez and Yuliya Komarova Loureiro. 2014. When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. *Journal of Behavioral and Experimental Finance* 2 (2014), 44–58.

- [30] Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. The crowd is a collaborative network. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. ACM, 134–147.
- [31] Mihajlo Grbovic. 2017. Search ranking and personalization at Airbnb. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 339–340.
- [32] Randi L Hagen and Arnold Kahn. 1975. Discrimination Against Competent Women 1. *Journal of Applied Social Psychology* 5, 4 (1975), 362–376.
- [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. In *CSCW*. 1914–1933.
- [34] WILLIAMT HEIR. 2000. Minimizing workplace gender and racial bias. *Psychology* 37 (2000), 233–56.
- [35] Yili Hong and Paul A Pavlou. 2012. *Are global online labor markets truly “flat”? Global frictions and global labor arbitrage*. Technical Report. SSRN Working Paper.
- [36] Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 611–620.
- [37] Nicolas Jones, Armelle Brun, and Anne Boyer. 2011. Comparisons instead of ratings: Towards more stable preferences. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE, 451–456.
- [38] Robert L Kaufman. 2010. *Race, gender, and the labor market: Inequalities at work*. Lynne Rienner Publishers Boulder, CO.
- [39] Rex B Kline. 2013. *Beyond significance testing: Statistics reform in the behavioral sciences*. American Psychological Association.
- [40] Farshad Kooti, Mihajlo Grbovic, Luca Maria Aiello, Nemanja Djuric, Vladan Radosavljevic, and Kristina Lerman. 2017. Analyzing Uber’s ride-sharing economy. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 574–582.
- [41] Tamar Kricheli-Katz and Tali Regev. 2016. How many cents on the dollar? Women and men in product markets. *Science advances* 2, 2 (2016), e1500599.
- [42] Cliff Lampe and R Kelly Garrett. 2007. It’s all news to me: The effect of instruments on ratings provision. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS’07)*. IEEE, 180b–180b.
- [43] Jacqueline Landau. 1995. The relationship of race and gender to managers’ ratings of promotion potential. *Journal of Organizational Behavior* 16, 4 (1995), 391–400.
- [44] Walter S Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P Bigham, and Michael S Bernstein. 2015. Apparition: Crowdsourced user interfaces that come to life as you sketch them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1925–1934.
- [45] Walter S Lasecki, Kyle I Murray, Samuel White, Robert C Miller, and Jeffrey P Bigham. 2011. Real-time crowd control of existing interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 23–32.
- [46] Russell V Lenth. 2001. Some practical guidelines for effective sample size determination. *The American Statistician* 55, 3 (2001), 187–193.
- [47] Hanna Levenson, Brent Burford, Bobbie Bonno, and Loren Davis. 1975. Are women still prejudiced against women? A replication and extension of Goldberg’s study. *The journal of Psychology* 89, 1 (1975), 67–71.
- [48] 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 (2015), 1122–1135.
- [49] Dan Marom, Alicia Robb, and Orly Sade. 2016. Gender dynamics in crowdfunding (Kickstarter): Evidence on entrepreneurs, investors, deals and taste-based discrimination. *Investors, Deals and Taste-Based Discrimination (February 23, 2016)* (2016).
- [50] Wendy W Moe and Michael Trusov. 2011. The value of social dynamics in online product ratings forums. *Journal of Marketing Research* 48, 3 (2011), 444–456.
- [51] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty’s subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* 109, 41 (2012), 16474–16479.
- [52] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.
- [53] Veronica F Nieva and Barbara A Gutek. 1980. Sex effects on evaluation. *Academy of management Review* 5, 2 (1980), 267–276.
- [54] Amanda Pallais. 2014. Inefficient hiring in entry-level labor markets. *American Economic Review* 104, 11 (2014), 3565–99.
- [55] Gail I Pheterson, Sara B Kiesler, and Philip A Goldberg. 1971. Evaluation of the performance of women as a function of their sex, achievement, and personal history. *Journal of Personality and Social Psychology* 19, 1 (1971), 114.
- [56] Devin G Pope and Justin R Sydnor. 2011. What’s in a Picture? Evidence of Discrimination from Prosper. com. *Journal of Human resources* 46, 1 (2011), 53–92.

- [57] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S Bernstein. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 75–85.
- [58] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.
- [59] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and others. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 1621–1630.
- [60] Matthew J Salganik and Duncan J Watts. 2008. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social psychology quarterly* 71, 4 (2008), 338–355.
- [61] Janet Swim, Eugene Borgida, Geoffrey Maruyama, and David G Myers. 1989. Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin* 105, 3 (1989), 409.
- [62] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Raine, Emerson Murphy-Hill, Chris Parnin, and Jon Stallings. 2017. Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science* 3 (2017), e111.
- [63] Jacob Thebault-Spieker, Daniel Kluver, Maximilian A Klein, Aaron Halfaker, Brent Hecht, Loren Terveen, and Joseph A Konstan. 2017a. Simulation Experiments On (The Absence of) Ratings Bias in Reputation Systems. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 101.
- [64] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017b. Toward a geographic understanding of the sharing economy: Systemic biases in UberX and TaskRabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 3 (2017), 21.
- [65] Bruce Thompson. 1987. Five Methodology Errors in Educational Research: The. *psychology* 45 (1987), 721–734.
- [66] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2012. Gender, representation and online participation: A quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*. IEEE, 332–338.
- [67] Ming Yin, Mary L Gray, Siddharth Suri, and Jennifer Wortman Vaughan. 2016. The communication network within the crowd. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1293–1303.