Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr

Anikó Hannák
Northeastern University
ancsaa@ccs.neu.edu

Claudia Wagner
GESIS Leibniz Institute for the Social Sciences & U. of Koblenz-Landau
claudia.wagner@gesis.org

David Garcia
ETH Zürich, Switzerland
dgarcia@ethz.ch

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Markus Strohmaier
GESIS Leibniz Institute for the Social Sciences & U. of Koblenz-Landau
markus.strohmaier@gesis.org

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT
Online freelancing marketplaces have grown quickly in recent years. In theory, these sites offer workers the ability to earn money without the obligations and potential social biases associated with traditional employment frameworks. In this paper, we study whether two prominent online freelance marketplaces—TaskRabbit and Fiverr—are impacted by racial and gender bias. From these two platforms, we collect 13,500 worker profiles and gather information about workers’ gender, race, customer reviews, ratings, and positions in search rankings. In both marketplaces, we find evidence of bias: we find that perceived gender and race are significantly correlated with worker evaluations, which could harm the employment opportunities afforded to the workers. We hope that our study fuels more research on the presence and implications of discrimination in online environments.

ACM Classification Keywords
H.3.5 Online Information Services: Web-based services; J.4 Social and Behavioral Sciences: Sociology; K.4.2 Social Issues: Employment

INTRODUCTION
Online freelance marketplaces such as Upwork, Care.com, and Freelancer have grown quickly in recent years. These sites facilitate additional income for many workers, and even provide a primary income source for a growing minority. In 2014, it was estimated that 25% of the total workforce in the US was involved in some form of freelancing, and this number is predicted to grow to 40% by 2020 [37, 34].

Online freelancing offers two potential benefits to workers. The first, flexibility, stems from workers’ ability to decide when they want to work, and what kinds of tasks they are willing to perform [33]. Indeed, online freelancing websites provide job opportunities to workers who may be disenfranchised by the rigidity of the traditional labor market, e.g., new parents who can only spend a few hours working on their laptops at night, or people with disabilities [66].

The second potential benefit of online freelance marketplaces is the promise of equality. Many studies have uncovered discrimination in traditional labor markets [12, 22, 8], where conscious and unconscious biases can limit the opportunities available to workers from marginalized groups. In contrast, online platforms can act as neutral intermediaries that preclude human biases. For example, when a customer requests a personal assistant from Fancy Hands, they do not select which worker will complete the task; instead, an algorithm routes the task to any available worker. Thus, in these cases, customers’ preexisting biases cannot influence hiring decisions.

While online freelancing marketplaces offer the promise of labor equality, it is unclear whether this goal is being achieved in practice. Many online freelancing platforms (e.g., TaskRabbit, Fiverr, Care.com, TopCoder, etc.) are still designed around a “traditional” workflow, where customers search for workers and browse their personal
profiles before making hiring decisions. Profiles often contain the worker's full name and a headshot, which allows customers to make inferences about the worker's gender and race. Crucially, perceived gender and race may be enough to bias customers, e.g., through explicit stereotyping, or subconscious preconceptions.

Another troubling aspect of existing online freelancing marketplaces concerns social feedback. Many freelancing websites (including the four listed above) allow customers to rate and review workers. This opens the door to negative social influence by making (potentially biased) collective, historical preferences transparent to future customers. Additionally, freelancing sites may use rating and review data to power recommendation and search systems. If this input data is impacted by social biases, the result may be algorithmic systems that reinforce real-world hiring inequalities.

In this study, our goal is to examine bias on online freelancing marketplaces with respect to perceived gender and race. We focus on the perceived demographics of workers since this directly corresponds to the experience of customers when hiring workers, i.e., examining and judging workers based solely on their online profiles. We control for workers' behavior-related information (e.g., how many tasks they have completed) in order to fairly compare workers with similar experience, but varying perceived demographic traits. In particular, we aim to investigate the following questions:

1. How do perceived gender, race, and other demographics influence the social feedback workers receive?

2. Are there differences in the language of the reviews for workers of different perceived genders and races?

3. Do workers’ perceived demographics correlate with their position in search results?

These questions are all relevant, as they directly impact workers’ job opportunities, and thus their ability to earn a livelihood from freelancing sites.

As a first step toward answering these questions, we present case studies on two prominent online freelancing marketplaces: TaskRabbit and Fiverr. We chose these services because they are well established (founded in 2008 and 2009, respectively), and their design is representative of a large class of freelancing services, such as Upwork, Amazon Home Services, Freelancer, TopCoder, Care.com, Honor, and HomeHero. Additionally, TaskRabbit and Fiverr allow us to contrast if and how biases manifest in markets that cater to physical tasks (e.g., home cleaning) and virtual tasks (e.g., logo design) [59].

For this study, we crawled data from TaskRabbit and Fiverr in late 2015, collecting over 13,500 worker profiles. These profiles include the tasks workers are willing to complete, and the ratings and reviews they have received from customers. Since workers on these sites do not self-report gender or race,\(^1\) we infer these variables by having humans label their profile images. Additionally, we also recorded each workers’ rank in search results for a set of different queries. To analyze our dataset, we use standard regression techniques that control for independent variables, such as when a worker joined the marketplace and how many tasks they have completed.

Our analysis reveals that perceived gender and race have significant correlations with the amount and the nature of social feedback workers receive on TaskRabbit and Fiverr. For example, on both services, workers who are perceived to be Black receive worse ratings than similarly qualified workers who are perceived to be White. More problematically, we observe algorithmic bias in search results on TaskRabbit: perceived gender and race have significant negative correlations with search rank, although the impacted group changes depending on which city we examine.

Our findings illustrate that real-world biases can manifest in online labor markets and, on TaskRabbit, impact the visibility of some workers. This may cause negative outcomes for workers, e.g., reduced job opportunities and income. We concur with the recommendations of other researchers [23, 62, 58], that online labor markets should be proactive about identifying and mitigating biases on their platforms.

**Limitations.** It is important to note that our study has several limitations. First, our data on worker demographics is based on the judgement of profile images by human labelers. In other words, we do not know the true gender or race of workers. Fortunately, our methodology closely corresponds to how customers perceive workers in online contexts.

Second, although our study presents evidence that perceived gender and race are correlated with social feedback, our data does not allow us to investigate the causes of these correlations, or the impact of these mechanisms on workers’ hireability. Prior work has shown that status differentiation and placement in rankings do impact human interactions with online systems [49, 18], which suggests that similar effects will occur on online freelance marketplaces, but we lack the data to empirically confirm this.

Third, since we do not know customers’ geolocations, we are unable to control for some location effects. For example, a customer may prefer to only hire workers who live in their own town for the sake of expedience, but if the racial demographics of that town are skewed, this may appear in our models as racial bias.

Lastly, we caution that our results from TaskRabbit and Fiverr may not generalize to other freelancing services. This work is best viewed as a case study of two services at a specific point in time, and we hope that our findings

\(^1\)We refer to this variable as “race” rather than “ethnicity” since it is only based on people’s skin color.
will encourage further inquiry and discussion into labor equality in online marketplaces.

RELATED WORK
In this section, we set the stage for our study by presenting related work. First, we introduce online freelance marketplaces and academic work that has examined them. Second, we briefly overview studies that have uncovered bias in online systems, and the mechanisms that lead to biased outcomes. Finally, we put our work into context within the larger framework of algorithmic auditing.

Online Freelance Marketplaces
In recent years, online, on-demand labor marketplaces have grown in size and importance. These marketplaces are sometimes referred to collectively as the “gig economy” [56], since workers are treated as “freelancers” or “independent contractors”. Whereas in pre-digital times it was challenging for independent workers to effectively advertise their services, and for customers to locate willing workers, today’s online marketplaces greatly simplify the process of matching customers and workers. The fluidity of online, on-demand labor marketplaces gives workers the flexibility to choose what jobs to do and when they are willing to work, while customers have the ability to request tasks that are more complex from very simple (e.g., label an image) to extremely complex (e.g., install new plumbing in a house).

Teodoro et al. propose a classification scheme for on-demand labor marketplaces that divides them along two dimensions: 1) task complexity, ranging from simple to complex, and 2) nature of the tasks, ranging from virtual (i.e., online) to physical (i.e., requiring real-world presence) [59]. For example, Amazon Mechanical Turk is the most prominent example of a microtasking website [66] that falls into the simple/virtual quadrant of the space.

In this study, we focus on two services that fall into the complex half of Teodoro’s classification scheme [59]. TaskRabbit caters to complex/physical jobs such as moving and housework, and is emblematic of similar marketplaces like Care.com and NeighborFavor. In contrast, Fiverr hosts complex/virtual jobs like video production and logo design, and is similar to marketplaces like Freelancer and TopCoder. For ease of exposition, we collectively refer to services in the complex half of Teodoro’s classification as freelancing marketplaces.

Since our goal is to examine racial and gender bias, we focus on freelancing marketplaces in this study. On microtask markets, there is little emphasis on which specific workers are completing tasks, since the price per task is so low (often less than a dollar). In fact, prices are so low that customers often solicit multiple workers for each job, and rely on aggregation to implement quality-control [64, 54, 5]. In contrast, jobs on complex markets are sufficiently complicated and expensive that only a single worker will be chosen to complete the work, and thus facilities that enable customers to evaluate individual workers are critical (e.g., detailed worker profiles with images and work histories). However, the ability for customers to review and inspect workers raises the possibility that preexisting biases may impact the hiring prospects of workers from marginalized groups.

Measuring Freelancing Marketplaces. Given the growing importance of the gig-economy, researchers have begun empirically investigating online freelance marketplaces. Several studies have used qualitative surveys to understand the behavior and motivations of workers on services like Gigwalk [59], TaskRabbit [59, 60], and Uber [39]. Zyskowski et al. specifically examine the benefits and challenges of online freelance work for disabled workers [66]. Other studies present quantitative results from observational studies of workers [47, 14]. This study also relies on observed data; however, to our knowledge, ours is the first study that specifically examines racial and gender inequalities on freelancing marketplaces.

Discrimination
Real-world labor discrimination is an important and difficult problem that has been extensively studied [61]. Some researchers approach the problem from the perception side, by conducting surveys [8] or performing controlled experiments [12, 22]. Others focus on measuring the consequences of labor discrimination by using large, observational data sets to find systematic disparities between groups [1, 2].

Although we are unaware of any studies that examine labor discrimination on online freelance marketplaces, studies have found racial and gender discrimination in other online contexts. For example, Latanya Sweeney found that Google served ads that disparaged African Americans [58], while Datta et al. found that Google did not show ads for high-paying jobs from women [20]. Similarly, two studies have found that female and Black sellers on eBay earn less that male and White sellers, respectively [4, 36]. Edelman et al. used field experiments to reveal that hosts on Airbnb are less likely to rent properties to racial minorities [23]. Finally, Wagner et al. found that biased language was used to describe women in Wikipedia articles [63].

Two studies that are closely related to ours examine discrimination by workers against customers in freelancing markets. Thebault et al. surveyed workers on TaskRabbit from the Chicago metropolitan area, and found that they were less likely to accept requests from customers in the socioeconomically disadvantaged South Side area, as well as from the suburbs [60]. Similarly, Ge et al. found that Uber drivers canceled rides for men with Black-sounding names more than twice as often as for other men [27]. In contrast, our study examines discrimination by customers against workers, rather than by workers against customers.
Mechanisms of Discrimination. Our study is motivated by prior work that posits that the design of websites may exacerbate preexisting social biases. Prior work has found that this may occur through the design of pricing mechanisms [24], selective revelation of user information [45], or the form in which information is disclosed [10, 13, 19, 26].

Many studies in social science have focused on the consequences of status differentiation. High status individuals tend to be more influential and receive more attention [6, 7], fare better in the educational system, and have better prospects in the labor market [46, 53, 42]. Other studies show that men are assumed to be more worthy than women [21, 11, 32, 46, 50] or that Whites are seen as more competent [16, 55]. Status differentiation is thus considered a major source of social inequality that affects virtually all aspects of society [51].

In this study, we examine two freelancing websites that present workers in ranked lists in response to queries from customers. Work from the information retrieval community has shown that the items at the top of search rankings are more likely to be clicked by users [49, 18]. When the ranked items are human workers in a freelancing market, the ranking algorithm can viewed as creating status differentiation. This opens the door for the reinforcement of social biases, if the ranking algorithm itself is afflicted by bias.

Algorithm Auditing
Recently, researchers have begun looking at the potential harms (such as gender and racial discrimination) posed by opaque, algorithmic systems. The burgeoning field of algorithm auditing [52] aims to produce tools and methodologies that enable researchers and regulators to examine black-box systems, and ultimately understand their impact on users. Successful prior audits have looked at personalization on search engines [30, 35], localization of online maps [54], social network news-feeds [25], online price discrimination [31, 43, 44], dynamic pricing in e-commerce [15], and the targeting of online advertisements [29, 38].

Sandvig et al. propose a taxonomy of five methodologies for conducting algorithm audits [52]. In this taxonomy, our study is a “scraping audit”, since we rely on crawled data. Other audit methodologies are either not available to us, or not useful. For example, we cannot perform a “code audit” without privileged access to TaskRabbit and Fiverr’s source code. It is possible for us to perform a “user” or “collaborative audit” (i.e., by enlisting real users to help us collect data), but this methodology offers no benefits (since the data we require from TaskRabbit and Fiverr is public) while incurring significant logistical (and possibly monetary) costs.

BACKGROUND
In this section, we introduce the online freelancing marketplaces TaskRabbit and Fiverr. We discuss the similarities and differences between these markets from the perspective of workers and customers.

TaskRabbit
TaskRabbit, founded in 2008, is an online marketplace that allows customers to outsource small, household tasks such as cleaning and running errands to workers. TaskRabbit focuses on physical tasks [59], and as of December 2015, it was available in 30 US cities.

Worker’s Perspective. To become a “tasker”, a worker must go through three steps. First, they must sign up and construct a personal profile that includes a profile image and demographic information. Second, the worker must pass a criminal background check. Third, the worker must attend an in-person orientation at a TaskRabbit regional center [57].

Once these steps are complete, the worker may begin advertising that they are available to complete tasks. TaskRabbit predefines the task categories that are available (e.g., “cleaning” and “moving”), but workers are free to choose 1) which categories they are willing to perform, 2) when they are willing to perform them, and 3) their expected hourly wage for each category.

Customer’s Perspective. When a customer wants to hire a “tasker”, they must choose a category of interest, give their address, and specify dates and times when they would like the task to be performed. These last two stipulations make sense given the physical nature of the tasks on TaskRabbit. Once the customer has input their constraints, they are presented with a ranked list of workers who are willing to perform the task. The list shows the workers’ profile images, expected wages, and positive reviews from prior tasks.

After a customers has hired a tasker, they may write a free-text review on that worker’s profile and rate them with a “thumbs up” or “thumbs down”. Workers’ profiles list their reviews, the percentage of positive ratings they received, and the history of tasks they have completed.

Fiverr
Fiverr is a global, online freelancing marketplace launched in 2009. On Fiverr, workers advertise “micro-gigs” that they are willing to perform, starting at a cost of $5 per job performed (from which the site derives its name). For the sake of simplicity, we will refer to micro-gigs as tasks².

Unlike TaskRabbit, Fiverr is designed to facilitate virtual tasks [59] that can be conducted entirely online. In December 2015, Fiverr listed more than three million tasks in 11 categories such as design, translation, and online marketing. Example tasks include “a career consultant will create an eye-catching resume design”,

²Since Nov 2015 the site has an open price model though most tasks still cost $5.
“help with HTML, JavaScript, CSS, and JQuery”, and “I will have Harold the Puppet make a birthday video”.

Worker’s Perspective. To post a task on Fiverr, a worker first fills out a user profile including a profile image, the country they are from, the languages they speak, etc. Unlike TaskRabbit, no background check or other preconditions are necessary for a person to begin working on Fiverr. Once a worker’s profile is complete, they can begin advertising tasks to customers. Each task must be placed in one of the predetermined categories/subcategories defined by Fiverr, but these categories are quite broad (e.g., “Advertising” and “Graphics & Design”). Unlike TaskRabbit, workers on Fiverr are free to customize their tasks, including their titles and descriptive texts.

Customer’s Perspective. Customers locate and hire workers on Fiverr using free-text searches within the categories/subcategories defined by Fiverr. After searching, the customer is presented with a ranked list of tasks matching their query. Customers can refine their search using filters, such as narrowing down to specific subcategories, or filtering by worker’s delivery speed.

If a customer clicks on a task, they are presented with a details page, including links to the corresponding worker’s profile page. The worker’s profile page lists other tasks that they offer, customer reviews, and their average rating. Although profile pages on Fiverr do not explicitly list workers’ demographic information, customers may be able to infer this information from a given worker’s name and profile image.

Like TaskRabbit, after a worker has been hired by a customer, the customer may review and rate the worker. Reviews are written as free-text and ratings range from 1 to 5. Similarly, a worker’s reviews and ratings are publicly visible on their profile.

Summary

Similarities. Overall, TaskRabbit and Fiverr have many important similarities. Both markets cater to relatively expensive tasks, ranging from a flat fee of $5 to hundreds of dollars per hour. Both websites also allow workers to fill out detailed profiles about themselves (although only TaskRabbit formally verifies this information). Customers are free to browse workers’ profiles, including the ratings and free-text reviews they have received from previous customers.

Both websites have similar high-level designs and workflows for customers. TaskRabbit and Fiverr are built around categories of tasks, and customers search for workers and tasks, respectively, within these categories. On both sites, search results are presented as ranked lists, and the ranking mechanism is opaque (i.e., by default, workers are not ordered by feedback score, price, or any other simple metric). Once tasks are completed, customers are encouraged to rate and review workers.

Differences. The primary difference between TaskRabbit and Fiverr is that the former focuses on physical tasks, while the latter caters to virtual tasks. Furthermore, TaskRabbit has a strict vetting process for workers, due to the inherent risks of tasks that involve sending workers into customers’ homes. As we will show, this confluence of geographic restrictions and background checks cause TaskRabbit to have a much smaller worker population than Fiverr.

Another important difference between these marketplaces is that workers on Fiverr may hide their gender and race, while workers on TaskRabbit cannot as a matter of practice. On TaskRabbit, we observe that almost all workers have clear headshots on their profiles. However, even without these headshots, customers will still meet hired workers face-to-face in most cases, allowing customers to form impressions about workers’ gender and race. In contrast, since tasks on Fiverr are virtual, workers need not reveal anything about their true physical characteristics. We observe that many workers take advantage of the anonymity offered by Fiverr and do not upload a picture that depicts a person (29%) or do not upload a picture at all (12%).

DATA COLLECTION

We now present our data collection and labeling methodology. Additionally, we give a high-level overview of our dataset, focusing specifically on how the data breaks down along gender and racial lines.

Crawling

To investigate bias and discrimination, we need to collect 1) demographic data about workers on these sites, 2) ratings and reviews of workers, and 3) workers’ rank in search results. To gather this data, we perform extensive crawls of TaskRabbit and Fiverr.

At the time of our crawls, TaskRabbit provided site maps with links to the profiles of all workers in all 30 US cities that were covered by the service. Our crawler gathered all worker profiles, including profile pictures, reviews, and ratings. Thus, our TaskRabbit dataset is complete. Furthermore, we used our crawler to execute search queries across all task categories in the 10 largest cities that TaskRabbit is available in, to collect workers’ ranks in search results.

In contrast, Fiverr is a much larger website, and we could not crawl it completely. Instead, we selected a random subcategory from each of the nine main categories on the site, and collected all tasks within that subcategory. These nine subcategories are: “Databases”, “Animation and 3D”, “Financial Consulting”, “Diet and Weight Loss”, “Web Analytics”, “Banner Advertising”, “Singers and Songwriters”, “T-Shirts”, and “Translation.”
Table 1: Overview of our data sets from TaskRabbit and Fiverr. “Number of Search Results” refers to user profiles that appeared in the search results in response to our queries.

<table>
<thead>
<tr>
<th>Website</th>
<th>Founded</th>
<th># of Workers</th>
<th># of Search Results</th>
<th>Unknown Demographics (%)</th>
<th>Gender (%)</th>
<th>Race (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>taskrabbit.com</td>
<td>2008</td>
<td>3,707</td>
<td>13,420</td>
<td>12%</td>
<td>42%</td>
<td>58%</td>
</tr>
<tr>
<td>fiverr.com</td>
<td>2009</td>
<td>9,788</td>
<td>7,022</td>
<td>56%</td>
<td>37%</td>
<td>63%</td>
</tr>
</tbody>
</table>

The crawler recorded the rank of each task in the search results, then crawled the profile of the worker offering each task.

Overall, we are able to gather 3,707 and 9,788 workers on TaskRabbit and Fiverr, respectively. It is not surprising that TaskRabbit has a smaller worker population, given that the tasks are geographically restricted within 30 cities, and workers must pass a background check. In contrast, tasks on Fiverr are virtual, so the worker population is global, and there are no background check requirements.

We use Selenium to implement our crawlers. We crawled Fiverr in November and December 2015, and TaskRabbit in December 2015. Fiverr took longer to crawl because it is a larger site with more tasks and workers.

**Extracted Features**

Based on the data from our crawls, we are able to extract the following four types of information about workers:

1. **Profile metadata:** We extract general information from workers’ profiles, including: location, languages spoken, a freetext “About” box, and links to Facebook and Google+ profiles. However, not all workers provide all of this information.

2. **Perceived demographics:** Workers on TaskRabbit and Fiverr do not self-identify their gender and race. Instead, we asked workers on Amazon Mechanical Turk to label the gender and race of TaskRabbit and Fiverr workers based on their profile images. Each profile image was labeled by two workers, and in case of disagreement we evaluated the image ourselves. We found disagreement in less than 10% of cases. Additionally, there are a small fraction of images for which race and/or gender cannot be determined (e.g., images containing multiple people, cartoon characters, or objects). This occurred in < 5% of profile images from TaskRabbit, and <18% on Fiverr.

3. **Activity and feedback:** We extract information describing each worker’s career, including the date they joined the site, the tasks they have completed in the past, when they last logged-in to the site, and social feedback in the form of freetext reviews and numeric ratings. Workers on TaskRabbit who have 98% positive reviews and high activity in a 30 day period are marked as “Elite”, which we also record.

4. **Rank:** We record the rank of each worker in response to different search queries. We construct search queries differently on each site, as their search functionality is different. On Fiverr, we search within each subcategory and obtain the ranking of all tasks. On TaskRabbit, we have to provide search parameters, so we select the 10 largest cities, all task types, and dates one week in the future relative to the crawl date. Since we run many queries in different task categories (and geographic locations on TaskRabbit), it is common for workers to appear in multiple result lists.

**Ethics**

While conducting this study, we were careful to collect data in an ethical manner. First, we made sure to respect robots.txt and impose minimal load on TaskRabbit and Fiverr servers during our crawls. Although both sites have Terms of Service that prohibit crawling, we believe that algorithm audits are necessary to ensure civil rights in the digital age. Second, we did not affect the workers on either website since we did not book any tasks or interact with the workers in any way. Third, we minimized our data collection whenever possible; for example, we did not collect workers’ names. Finally, we note that although all information on the two websites is publicly available, we do not plan to release our dataset, since this might violate workers’ contextual expectations about their data.
(a) Negative binomial regression using number of reviews as the dependent variable. Being an Elite worker, active, experienced, and high rating scores have positive effects. The perception of being a woman has significant negative correlation with the number of reviews, particularly so among those also perceived to be White.

<table>
<thead>
<tr>
<th></th>
<th># of Reviews (w/o Interactions)</th>
<th># of Reviews (w/ Interactions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.601***</td>
<td>-2.593***</td>
</tr>
<tr>
<td>Completed Tasks</td>
<td>0.009***</td>
<td>0.009***</td>
</tr>
<tr>
<td>Elite</td>
<td>0.368***</td>
<td>0.371***</td>
</tr>
<tr>
<td>Member Since</td>
<td>-0.308***</td>
<td>-0.308***</td>
</tr>
<tr>
<td>Recent Activity</td>
<td>0.005***</td>
<td>0.005***</td>
</tr>
<tr>
<td>Rating Score</td>
<td>0.049***</td>
<td>0.049***</td>
</tr>
<tr>
<td>Female</td>
<td>-0.087***</td>
<td>-0.105***</td>
</tr>
<tr>
<td>Asian</td>
<td>0.092</td>
<td>-0.145**</td>
</tr>
<tr>
<td>Black</td>
<td>-0.051</td>
<td>0.037</td>
</tr>
<tr>
<td>Asian Women</td>
<td>0.127</td>
<td></td>
</tr>
<tr>
<td>Black Women</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>3,512</td>
<td>3,512</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-11,758</td>
<td>-11,757</td>
</tr>
</tbody>
</table>

(b) Ordinal regression using ratings as the dependent variable shows that being an Elite worker and active have positive effects. Workers perceived to be Black receive significantly fewer reviews than workers perceived to be White. This effect is pronounced among workers perceived to be male.

<table>
<thead>
<tr>
<th></th>
<th>Rating Score (w/o Interactions)</th>
<th>Rating Score (w/ Interactions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed Tasks</td>
<td>0.002*</td>
<td>-0.002*</td>
</tr>
<tr>
<td>Elite</td>
<td>0.585***</td>
<td>0.587***</td>
</tr>
<tr>
<td>Member Since</td>
<td>-0.092*</td>
<td>-0.100*</td>
</tr>
<tr>
<td>Number of Reviews</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>Recent Activity</td>
<td>0.017***</td>
<td>0.017***</td>
</tr>
<tr>
<td>Female</td>
<td>-0.041</td>
<td>-0.08</td>
</tr>
<tr>
<td>Asian</td>
<td>-0.068</td>
<td>-0.149</td>
</tr>
<tr>
<td>Black</td>
<td>-0.306***</td>
<td>-0.347***</td>
</tr>
<tr>
<td>Asian Women</td>
<td>0.206</td>
<td></td>
</tr>
<tr>
<td>Black Women</td>
<td>0.092</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>3,513</td>
<td>3,513</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-5,660</td>
<td>-5,658.14</td>
</tr>
</tbody>
</table>

Table 2: Variables and their relations with reviews and ratings on TaskRabbit. Note: *p < 0.05; **p < 0.01; ***p < 0.001

Labeling Profile Images

Workers do not self-report gender or race on TaskRabbit and Fiverr. Thus, to classify workers’ demographics, we rely on profile image-based inference from workers on Amazon Mechanical Turk (AMT). Each image is evaluated by two AMT workers residing in the US.

We asked AMT workers to answer two questions about each profile image. The first asked the AMT worker if the image depicted a human, multiple humans, or some other non-human image. If the AMT worker determined that the image depicted a single human, then we asked them to classify the race and gender of the person on the image. The AMT workers had to select from pre-defined categories of race and gender. For the racial categories, we picked the three largest groups that are recognized by the United States Census Bureau: White, Black, and Asian [3]. The fourth largest group, Hispanic, is an ethnonym that covers people with diverse racial backgrounds.

Overall, the two raters agreed on 88% of the TaskRabbit images and 85% of the Fiverr images. They had the most difficulty differentiating between White and Asian faces of the same gender; these cases account for over two thirds of all disagreements. In these cases, we manually assessed the picture and either removed or labeled it correctly.

It is important to point out that the true characteristics of workers (i.e., the gender and race they self-identify with), and the characteristics perceived by our human labelers, may not agree. In an online context, customers form their impressions of workers based on the provided profile images, which could potentially differ from reality. In this paper, we use the terms “gender” and “race” to describe these perceived characteristics of workers. Our assumption is that the gender and race labels provided by AMT workers are a close estimate of the impressions that real customers form based on the same images.

Dataset Overview

Table 1 presents an overview of our TaskRabbit and Fiverr datasets, focusing on summary statistics and the gender and racial breakdowns of workers. Our exhaustive crawl collected all workers from TaskRabbit, whereas on Fiverr we only collected workers that had at least one task in our random sample of nine subcategories. Despite this, we see that Fiverr is more popular overall, with our data containing 9,788 workers, versus 3,707 for TaskRabbit.

As shown in Table 1, 12% and 56% of workers on TaskRabbit and Fiverr, respectively, could not be labeled with race and gender. The large fraction of unlabeled workers on Fiverr fall into two categories: 12% have no profile image at all, while 29% have an image that does not contain any human. We include both of these categories into our subsequent analysis since they capture cases where customers cannot perceive workers’ gender or race. Overall, Table 1 shows that Whites and males are the largest identifiable perceived race and gender classes on these websites.

Figure 1 explores the growth of the worker populations on TaskRabbit and Fiverr. The subfigures break down the population by the perceived gender and race of workers. Overall, we observe rapid population growth on both sites, which indicates that online freelancing is becoming an increasingly popular occupation.

Finally, we note that our population data does not include workers who deactivated their accounts prior to our crawls. This raises the question of whether observed imbalances in perceived gender and race are due to 1) unequal numbers of workers joining the sites, 2) certain classes of workers abandoning these sites at greater rates than others, or 3) some combination of the two? In future work, we may
be able to answer this question by using periodic crawls to identify users who deactivate their accounts.

**RESULTS**

We now explore race and gender bias on TaskRabbit and Fiverr. First, we focus on social feedback by analyzing how different variables are correlated with the number of reviews and ratings received by workers. Second, we take a deeper look at the content of customer reviews using linguistic analysis techniques. Both of these investigations reveal significant differences that are correlated with perceived gender and race. This motivates our third analysis, which examines whether perceived gender and race are correlated with workers’ ranking in search results.

**Review and Rating Bias**

To what extent are perceived gender, race, and other demographic variables correlated with the social feedback (in the form of reviews and ratings) workers receive? This is an important question, because social feedback may influence the hiring decisions of future customers. If these social feedback mechanisms are impacted by bias, this may negatively affect the job opportunities available to workers.

To ensure that the effects of perceived gender and race on social feedback are not simply due to other variables correlated with gender/race, we control for a number of factors having to do with 1) demographic information and 2) workers’ experience on the site (e.g., number of completed tasks). Of course, we cannot exclude the possibility that unobserved confounding variables exist, but we do control for all observable cues on the websites in our models.

**Review Bias on TaskRabbit**

Table 2a depicts the results of a negative binomial regression model using the number of reviews as dependent variable and perceived gender and race as independent variables. The first column presents a model without interactions, while the second includes interactions between perceived race and gender. In our models, we use “male” and “White” as the baseline perceived gender and race, i.e., all comparisons are made relative to these categories. For example, the “Female” row in Table 2a compares workers that are perceived to be female versus workers that are perceived to be male in the non-interaction model, and workers that are perceived to be White females versus workers that are perceived to be White males in the interaction model. We control for other factors such as being an elite worker, how long the worker has been a member of TaskRabbit, the last time the worker was online (i.e., activity level), their average rating score, and how many tasks they have completed in the past. The “Member Since” variable of a worker is encoded as the difference in years from 2015 (i.e., 2014 is −1, 2013 is −2, etc.). “Recent Activity” is encoded as the difference in days from the day we collected the data.

First, we examine the model without interactions. Table 2a reveals that all factors besides perceived race have significant statistical relationships with the number of reviews a worker receives. Unsurprisingly, the join date has a significant negative coefficient, which means that workers who joined recently (and therefore have less negative values than those who joined a long time ago) are less likely to have received many reviews. Conversely, recent activity has a significant positive correlation with the number of reviews, since active workers receive more reviews. As we would expect, the number of completed tasks is also positively correlated with the number of reviews. All of these results are intuitive: long-term workers who are very active accrue more reviews than new or infrequent workers.

We also find that the perception of being female is associated with fewer reviews: White women receive 10% fewer reviews than White men (IRR = 0.90). The mean (median) number of reviews for workers perceived to be women is 33 (11), while it is 59 (15) for workers perceived to be men.

Next, we examine the model with interactions. In this model, the gender-coefficient captures the effect of perceived gender for White people, while the race-coefficient captures the effect of perceived race on the number of reviews for men. Table 2a shows that the perception of being female given that a worker is perceived to be White is associated with fewer reviews. Specifically, workers perceived to be White women receive 10% fewer reviews than those perceived to be White men (IRR = 0.90). For all three races we observe that workers perceived to be women receive fewer reviews on average: the mean (median) number of reviews White women receive is 35 (12), while White men get 57 (15) reviews. Black women receive 28 (10) reviews while Black men receive 65 (16) reviews. Asian women receive 32 (10) and Asian men accrue 57 (11) reviews.

We do not observe any significant main correlations for perceived race, but the interaction model shows that workers perceived to be Asian men receive 13% fewer reviews than those perceived to be White men (IRR = 0.87).

Although receiving many reviews may indicate that a worker is hired frequently, we note that reviews are not necessarily positive. In the section “Linguistic Bias” we examine the substantive content of reviews.

**Ratings Bias on TaskRabbit**

Alongside reviews, ratings are another form of social feedback on TaskRabbit. Table 2b shows the results of an ordinal model using ratings as outcome variable on TaskRabbit. As before, we present results from models without and with perceived gender/race interactions. In the no interaction model, we observe that the perception of being Black has a significant statistical relationship with rating scores. However, we see no significant correlation in the case of gender. Furthermore, as shown by the interaction model, workers specifically perceived to be Black men receive worse ratings than Black workers overall.
Furthermore, we see that activity on Fiverr (low average response time, very fast response time) is associated with having a profile image, a lengthy bio, and being a long-time member, which aligns with our overall findings. However, we caution that these images are advertisements containing text about the worker, so we include “no image” and “not human image” in our regression models.

Next, we examine social feedback on Fiverr, starting with reviews. In contrast with TaskRabbit, on Fiverr a significant fraction of users have no profile image or use an image that does not depict a human (many of these images are advertisements containing text about a task). Both of these image choices may impact customers’ perceptions about a worker, so we include “no image” and “not human image” in our regression models. Furthermore, recall that on Fiverr all tasks are virtual, meaning that customers and workers never meet in person (unlike TaskRabbit). This gives workers flexibility to obfuscate their true demographics from customers, which may also impact customer’s perceptions and therefore social feedback.

Table 3a depicts the results of a negative binomial regression using the number of reviews as the dependent variable. Having a lengthy bio, quick response time, being verified on Google+, Facebook and being a long-time member have positive correlations. Having no profile image has a strong negative correlation. Workers perceived to be Black receive worse ratings than those perceived to be male, while workers perceived to be Asian and Black receive worse rating scores than those perceived to be White.

To summarize, we observe that workers perceived to be female on TaskRabbit receive less attention (fewer number of reviews and ratings) than those perceived to be men, and that workers perceived to be Black receive slightly worse ratings than workers perceived to be White. The mean (median) normalized rating score for White workers is 0.98 (1), while it is 0.97 (1) for Black workers.

### Disparities by City on TaskRabbit

Thus far, our analysis of TaskRabbit has focused on our entire dataset, which covers workers from 30 cities. However, given the geographical nature of tasks on TaskRabbit and varying demographic breakdowns across US cities, it is unclear whether our findings are representative of individual cities.

To examine if our findings are consistent across cities, we built separate models per city and repeated each of the above analyses (number of reviews and rating score) on each geographic subset of workers. Unfortunately, most of these models produce no statistically significant results, since the sample sizes are very small (<209 workers). Instead, we present results from four of the largest TaskRabbit cities (New York City, San Francisco, Los Angeles, and Chicago) in Tables 6 and 7 in the Appendix.

We find that the perception of being female is negatively correlated with the number of reviews in every city, which aligns with our overall findings. However, we caution that only two of these correlations are statistically significant (in San Francisco and Chicago). Furthermore, we see that the perception of being Black is associated with worse ratings across all four cities, although this correlation is only significant in New York City. Overall, the correlations that we find on a city-level with respect to perceived gender and race are in agreement with our results from TaskRabbit on the aggregate-level, though with less statistical confidence due to the smaller sample sizes.

### Review Bias on Fiverr

Next, we examine social feedback on Fiverr, starting with reviews. In contrast with TaskRabbit, on Fiverr a significant fraction of users have no profile image or use an image that does not depict a human (many of these images are advertisements containing text about a task). Both of these image choices may impact customers’ perceptions about a worker, so we include “no image” and “not human image” in our regression models.
worker receives. The model also shows a strong negative correlation with not having a profile image. Additionally, we observe a positive correlation when workers show a picture that does not depict a person. As previously mentioned, these images are often advertisements for the worker’s task, so it is plausible that these ads are effective at attracting customers and reviews.

With respect to perceived gender and race, we observe that workers perceived to be Black receive significantly fewer reviews than those perceived to be White (IRR=0.62 which means Black workers receive on average 38% fewer reviews than White workers). The mean (median) number of review for Black workers is 65 (4), while it is 104 (6) for White workers, 101 (8) for Asian workers, 94 (10) for non-human images and 18 (0) for users with no image. This clearly shows that only users with no picture receive fewer reviews than workers perceived to be Black, on average.

Next, we move on to the interaction model, which only includes workers for whom we could label gender and race, i.e., those workers who had human profile pictures. We omit “no image” and “non-human image” workers from the interaction model because we have no way to label their gender or race, so we cannot possibly examine interactions between these variables. Table 3a shows that having a lengthy bio, quick response time, being verified on Google+, and being a long-time member have positive correlations with number of reviews. The interaction model indicates that workers perceived to be Black men receive, on average, 32% fewer reviews than workers perceived to be White men (IRR=0.68).

**Ratings Bias on Fiverr**

Next, we examine ratings on Fiverr. As before, we fit an ordinal regression model to the ratings, using perceived gender and race as independent variables, and control for various other features. We present results similarly to those for TaskRabbit.

Table 3b shows that a lengthy bio, low average response time and having an old account have a positive correlation with the rating scores workers receive. Not having a picture has a strong negative correlation with ratings, but having a non-human image does not significantly correlate with ratings. Additionally, we find that the perception of being female is positively correlated with the rating score. The mean (median) rating score for women is 3.4 (4.8) while it is 3.3 (4.8) for men, 1.7 (0.0) for users with no picture, and 3.6 (4.8) for user with non-human image. We see that in general, users tend to give very positive ratings and only small differences can be observed.

We observe evidence of racial bias in ratings: the perception of being Black or Asian is significantly correlated with worse ratings on Fiverr, compared to workers who are perceived as White. In fact, the mean (median) rating of White workers is 3.3 (4.8), while it is 3.0 (4.6) for Black workers, 3.3 (4.8) for Asian workers, 3.6 (4.8) for workers with a picture that does not depict a person, and 1.7 (0.0) for workers with no image.

When looking at the interaction model in Table 3b, we see significant correlations with perceived gender and race as well. The perception of being a White woman is associated with better rating scores, while workers perceived as male and non-White receive worse ratings.

We were surprised that workers with female profile images received higher ratings than those with male images (as compared to TaskRabbit, where the opposite is true), so we examined our data more closely. It is a commonly argued theory that women need to be exceptionally productive in male-dominated areas in order to succeed, and we see some evidence for this in our data [17, 40]. We observe that across the nine task categories we crawled on Fiverr, workers perceived to be women received dramatically higher ratings than those perceived to be men (on average) in the “Databases” and “Web Analytics” categories. For example, the mean (median) rating for women in the “Databases” category is 3.5 (4.8) while it is 2.8 (4.5) for men. We also observe similar trends in terms of the number of reviews workers perceived to be female receive. In Databases, Web Analytics, and Financial Consulting, women receive more reviews, while in all other categories we see the opposite trend. Furthermore, in these categories the fraction of workers perceived to be women is smaller than the overall average; for example, women are 14% of the population in the “Databases”, versus 37% of the overall population on Fiverr. Motivated by these statistics, we analyze individual task categories on Fiverr in the next section.

**Disparities By Category on Fiverr**

Although tasks on Fiverr are not geographically constrained, they are divided among many diverse categories. This raises the question of whether our findings for Fiverr as-a-whole hold when examining individual categories of tasks.

To answer this question, we built individuals models for all nine task categories that we crawled on Fiverr (with separate models for reviews and ratings). The results for eight categories are shown in Tables 9 and 10 in the Appendix (we omit the ninth category due to space constraints).

Overall, we observe that very few coefficients are significant, thus our per-category analysis is inconclusive. However, it is important to note that by dividing the dataset into nine categories, each is left with few data points, which weakens the statistical power of the categorical analyses.

**Linguistic Bias**

In the previous section, we demonstrate that perceived race and gender have significant correlations with the social feedback received by workers. Next we ask: *Do perceived gender and race correlate with the content of reviews received by workers?*
While coefficients are not significant on TaskRabbit, perceived gender and race have significant effects in Fiverr. Note: for adjectives among negative words. We model the dependence of positive and negative words being adjectives as two logistic regression models in which the probability of a positive or negative word being an adjective depends on the perceived race and gender of the reviewed worker: 

\[
l(P(a_+)) = a \cdot \delta_F + b_1 \cdot \delta_B + b_2 \cdot \delta_A + c_1 \cdot \delta_F \cdot \delta_B + c_2 \cdot \delta_F \cdot \delta_A \quad (1)
\]

where \(l(P(w)) = \ln(P(w)/(1 − P(w)))\), and \(\delta_F\), \(\delta_B\), and \(\delta_A\) are 1 if and only if the reviewed worker is perceived to be female, Black, or Asian, respectively. This model includes the interaction terms \(c_1\) and \(c_2\), which allow us to test if a combination of perceived race and gender is subject to linguistic biases. Similarly, we fit the model for adjectives among negative words (\(P(a_-)\)). Finally, we repeat the fits using the same controls as in our previous analyses, testing for possible confounds with experience, amount of reviews, average rating, etc.

We analyze all English words in reviews on TaskRabbit and Fiverr for which the gender and race of the reviewed worker could be labeled. After applying these filters, our
analysis includes 53,901 positive words and 5,273 negative words drawn from TaskRabbit, and 319,864 positive and 21,429 negative words from Fiverr.

**Linguistic Bias on TaskRabbit and Fiverr**

We present the results of logistic regression in Table 4, reporting the point estimate of each parameter in the models with and without controls. Note that the parameters of a logistic model are log odds ratios, measuring the ratios of probabilities of positive and negative words being adjectives as a function of the perceived race and gender of the reviewed worker.

Overall, the fit for TaskRabbit shows no clear signs of linguistic biases. However, some of the gender and race-related coefficients of the Fiverr model are significant and do not greatly change by introducing controls.

To interpret the effects better, we computed the effect size on each simple model over the predicted values of the dependent variable for the six combinations of perceived gender and race. Figure 2 shows these estimates. Reviews on TaskRabbit do not show large effects, besides a relatively higher frequency of adjectives being used as positive words for workers perceived to be Asian. On Fiverr, we observe that workers perceived to be Black women are less likely to be described with adjectives as positive words. With respect to the use of adjectives as negative words, the effect is most pronounced as positive and significant for workers perceived to be either Black males or females on Fiverr. Not having a recognizable gender in the image or not having an image at all does not have a large effect, but shows a bit of a negative tendency in the use of abstract words for both positive and negative expression.

**Discussion**

The results in Table 4 indicate the existence of linguistic biases depending on perceived gender and race on Fiverr. These results are robust to interactions between perceived gender and race and to the inclusion of controls related to average rating and experience. The absence of effects on TaskRabbit suggest that there is some fundamental difference between the two communities: we hypothesize that this may be due to the different types of tasks the sites offer (i.e., physical versus virtual tasks). It could be that people are more likely to write harsh reviews about taskers they never met personally. Further, different gender and ethnic ratios may exist in the populations of customers and workers.

**Limitations.** A dataset with gender and race annotations of reviewers (in addition to workers) would enable us to test the role of similarity in linguistic biases, including in- and out-group identity effects to fully test linguistic intergroup bias [41]. It is also important to note that our analysis only studies review texts in English. We have no indication of how our results generalize to non-English communication in Fiverr. Future studies could add important features to the analysis, such as the role of non-native speakers, dialects, and the demographics of the authors of reviews.

**Search Ranking Bias**

Finally we ask: Do workers’ perceived race or gender correlate with their rank in search results on TaskRabbit or Fiverr? The motivation for this question is that customers rely on the website’s search engine to locate suitable workers, the same way people rely on Google to surface relevant links. If the ranking algorithms used by TaskRabbit and Fiverr are influenced by demographic variables, this might cause specific classes of workers to be consistently ranked lower, potentially harming their job prospects. It is important to note that even if demographic variables are not explicitly taken into account by a ranking algorithm, the ranking may still be implicitly influenced if it incorporates variables like reviews and ratings, which we have shown are correlated with perceived demographics.

To answer this question, we ran extensive searches on TaskRabbit and Fiverr and recorded workers’ ranks in the results (refer to the “Crawling” section for more details). This enables us to analyze correlations between workers’ rank in search results and other variables. For the purposes of our discussion, “high” ranks are the desirable positions at the top of search results, while “low” ranks are undesirable positions towards the bottom.

**Search Ranking Bias on TaskRabbit**

Table 5 shows the results of an ordinal regression model using workers’ rank in search results as the dependent variable. As before, we have separate models without and with interaction effects. We observe that the number of completed tasks, the membership length, and recent activity have a positive correlation with rank, i.e., active workers and workers who recently joined tend to rank higher than those who have been active for a longer period or have joined recently. Additionally, we note that workers perceived to be Black rank lower than those perceived to be White, while workers perceived to be Black rank lower than those perceived to be White. The interaction model reveals that being perceived as female has a negative relation with rank for White workers but positive for Black workers. Note: *p < 0.05; **p < 0.01; ***p < 0.001.

<table>
<thead>
<tr>
<th>Search Rank</th>
<th>Search Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w/o Interactions)</td>
<td>(w/ Interactions)</td>
</tr>
<tr>
<td>Avg. Rating</td>
<td>0.003***</td>
</tr>
<tr>
<td>Completed Tasks</td>
<td>0.003***</td>
</tr>
<tr>
<td>Member Since</td>
<td>0.457***</td>
</tr>
<tr>
<td>Recent Activity</td>
<td>0.105***</td>
</tr>
<tr>
<td>Reviews</td>
<td>-0.000</td>
</tr>
<tr>
<td>Female</td>
<td>-0.066</td>
</tr>
<tr>
<td>Asian</td>
<td>0.283***</td>
</tr>
<tr>
<td>Black</td>
<td>-0.076***</td>
</tr>
<tr>
<td>Asian Female</td>
<td>0.364*</td>
</tr>
<tr>
<td>Black Female</td>
<td>1.3***</td>
</tr>
<tr>
<td>Observations</td>
<td>12,663</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-45.947</td>
</tr>
</tbody>
</table>

*Note:*
Workers perceived to be female have lower median rank. With respect to race, we observe that workers perceived to be Black tend to be shown at lower ranks relative to those perceived to be White, while workers perceived to be Asian tend to be shown at significantly higher ranks. Overall, we do not observe a significant correlation with perceived gender.

However, the results in Table 5 become more nuanced once we examine the interactions of perceived race and gender. We observe that the perception of being a White woman or a Black man has a significant negative correlation with rank. Conversely, the perception of being a Black woman has a significant positive correlation with rank. Finally, workers perceived to be Asian tend to rank highly regardless of gender.

Search Ranking by City on TaskRabbit
Although the results in Table 5 are significant, they are somewhat confusing: it is unclear why TaskRabbit’s search algorithm would produce rankings that are biased along these axes. To delve into these results further, we built separate regression models for each TaskRabbit city. Table 8 in the Appendix shows the results for four of the largest TaskRabbit cities where the model produces significant results.

Table 8 reveals that the biased rankings produced by TaskRabbit’s search algorithm vary city-to-city. This suggests that the algorithm may take variables into account that we cannot observe (e.g., the click behavior of users in different cities). It is also possible that the ranking algorithm heavily weights negative feedback, which would explain why we observe workers perceived to be Black men appearing at lower ranks in several cities.

Search Ranking by Category on TaskRabbit
Next, we examine rankings within individual task categories, since task categories could function as confounding factors. Figure 3 plots the search rank distribution based on perceived gender in four different categories on TaskRabbit. Note that zero is the highest rank in this figure, i.e., the result at the top of the search results. Each bar captures the $0^{th}$, $25^{th}$, $50^{th}$, $75^{th}$, and $100^{th}$ percentiles. We observe that workers perceived to be women are more likely to appear at lower ranks across all four categories. The gender gap is biggest in the “Parties” category and smallest in “Moving”, but overall workers perceived to be men have higher $25^{th}$ percentile, median, and $75^{th}$ percentile ranks in all categories.

Search Ranking Bias on Fiverr
Our analysis of search ranking on Fiverr differs from our analysis of TaskRabbit in two ways, due to differences between the two websites. First, search results on Fiverr list tasks rather than workers; although each task is offered by one worker, one worker may offer multiple tasks. Therefore, we define the rank of a worker as the average rank between the two websites. However, we found that no variable had a significant correlation with rank.

Thus, we tried a different method. We created a binary variable for each worker, corresponding to whether the worker appeared in the first $X\%$ of the search results or not. We built a mixed-effects model predicting this variable for varying values of $X$ (5%, 10%, 25% and 50%). Since there is variance in perceived gender and race distributions depending on the task category, we control for task categories in our model. However, again we found that no variable exhibited significant correlation with rank.

Although Fiverr’s website claims to rank workers by ratings by default, it is clear from our results that the actual ranking algorithm is more subtle. Based on manual examination of the Fiverr website, it is clear that the ranking algorithm is deterministic (i.e., repeated searches over short timespans return the same tasks in the same order), however there is no clear rationale behind the ordering. On one hand, this result is unsatisfying; on the other hand, whatever hidden variable Fiverr is using to rank workers does not appear to be correlated with perceived gender or race.

CONCLUDING DISCUSSION
In this work we collected and analyzed data from two online freelance marketplaces and quantified race- and gender-based biases. In this section, we briefly summa-
rize our key findings, and discuss implications of these findings.

Summary of Results
Using controlled regression models, we explored the correlations between perceived gender and race with social feedback on TaskRabbit and Fiverr. The models reveal that social feedback on these sites often has a significant statistical relationship with perceived gender and race. Specifically, on TaskRabbit we find:

- Workers perceived to be women, especially White women, receive 10% fewer reviews than workers perceived to be men with equivalent work experience.
- Workers perceived to be Black, especially men, receive significantly lower feedback scores (i.e., ratings) than other workers with similar attributes.

On Fiverr, we find:

- Workers perceived to be Black, especially men, receive ~32% fewer reviews than other men. They also receive significantly lower rating scores. Only workers with no profile image receive lower ratings than Black workers on average.
- Linguistic analysis shows that reviews for workers perceived to be Black women include significantly fewer positive adjectives, while reviews for Black workers in general use significantly more negative adjectives.
- Workers perceived to be Asian, especially men, receive significantly higher rating scores than other workers.

Overall, these results are remarkable for their consistency. Even though TaskRabbit and Fiverr cater to different types of tasks (physical versus virtual), unfortunately, social feedback is biased against workers perceived to be Black on both platforms.

In addition to examining social feedback, we also analyze gender and racial bias in the search algorithms used by TaskRabbit and Fiverr. We find that TaskRabbit’s algorithm produces results that are significantly correlated with perceived race and gender, although the specific groups that are ranked lower change from city-to-city.

It is unclear, based on our analysis, why TaskRabbit’s search algorithm exhibits bias. We find no evidence that the algorithm was intentionally designed to exhibit this behavior, and we consider this to be unlikely. Instead, a more plausible explanation is that the algorithm is designed to take customer behavior into account (e.g., ratings, reviews, and even clicks on profiles). Unfortunately, as we have shown, customer feedback on TaskRabbit is biased, which may implicitly cause the search algorithm to exhibit bias.

Implications for Designers
Although our findings demonstrate that social feedback on online freelancing marketplaces can be biased, simply getting rid of social feedback is not an option for many marketplace proprietors. Customers have come to rely on reviews as key decision aids when shopping online, especially on systems like Fiverr that are entirely virtual. Given that feedback must be presented to customers, marketplace proprietors should take steps to mitigate inherent biases in the data.

One option for web designers is to more selectively reveal review information [45, 10, 13, 19, 26]. For example, we observe that workers perceived to be women on TaskRabbit and perceived to be Black on Fiverr receive significantly less reviews. To mitigate this, designers could consider only showing the most recent r reviews for each worker, while hiding the rest (along with the total number of reviews per worker). This design levels the playing field for workers, while still giving customers access to timely testimonial feedback.

Interestingly, TaskRabbit offers a feature on their service that sidesteps some of the negative consequences of biased feedback. In addition to the “search” workflow for customers to locate workers, TaskRabbit has a “Quick Assign” feature where customers can simply request that a task be completed within a given timeframe, at a given price, by any available worker. Intuitively, “Quick Assign” is similar to Uber, which automatically matches customers to drivers using an algorithm. This system design removes customers’ ability to hand-pick workers, thus mooting the issue of biased hiring decisions. Of course, this design does not fix all issues (e.g., workers can still potentially discriminate against customers), but it does represent a viable alternative in the design space that mitigates issues that stem from biased social feedback.

Lastly, perhaps the most direct approach online freelance marketplaces could take to mitigate biased feedback is to adjust individual worker’s ratings to compensate for measurable sources of bias. For example, in our dataset we observe that workers perceived to be Black (especially men) receive systematically lower ratings than other groups. This deviation is quantifiable, and Black workers’ ratings could be weighted upwards to compensate. Although such a system would almost certainly be controversial (it could be construed as unfair “reverse discrimination”), it would directly mitigate the effect of societal biases without necessitating changes in customer behavior.

Future Work
Our case study on TaskRabbit and Fiverr leaves open several directions for future work. One open question is whether adverse working conditions for women and minorities cause them to drop-out of the freelancing workforce at greater rates than men. This question could be answered by conducting a longitudinal observational study of worker behavior over time.

Another critical question left open by our work is the precise impact of social feedback on customers’ hiring decisions. One possible way to answer this question is through an in-person experiment. Specifically, study pur-
Participants could be recruited, shown an online freelancing website created by the researchers, and asked to make hiring decisions in response to controlled prompts. The data on the constructed website could be derived from real freelancing websites, thus preserving the diversity of workers, tasks, and social feedback that customers would encounter on real marketplaces.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their extremely helpful comments. This research was supported in part by NSF grants IIS-1408345 and IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES


http://dx.doi.org/10.3386/w19021
http://dx.doi.org/10.3386/w22776


**Appendix**
The tables in this section provide additional analysis of our TaskRabbit and Fiverr datasets. Tables 6–8 examine reviews, ratings, and search rank, respectively, for workers on TaskRabbit in four different US cities. Tables 9 and 10 examine reviews and ratings, respectively, for workers on Fiverr in eight different task categories.
<table>
<thead>
<tr>
<th></th>
<th>NYC w/o Int.</th>
<th>NYC w/ Int.</th>
<th>SF w/o Int.</th>
<th>SF w/ Int.</th>
<th>LA w/o Int.</th>
<th>LA w/ Int.</th>
<th>Chicago w/o Int.</th>
<th>Chicago w/ Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed Tasks</td>
<td>0.01***</td>
<td>0.01***</td>
<td>0.006***</td>
<td>0.006***</td>
<td>0.012**</td>
<td>0.012***</td>
<td>0.016***</td>
<td>0.016***</td>
</tr>
<tr>
<td>Elite</td>
<td>0.372**</td>
<td>0.375**</td>
<td>0.438***</td>
<td>0.436***</td>
<td>0.232</td>
<td>0.222</td>
<td>0.384</td>
<td>0.405</td>
</tr>
<tr>
<td>Member Since</td>
<td>-0.321***</td>
<td>-0.322***</td>
<td>-0.303***</td>
<td>-0.303***</td>
<td>-0.286***</td>
<td>-0.286***</td>
<td>-0.277***</td>
<td>-0.287***</td>
</tr>
<tr>
<td>Recent Activity</td>
<td>0.008*</td>
<td>0.009*</td>
<td>0.003</td>
<td>0.003</td>
<td>0.001</td>
<td>0.001</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>Rating Score</td>
<td>0.051***</td>
<td>0.05***</td>
<td>0.047***</td>
<td>0.047***</td>
<td>0.047***</td>
<td>0.047***</td>
<td>0.055***</td>
<td>0.055***</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td>-0.127**</td>
<td>-0.109</td>
<td>-0.017</td>
<td>-0.049</td>
<td>-0.186</td>
<td>-0.31*</td>
</tr>
<tr>
<td>Asian</td>
<td>0.126</td>
<td>0.004</td>
<td>-0.245**</td>
<td>-0.201</td>
<td>-0.105</td>
<td>-0.043</td>
<td>-0.632**</td>
<td>-1.379***</td>
</tr>
<tr>
<td>Black</td>
<td>0.137*</td>
<td>0.166*</td>
<td>0.01</td>
<td>0.04</td>
<td>0.057</td>
<td>0.042</td>
<td>0.159</td>
<td>0.082</td>
</tr>
<tr>
<td>Asian Female</td>
<td>0.256</td>
<td></td>
<td>-0.1</td>
<td>-0.199</td>
<td></td>
<td>1.189**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black Female</td>
<td>-0.074</td>
<td></td>
<td>-0.065</td>
<td>0.204</td>
<td></td>
<td>0.163</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1194</td>
<td>1194</td>
<td>845</td>
<td>845</td>
<td>582</td>
<td>582</td>
<td>211</td>
<td>211</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-3587.8</td>
<td>-3587</td>
<td>-3375</td>
<td>-3374.8</td>
<td>-1777.1</td>
<td>-1776.6</td>
<td>-609.56</td>
<td>-608.08</td>
</tr>
</tbody>
</table>

Table 6: Negative binomial regression on TaskRabbit using number of reviews as the dependent variable. We show results without and with interactions for four different cities. *Note: *p < 0.05; **p < 0.01; ***p < 0.001

<table>
<thead>
<tr>
<th></th>
<th>NYC w/o Int.</th>
<th>NYC w/ Int.</th>
<th>SF w/o Int.</th>
<th>SF w/ Int.</th>
<th>LA w/o Int.</th>
<th>LA w/ Int.</th>
<th>Chicago w/o Int.</th>
<th>Chicago w/ Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed Tasks</td>
<td>-0.005</td>
<td>-0.005</td>
<td>0</td>
<td>0</td>
<td>-0.006</td>
<td>-0.006</td>
<td>-0.017</td>
<td>-0.017</td>
</tr>
<tr>
<td>Elite</td>
<td>0.683*</td>
<td>0.683*</td>
<td>0.464</td>
<td>0.46</td>
<td>0.64</td>
<td>0.477</td>
<td>0.318</td>
<td>0.32</td>
</tr>
<tr>
<td>Member Since</td>
<td>-0.148</td>
<td>-0.147</td>
<td>0.107</td>
<td>0</td>
<td>-0.134</td>
<td>-0.142</td>
<td>-0.532</td>
<td>-0.536</td>
</tr>
<tr>
<td>Number of Reviews</td>
<td>0.006</td>
<td>0.006</td>
<td>0</td>
<td>0</td>
<td>0.007</td>
<td>0.008</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Recent Activity</td>
<td>0.033***</td>
<td>0.033***</td>
<td>-0.002</td>
<td>-0.002</td>
<td>0.019*</td>
<td>0.019*</td>
<td>0.074***</td>
<td>0.074***</td>
</tr>
<tr>
<td>Female</td>
<td>-0.069</td>
<td>-0.189</td>
<td>-0.004</td>
<td>-0.01</td>
<td>-0.132</td>
<td>-0.163</td>
<td>0.331</td>
<td>0.312</td>
</tr>
<tr>
<td>Asian</td>
<td>-0.211</td>
<td>-0.314</td>
<td>0.111</td>
<td>-0.013</td>
<td>-0.468</td>
<td>-0.631</td>
<td>2.395**</td>
<td>2.719*</td>
</tr>
<tr>
<td>Black</td>
<td>-0.292*</td>
<td>-0.41*</td>
<td>-0.301</td>
<td>-0.0164</td>
<td>-0.07</td>
<td>-0.062</td>
<td>-0.561</td>
<td>-0.621</td>
</tr>
<tr>
<td>Asian Female</td>
<td>0.237</td>
<td></td>
<td>0.371</td>
<td></td>
<td>0.495</td>
<td></td>
<td>0.663</td>
<td></td>
</tr>
<tr>
<td>Black Female</td>
<td>0.284</td>
<td></td>
<td>-0.289</td>
<td></td>
<td>-0.006</td>
<td></td>
<td>0.118</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1194</td>
<td>1194</td>
<td>845</td>
<td>845</td>
<td>611</td>
<td>611</td>
<td>211</td>
<td>211</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-1858.36</td>
<td>-1858.18</td>
<td>-1448.24</td>
<td>-1447.58</td>
<td>-934.73</td>
<td>-934.44</td>
<td>-293.24</td>
<td>-293.12</td>
</tr>
</tbody>
</table>

Table 7: Ordinal regression on TaskRabbit using ratings as the dependent variable. We show results without and with interactions for four different cities. *Note: *p < 0.05; **p < 0.01; ***p < 0.001

<table>
<thead>
<tr>
<th></th>
<th>NYC w/o Int.</th>
<th>NYC w/ Int.</th>
<th>SF w/o Int.</th>
<th>SF w/ Int.</th>
<th>LA w/o Int.</th>
<th>LA w/ Int.</th>
<th>Chicago w/o Int.</th>
<th>Chicago w/ Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Rating</td>
<td>-0.011***</td>
<td>-0.012***</td>
<td>0.003</td>
<td>0.004</td>
<td>0.008***</td>
<td>0.01***</td>
<td>-0.013***</td>
<td>-0.013***</td>
</tr>
<tr>
<td>Completed Tasks</td>
<td>0</td>
<td>0.001</td>
<td>-0.008***</td>
<td>-0.008***</td>
<td>-0.018***</td>
<td>-0.017***</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>Member Since</td>
<td>-0.887***</td>
<td>-0.85***</td>
<td>-0.38***</td>
<td>-0.391***</td>
<td>-0.24***</td>
<td>-0.306***</td>
<td>-0.815***</td>
<td>-0.788***</td>
</tr>
<tr>
<td>Number of Reviews</td>
<td>0.004</td>
<td>-0.005***</td>
<td>0.009***</td>
<td>0.009***</td>
<td>0.017***</td>
<td>0.016***</td>
<td>0.007</td>
<td>0.008</td>
</tr>
<tr>
<td>Recent Activity</td>
<td>0.128</td>
<td>0.127</td>
<td>-0.092**</td>
<td>-0.121***</td>
<td>-0.209**</td>
<td>-0.16***</td>
<td>-0.41***</td>
<td>-0.41***</td>
</tr>
<tr>
<td>Female</td>
<td>-1.462***</td>
<td>-0.595***</td>
<td>0.898***</td>
<td>0.89***</td>
<td>0.023</td>
<td>0.628***</td>
<td>0.521***</td>
<td>0.716***</td>
</tr>
<tr>
<td>Asian</td>
<td>-0.064</td>
<td>-1.639***</td>
<td>0.087</td>
<td>0.148</td>
<td>-0.867***</td>
<td>1.883***</td>
<td>-0.415</td>
<td>-0.38</td>
</tr>
<tr>
<td>Black</td>
<td>-0.777***</td>
<td>-0.001</td>
<td>0.158</td>
<td>0.124</td>
<td>0.83***</td>
<td>1.155***</td>
<td>0.266</td>
<td>0.386*</td>
</tr>
<tr>
<td>Asian Female</td>
<td>1.669*</td>
<td></td>
<td>-0.68</td>
<td></td>
<td>-3.754*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black Female</td>
<td>-0.556*</td>
<td></td>
<td>0.289</td>
<td></td>
<td>-1.465*</td>
<td></td>
<td>-0.416</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>2257</td>
<td>2257</td>
<td>2801</td>
<td>2801</td>
<td>2299</td>
<td>2299</td>
<td>860</td>
<td>860</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-6209.79</td>
<td>-6199.6</td>
<td>-8445.9</td>
<td>-8444.01</td>
<td>-6792.3</td>
<td>-6743.3</td>
<td>-3009.02</td>
<td>-3007.88</td>
</tr>
</tbody>
</table>

Table 8: Ordinal regression on TaskRabbit using search result rank as the dependent variable. We show results without and with interactions for four different cities. *Note: *p < 0.05; **p < 0.01; ***p < 0.001
Table 9: Negative binomial regression on Fiverr using the number of reviews as the dependent variable. We show results with interactions for eight different task categories. Note: *p < 0.05; **p < 0.01; ***p < 0.001

<table>
<thead>
<tr>
<th></th>
<th>Databases w/ Int.</th>
<th>Animation w/ Int.</th>
<th>Financial w/ Int.</th>
<th>Dieting w/ Int.</th>
<th>Web Analytics w/ Int.</th>
<th>Banner Ads w/ Int.</th>
<th>Songwriters w/ Int.</th>
<th>T-shirts w/ Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>About Length</td>
<td>0.013*</td>
<td>-0.007</td>
<td>0.02**</td>
<td>0.003</td>
<td>0.014*</td>
<td>0.001</td>
<td>0.021***</td>
<td>0.021***</td>
</tr>
<tr>
<td>Avg. Response Time</td>
<td>0.001***</td>
<td>0.001***</td>
<td>0.002***</td>
<td>0.000</td>
<td>0.002***</td>
<td>0.01***</td>
<td>0.001</td>
<td>0.002***</td>
</tr>
<tr>
<td>Facebook Profile</td>
<td>-0.015</td>
<td>0.464*</td>
<td>0.689**</td>
<td>0.09</td>
<td>0.118</td>
<td>0.38</td>
<td>0.274</td>
<td>-0.096</td>
</tr>
<tr>
<td>Google+ Profile</td>
<td>0.25</td>
<td>0.303</td>
<td>0.184</td>
<td>-0.072</td>
<td>-0.074</td>
<td>0.087</td>
<td>0.25</td>
<td>0.125</td>
</tr>
<tr>
<td>Member Since</td>
<td>0.866***</td>
<td>0.81***</td>
<td>0.525***</td>
<td>0.720***</td>
<td>0.836***</td>
<td>0.740***</td>
<td>0.898***</td>
<td>1.055***</td>
</tr>
<tr>
<td>Rating Score</td>
<td>1.016***</td>
<td>1.138***</td>
<td>0.885***</td>
<td>1.002***</td>
<td>0.88***</td>
<td>1.018***</td>
<td>0.992***</td>
<td>1.198***</td>
</tr>
<tr>
<td>Spoken Languages</td>
<td>-0.221*</td>
<td>0</td>
<td>-0.116</td>
<td>0.153</td>
<td>-0.107*</td>
<td>-0.314*</td>
<td>0.004</td>
<td>-0.006</td>
</tr>
</tbody>
</table>

Female                      -0.34          0.273             0.428             -0.323          0.083                 0.688*              -0.222             0.583*           
Asian                       -0.193         -0.344             0.082             -0.301          -0.312                0.399               -0.166             0.52*            
Black                       -0.216         0.006             -0.651            -1.323*         0.346                 0.525               -0.142             0.459           
Asian Female                0.411          -0.164            0.142             0.385           0.968*                -0.745              0.089              -0.4            
Black Female                0.106          -0.555            0.081             1.374*          -0.576                -1.08               0.017              -0.291           
Observations                684           323               204               456             324                  378               521                561            
Log Likelihood              -2102.8       -1840.7          -358.38           -1155.9         -1074.8               -1541.5             -1772.4            -1684.4         

Table 10: Ordinal regression on Fiverr using ratings as the dependent variable. We show results with interactions for eight different task categories. Note: *p < 0.05; **p < 0.01; ***p < 0.001

<table>
<thead>
<tr>
<th></th>
<th>Databases w/ Int.</th>
<th>Animation w/ Int.</th>
<th>Financial w/ Int.</th>
<th>Dieting w/ Int.</th>
<th>Web Analytics w/ Int.</th>
<th>Banner Ads w/ Int.</th>
<th>Songwriters w/ Int.</th>
<th>T-shirts w/ Int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.016*</td>
<td>0.008</td>
<td>0.02*</td>
<td>0.005</td>
<td>0.006</td>
<td>0.021***</td>
<td>0.014**</td>
<td>0.013**</td>
</tr>
<tr>
<td>About Length</td>
<td>0.002***</td>
<td>0</td>
<td>0.001***</td>
<td>0.002***</td>
<td>0.001***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
</tr>
<tr>
<td>Avg. Response Time</td>
<td>0.286</td>
<td>-0.227</td>
<td>0.43</td>
<td>0.06</td>
<td>0.023</td>
<td>0.092</td>
<td>-0.141</td>
<td>0.239</td>
</tr>
<tr>
<td>Facebook Profile</td>
<td>0.403</td>
<td>0.225</td>
<td>0.261</td>
<td>0.152</td>
<td>0.959**</td>
<td>0.143</td>
<td>0.718**</td>
<td>0.276</td>
</tr>
<tr>
<td>Google+ Profile</td>
<td>0.284*</td>
<td>-0.058</td>
<td>0.098</td>
<td>0.159</td>
<td>0.49**</td>
<td>0.42***</td>
<td>0.301***</td>
<td>0.57**</td>
</tr>
<tr>
<td>Member Since</td>
<td>0.006***</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0.002**</td>
<td>0</td>
</tr>
<tr>
<td>Spoken Languages</td>
<td>0.179</td>
<td>-0.015</td>
<td>-0.253</td>
<td>0.259</td>
<td>0.11</td>
<td>0.081</td>
<td>0.212</td>
<td>-0.002</td>
</tr>
</tbody>
</table>

Female                      1.108*         0.085             0.283             0.307           0.313                 0.204               -0.147              0.126           
Asian                       0.143          0.343             0.086             0.223           -0.787**              -0.332              -0.377              -0.379          
Black                       -1.273         -0.024            -0.213            -0.216          -1.463*               -0.69               -0.723**             -0.136          
Asian Female                -0.327         -0.409            -0.26             -0.589          0.673                 -0.226              0.084              -0.35           
Black Female                -1.098         -0.929            -0.775            -0.16           0.602                 -1.678*             0.287               0.816           
Observations                374           323               204               241             324                  378               521                561            
Log Likelihood              -608.39       515.04            -345.96           -376.19         -576.06               -680.59             -780.58             -1012.55