Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr

Anikó Hannák[†] **Claudia Wagner**[*] **David Garcia**[‡]
Alan Mislove[†] **Markus Strohmaier**[*] **Christo Wilson**[†]

[†]Northeastern University [*]GESIS Leibniz Institute for the Social Sciences & U. of Koblenz-Landau
[‡]ETH Zürich, Switzerland

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 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

Author Keywords

Gig economy; discrimination; information retrieval; linguistic analysis

INTRODUCTION

Online freelance marketplaces such as Upwork, Care.com, Freelancer, and TopCoder 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 [36, 33].

Online freelancing offers two potential benefits to workers, the first of which is *flexibility*. Flexibility stems from workers' ability to decide when they want to work, and what kinds of tasks they are willing to perform [32]. 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 [65].

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.
CSCW'17, February 25–March 1, 2017, Portland, Oregon.
Copyright © 2014 ACM ISBN/14/04...$15.00.
DOI string from ACM form confirmation

The second potential benefit of online freelance marketplaces is the promise of *equality*. Many studies have uncovered discrimination in traditional labour 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 orders a ride from Uber, or 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, making it easy for customers' biases to come into play. Furthermore, many freelancing websites (including the four listed above) allow customers to rate and review workers. This further opens the door to negative social influence by making (potentially biased) collective, historical preferences transparent to future customers. Finally, 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 and discrimination on online freelancing marketplaces with respect to gender and race. Discrimination refers to the differential treatment of individuals based on their characteristics rather than actual behavior. We therefore rely on demographic characteristics of workers and control for behavior-related information in order to measure discrimination. 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 genders and races?

3. Do workers' 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 for three reasons: *first*, they are well established, having been founded in 2008 and 2009, respectively. This predates other well-known "sharing-economy" services like Airbnb and Uber. *Second*, their design is representative of a large class of freelancing services: on TaskRabbit, customers use free-text searches to browse rank-ordered lists of workers, while on Fiverr customers search and browse a list of tasks (each offered by one worker). Customers can rate and review workers, and browse workers' personal profiles before making hiring decisions. Other services with similar architecture include Upwork, Amazon Home Services, Freelancer, TopCoder, Care.com, Honor, and HomeHero. *Third*, 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) [58].

For this study, we crawled data from TaskRabbit and Fiverr in October and December of 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 labeling their profile images. Additionally, we also recorded each workers' rank in search results for a set of different queries and combinations of search filters.

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. To analyze the language used by customers in written reviews of workers, we use the methods of Otterbacher *et al.* [47] to detect abstract and subjective language. This is implemented as a logistic regression at the word level across review texts.

Our analysis reveals that gender and race have a significant correlation with the amount and the nature of social feedback workers receive. On TaskRabbit, we find that women receive significantly fewer reviews, especially White women. We also find evidence for racial bias: Black workers receive worse ratings than Asian and White workers, especially Black men. Most problematically, we find algorithmic bias in search results: gender and race have significant negative correlations with search rank, although the impacted group changes depending on which city is examined.

On Fiverr, we also find evidence of gender and racial biases in social feedback. In contrast to TaskRabbit, we find that women receive more positive rating scores than men. However, similar to TaskRabbit, Black workers on Fiverr receive fewer reviews than White workers (especially men), while Black and Asian workers receive significantly worse ratings than White workers. Furthermore, we find evidence of linguistic biases in written reviews on Fiverr, where Black women are less likely to be described with positive adjectives, while Asian and Black men are more likely to be described using negative adjectives.

Ultimately, 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 in the form of reduced job opportunities and income. We concur with the recommendations of other researchers [23, 61, 57], that online labor markets should be proactive about identifying and mitigating biases on their platforms.

Limitations. Although our study presents evidence that perceived gender and race are correlated with social feedback and search ranking in online freelancing marketplaces, our data does not allow us to directly 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 [48, 18], which suggests that similar effects will occur on online freelance marketplaces, but we lack the data to empirically confirm this.

Further, 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 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" [55], since workers are treated as "freelancers" or

[1]We refer to this variable as "race" rather than "ethnicity" since it is only based on people's skin color.

"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 give workers the flexibility to choose what jobs to they are willing to do, and when they are willing to work, while customers have the ability to request jobs that range in complexity 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) [58]. For example, Amazon Mechanical Turk is the most prominent example of a microtasking website [65] 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 [58]. 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 [63, 53, 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 freelancing marketplaces. Several studies have used qualitative surveys to understand the behavior and motivations of workers on services like Gigwalk [58], TaskRabbit [58, 59], and Uber [38]. Zyskowski *et al.* specifically examine the benefits and challenges of online freelance work for disabled workers [65]. Other studies present quantitative results from observational studies of workers [46, 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 studied for many years [60]. Some researchers approach the problem from the perception side, by conducting surveys [8] or performing controlled experiments [12, 22]. Other studies 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 [57], 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, 35]. 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 [62].

The study that is most closely related to ours is by Thebault *et al.* [59]. In this work, the authors surveyed workers on TaskRabbit from the Chicago metropolitan area, and found that they were less likely to accept requests from customer in the socioeconomically disadvantaged South Side area, as well as from the suburbs. 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 [44], 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 [45, 52, 41]. Other studies show that men are assumed to be more worthy than women [21, 11, 31, 45, 49] or that Whites are seen as more competent [16, 54]. Status differentiation is thus considered a major source of social inequality that affects virtually all aspects of individuals' lives [50].

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 far more likely to be clicked on by users [48, 18]. When the ranked items are human workers in a freelancing marketplace, 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* [51] 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 [29, 34], localization of online maps [53], social network news-feeds [25], online price discrimination [30, 42, 43], dynamic pricing in e-commerce [15], and the targeting of online advertisements [28, 37].

Sandvig *et al.* propose a taxonomy of five methodologies for conducting algorithm audits [51]. 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 [58], 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 [56].

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 *tasks*[2].

Unlike TaskRabbit, Fiverr is designed to facilitate `virtual` tasks [58] 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", "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

[2]Since Nov 2015 the site has an open price model though most tasks still cost $5.

tasks matching their query.[3] 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 that workers 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 much stricter vetting process for workers, due to the inherent risks of physical 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 learn workers' gender and race when they physically arrive to complete tasks. 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%).

The ability for workers on Fiverr to selectively reveal their demographics impacts how the results of our analysis should be interpreted. On TaskRabbit, we are able to correlate workers' true gender and race with customer ratings and reviews. In contrast, on Fiverr, we are measuring the correlation between workers' *perceived* gender and race (based on their name and profile image), with customer-generated performance metrics.

[3]Note that search results on Fiverr and TaskRabbit are slightly different: on Fiverr, searches return lists of tasks, each of which is offered by a worker; on TaskRabbit, searches return a list of workers.

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.

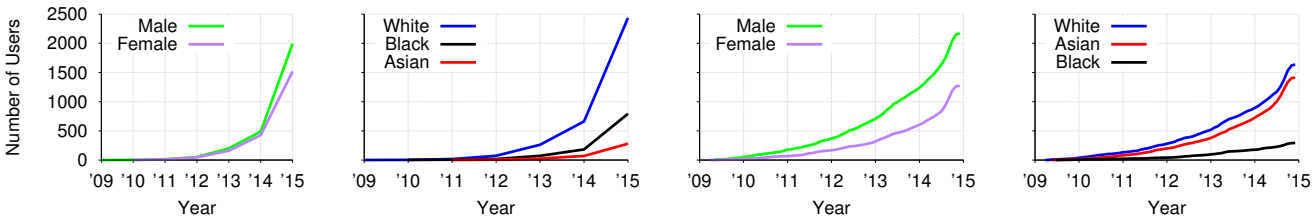
At the time of our crawls, TaskRabbit provided site maps with links to the profiles of all workers in all 30 US cities that were covered by the service. Our crawler gathered all worker profiles, including profile pictures, reviews, and ratings. Thus, **our TaskRabbit dataset is complete**. Furthermore, 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". 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.

Website	Founded	# of Workers	# of Search Results	Unknown Demographics (%)	Gender (%) Female	Male	Race (%) White	Black	Asian
taskrabbit.com	2008	3,707	13,420	12%	42%	58%	73%	15%	12%
fiverr.com	2009	9,788	7,022	56%	37%	63%	49%	9%	42%

Table 1: Overview of the two data sets from TaskRabbit and Fiverr. "Number of Search Results" refers to user profiles that appeared in the search results in response to our search queries. We cannot infer the gender or race for 12% and 56% of users, respectively.



(a) TaskRabbit, gender (b) TaskRabbit, race (c) Fiverr, gender (d) Fiverr, race

Figure 1: Member growth over time on TaskRabbit and Fiverr, broken down by gender and race.

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. *Demographic information:* 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 find 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. We do not consider these users in the remainder of our analysis.

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 site 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.

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.

	# of Reviews (w/o Interactions)	# of Reviews (w/ Interactions)
(Intercept)	$\mathbf{-2.601}^{***}$	$\mathbf{-2.593}^{***}$
Completed Tasks	$\mathbf{0.009}^{***}$	$\mathbf{0.009}^{***}$
Elite	$\mathbf{0.368}^{***}$	$\mathbf{0.371}^{***}$
Member Since	$\mathbf{-0.308}^{***}$	$\mathbf{-0.308}^{***}$
Recent Activity	$\mathbf{0.005}^{***}$	$\mathbf{0.005}^{***}$
Rating Score	$\mathbf{0.049}^{***}$	$\mathbf{0.049}^{***}$
Female	$\mathbf{-0.087}^{***}$	$\mathbf{-0.105}^{***}$
Asian	0.092	$\mathbf{-0.145}^{**}$
Black	-0.051	0.037
Asian Women		0.127
Black Women		0.033
Observations	3,512	3,512
Log Likelihood	−11,758	−11,757

(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. Being a woman has significant negative correlation with the number of reviews, particularly so among White workers.

	Rating Score (w/o Interactions)	Rating Score (w/ Interactions)
Completed Tasks	$\mathbf{0.002}^{*}$	$\mathbf{-0.002}^{*}$
Elite	$\mathbf{0.585}^{***}$	$\mathbf{0.587}^{***}$
Member Since	$\mathbf{-0.092}^{*}$	$\mathbf{-0.100}^{*}$
Number of Reviews	0.002	0.002
Recent Activity	$\mathbf{0.017}^{***}$	$\mathbf{0.017}^{***}$
Female	−0.041	−0.08
Asian	−0.068	−0.149
Black	$\mathbf{-0.306}^{***}$	$\mathbf{-0.347}^{***}$
Asian Women		0.206
Black Women		0.092
Observations	3,513	3,513
Log Likelihood	−5,660	−-5,658.14

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

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

Workers do not self-report gender or race on either website. 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 pick select predefined 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, 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.

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 depict a 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 race and gender on these websites.

Figure 1 explores the growth of the worker populations on TaskRabbit and Fiverr. The subfigures break down the population by gender and race. 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 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 gender and race. This motivates our *third* analysis, which examines whether gender and race are correlated with workers' ranking in search results.

Review and Rating Bias

To what extent are 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 customers' hiring decisions, especially in online scenarios where in-person evaluation is impossible before hiring. 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 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 gender and race as independent variables. The first column presents a model without interactions, while the second includes interactions between race and gender. 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 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 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 women is 33 (11), while it is 59 (15) for men.

Next, we examine the model with interactions. In this model, the gender-coefficient captures the effect of gender for White people, while the race-coefficient captures the effect of race on the number of reviews for men. Table 2a shows that being female given that a worker is White is associated with fewer reviews. Specifically, White women receive 10% fewer reviews than White men ($IRR = 0.90$). For all three races we observe that 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 race, but the interaction model shows that Asian men receive 13% fewer reviews than 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 gender/race interactions. In the no interaction model, we observe that 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, gender bias is most apparent in the case of men: Black men receive worse ratings than White men.

To summarize, we observe that women on TaskRabbit receive less attention (fewer number of reviews and ratings) than men, and that Black workers receive slightly worse ratings than White workers. 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 physical nature of tasks on TaskRabbit and varying demographic breakdowns across US cities, it is unclear whether our findings generalize to 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 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

	# of Reviews (w/o Interactions)	# of Reviews (w/ Interactions)
(Intercept)	$\mathbf{-2.3121}^{***}$	$\mathbf{-2.796}^{***}$
"About" Length	$\mathbf{0.017}^{***}$	$\mathbf{0.002}^{***}$
Avg. Response Time	$\mathbf{0.001}^{***}$	$\mathbf{0.001}^{***}$
Facebook Profile	$\mathbf{0.149}^{**}$	0.029
Google+ Profile	$\mathbf{0.122}^{*}$	$\mathbf{0.319}^{***}$
Member Since	$\mathbf{0.82}^{***}$	$\mathbf{0.843}^{***}$
Rating Score	$\mathbf{0.05}^{***}$	$\mathbf{1.095}^{***}$
Spoken Languages	−0.021	−0.054
No Image	$\mathbf{-0.1260}^{**}$	
Not Human Image	$\mathbf{0.073}^{*}$	
Female	0.062	0.11
Asian	-0.011	-0.015
Black	$\mathbf{-0.481}^{***}$	$\mathbf{-0.382}^{**}$
Asian Female		−0.07
Black Female		−0.2370
Observations	6,275	3342
Log Likelihood	−21,908	−12,146

(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 negative correlation, while having a non-human image is positively correlated with the number of reviews. Black workers receive fewer reviews than White workers, especially so in the case of men.

	Rating Score (w/o Interactions)	Rating Score (w/ Interactions)
"About" Length	$\mathbf{0.013}^{*}$	$\mathbf{0.002}^{***}$
Avg. Response Time	$\mathbf{0.002}^{***}$	$\mathbf{0.002}^{***}$
Facebook Profile	0.042	$\mathbf{0.193}^{*}$
Google+ Profile	$\mathbf{0.355}^{***}$	$\mathbf{0.368}^{***}$
Member Since	$\mathbf{0.36}^{***}$	$\mathbf{0.422}^{***}$
Spoken Languages	$\mathbf{0.69}^{**}$	0.014
No Image	$\mathbf{-0.608}^{***}$	
Not Human Image	−0.079	
Female	$\mathbf{0.175}^{*}$	$\mathbf{0.203}^{*}$
Asian	$\mathbf{-0.222}^{**}$	$\mathbf{-0.377}^{***}$
Black	$\mathbf{-0.45}^{***}$	$\mathbf{-0.367}^{*}$
Asian Female		0.15
Black Female		−0.156
Observations	6,275	3,342
Log Likelihood	−10,931.46	−5,603

(b) Ordinal regression using ratings as the dependent variable. Having a lengthy bio, quick response time, being verified on Google+ or Facebook and being a long-time member have positive effects. Having no profile image has a strong negative correlation. Females receive higher rating scores than males while Asian and Black workers receive worse rating scores than White workers.

Table 3: Analyzing variables that may impact reviews and ratings on Fiverr. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Francisco and Chicago). Furthermore, we see that 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 gender and race are in agreement with our results from TaskRabbit on the aggregate-level, though with less statistical confidence due to the smaller number of users.

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.

Table 3a depicts the results of a negative binomial regression using the number of reviews as the dependent variable and gender and race as independent variables. We control for other individual factors, including average response time to inquiries, number of spoken languages, and membership length on Fiverr. As before, we present results without interactions first.

We observe that activity on Fiverr (low average response time, lengthy profile description, and verified Google+ account) and experience ("Member Since" and ratings) have a positive correlation with the number of reviews a 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 depicting 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 gender and race, we observe that Black workers receive significantly fewer reviews than White workers (IRR=0.62 which means Black workers receive on average 38% fewer reviews than White workers), especially so in the case of men. 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 pictures and 18 (0) for users with no image. This clearly shows that only users with no picture receive fewer reviews than Black workers on average.

Next, we move on to the interaction model, which only includes workers for whom we could identify 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 do not know 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 Black men in particular receive on average 32% fewer reviews than 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 gender as independent variable, and control for various other features. We present results similarly to those above. However, note that workers on Fiverr may hide their true gender and race, since tasks are `virtual`. Thus, all

gender and race effects are due to *perceived* demographics, rather than workers' true attributes.

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 picture. 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 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, 39]. We observe that across the nine task categories we crawled on Fiverr, women received dramatically higher ratings than 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 women 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 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 Task 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 racial and gender biases have significant correlations with the social feedback received by workers. Next we ask: *Do gender and racial biases correlate with the content of reviews received by workers?*

Methods

We measure linguistic biases in reviews using the methods of Otterbacher *et al.* [47] to detect abstract and subjective language. Abstract expression manifests through the use of adjectives, which tend express time-independent properties of what is described [40, 27]. An illustrative comparison are the phrases "is a fantastic web programmer" and "implemented the web site very well": the former is more abstract through the use of an adjective to describe a generalized property, rather than a concrete fact which is usually depicted through the usage of verbs. We detect adjectives in reviews by applying the Parts-Of-Speech tagger of NLTK [9]. We identify subjectivity through the MQPA *subjectivity clues* lexicon [64], composed of more than 8,000 terms classified by polarity. For each word in a review, we match its appearance in the lexicon, and identify if it is positive or negative.

We test for the existence of linguistic biases through a logistic regression at the word-level. 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 race and gender of the reviewed worker:

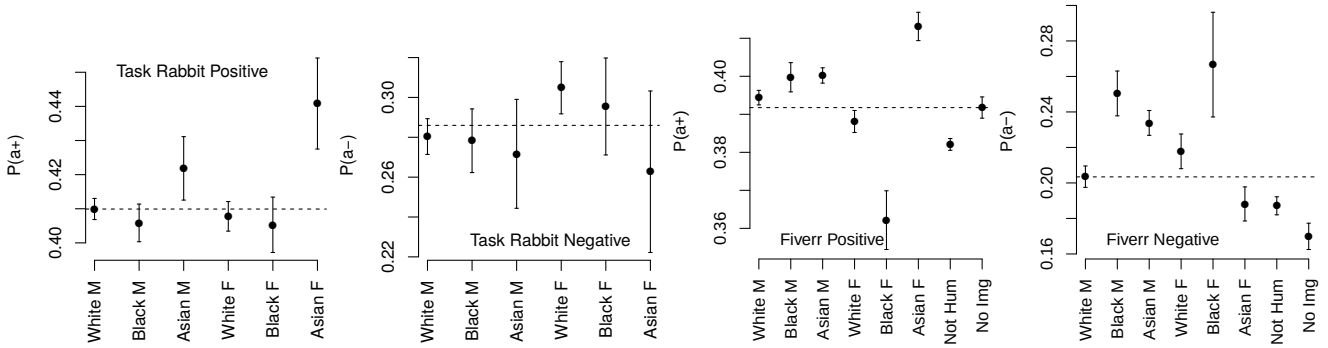
$$l(P(a_+)) = a{\cdot}\delta_F + b_1{\cdot}\delta_B + b_2{\cdot}\delta_A + c_1{\cdot}\delta_F{\cdot}\delta_B + c_2{\cdot}\delta_F{\cdot}\delta_A \quad (1)$$

where $l(P(w)) = ln(P(w)/(1-P(w)))$, and δ_F, δ_B, and δ_A are 1 if and only if the reviewed worker is female, Black, or Asian, respectively (or perceived to be so, in the case of Fiverr). This model includes the interaction terms c_1 and c_2, which allow us to test if a combination of 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 identified. After applying these filters,

	TaskRabbit					Fiverr			
	Positive Adjectives		Negative Adjectives			Positive Adjectives		Negative Adjectives	
	w/ Cntrl	No Cntrl	w/ Cntrl	No Cntrl		w/ Cntrl	No Cntrl	w/ Cntrl	No Cntrl
(Intercept)	**-0.418*****	**-0.364*****	**-0.862*****	**-0.943*****	(Intercept)	0.025	**-0.429*****	**13.154***	**-1.364*****
Female	-0.009	-0.009	0.100	0.118	Female	**-0.037***	-0.026	0.100	0.086
Asian	0.047	0.049	-0.046	-0.043	Asian	0.015	**0.024***	**0.167****	**0.178*****
Black	-0.016	-0.017	-0.008	-0.010	Black	0.006	0.022	**0.283*****	**0.268*****
Asian Female	0.085	0.086	-0.160	-0.164	Asian Female	0.046	**0.08*****	**-0.426*****	**-0.361*****
Black Female	0.008	0.007	-0.048	-0.034	Black Female	**-0.101***	**-0.133*****	-0.041	-0.001
					Not Human	**-0.047*****	**-0.052*****	-0.041	**-0.105***
					No Image	-0.012	-0.011	**0.347*****	**-0.222*****
Last Online	0.000		-0.001		Response	-0.000		**-0.009*****	
Join Date	0.010		-0.003		Member Since	**-0.003*****		**-0.001****	
Elite	-0.040		0.013		About Len.	0.000		0.001***	
Experience	0.000		0.000		Google+	0.019		**0.282*****	
					Facebook	0.008		**-0.231*****	
Log Likelihood	-36475.9	-36477.6	-3152.8	-3154.2		-162352	-214102	-7866	-10786
Num. Obs.	53901	53901	5273	5273		242259	319864	15617	21429

Table 4: Results of logistic regression for TaskRabbit and Fiverr to detect linguistic biases, with and without controls. While coefficients are not significant on TaskRabbit, gender and race have significant effects in Fiverr. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$



(a) TaskRabbit (b) Fiverr

Figure 2: Fitted $P(a_+)$ and $P(a_-)$ depending on combinations of gender and race of the reviewed worker. Points show expected values and bars standard errors. In Fiverr, Black workers are less likely to be described with adjectives for positive words, and Black Male workers are more likely to be described with adjectives for negative words.

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 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 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 Asian workers. On Fiverr, we observe that 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 Black male and female workers in 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 gender and race on Fiverr. These results are robust to interactions between 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 different types of tasks they sites offer (physical

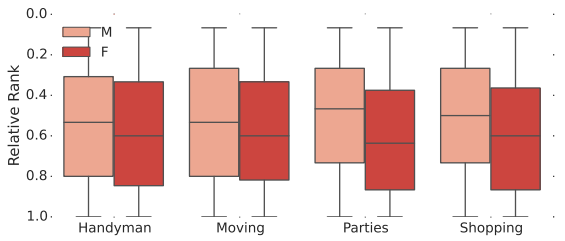


Figure 3: Search rank distributions for four task categories on TaskRabbit by gender. *Note that zero is the highest rank on the page, i.e., the first result.* Females have lower median ranks in all four categories. The gender gap is biggest for "Party Planning" while women are positioned least badly in "Moving".

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 ethnicity ratios may exist in the populations of costumers 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 [40]. 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' 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 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 Rank (w/o Interactions)	Search Rank (w/ Interactions)
Avg. Rating	$\mathbf{0.003}^{***}$	$\mathbf{0.003}^{***}$
Completed Tasks	$\mathbf{0.003}^{***}$	$\mathbf{0.003}^{***}$
Member Since	$\mathbf{0.457}^{***}$	$\mathbf{0.51}^{***}$
Recent Activity	$\mathbf{0.105}^{***}$	$\mathbf{0.089}^{***}$
Reviews	-0.000	-0.004
Female	-0.066	$\mathbf{-0.468}^{***}$
Asian	$\mathbf{0.283}^{***}$	$\mathbf{0.194}^{*}$
Black	$\mathbf{-0.076}^{*}$	$\mathbf{-0.428}^{***}$
Asian Female		$\mathbf{0.364}^{*}$
Black Female		$\mathbf{1.3}^{***}$
Observations	12,663	9,132
Log Likelihood	−45,947	−33,128

Table 5: Ordinal regression using search result rank as the dependent variable for TaskRabbit. The model without interactions reveals that Asian workers rank higher than White workers, while Black workers rank lower than White workers. The interaction model reveals that being a woman has a negative relation with rank for White workers but positive for Black workers. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

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. Additionally, ratings have a weak positive correlation, while reviews have a weak negative correlation with rank, indicating that workers with positive ratings rank higher than workers who simply have large quantities of feedback.

With respect to race, we observe that Black workers tend to be shown at lower ranks relative to White workers, while Asian workers tend be shown at significantly higher ranks. Overall, we do not observe a significant correlations with gender.

However, the results in Table 5 become more nuanced once we examine the interactions of race and gender. We observe that being a White women or a Black man has a significant negative correlation with rank. Conversely, being a Black woman has a significant positive correlation with rank. Finally, Asian workers 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 which we do not know and cannot unobserved (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 Black men appearing at lower ranks in several cities.

Search Ranking by Task 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 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 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 "Shopping", but overall 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 of all tasks he/she offers that match the search. *Second*, Fiverr returns thousands of results for each query, unlike TaskRabbit where results are constrained by location and availability.

Initially, we attempted to build an ordinal regression model using average rank as the dependent variable (much like the model we use to examine TaskRabbit in the previous section). 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 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 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 summarize our key findings, and discuss implications of these findings.

Summary of Results

Using controlled regression models, we explored the correlations between 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 gender and race. Specifically, on TaskRabbit we find:

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

On Fiverr, we find:

- Workers who are 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 Black workers on both platforms. Furthermore, it is important to note that customers do not meet workers in-person on Fiverr, so negative predispositions against Black workers are based purely on customer *perception.*

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 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 [44, 10, 13, 19, 26]. For example, we observe that women on TaskRabbit and Black workers 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 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 Black workers (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 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

1. Aleo, M., and Svirsky, P. Foreclosure fallout: The banking industry's attack on disparate impact race discrimination claims under the fair housing act and the equal credit opportunity act. *BU Pub. Int. LJ 18* (2008).
2. Altonji, J. G., and Blank, R. M. Chapter 48 race and gender in the labor market. In *Handbook of Labor Economics*, vol. 3, Part C. Elsevier, 1999, pp. 3143 – 3259.
3. Arias, E., Heron, M., Tejada-Vera, B., et al. National vital statistics reports. *National Vital Statistics Reports 61*, 9 (2013).
4. Ayres, I., Banaji, M., and Jolls, C. Race effects on eBay. *The RAND Journal of Economics 46*, 4 (2015), 891–917.
5. Bashir, M. A., Arshad, S., Robertson, W., and Wilson, C. Tracing information flows between ad exchanges using retargeted ads. In *In Proceedings of Usenix Security* (2016).
6. Berger, J., Fisek, M. H., Norman, R. Z., and Zelditch Jr, M. *Status characteristics and interaction: An expectation states approach*. New York: Elsevier, 1977.
7. Berger, J., Rosenholtz, S. J., and Zelditch, M. Status organizing processes. *Annual review of sociology* (1980), 479–508.
8. Bertrand, M., and Mullainathan, S. 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.
9. Bird, S., Klein, E., and Loper, E. *Natural Language Processing with Python*, 1st ed. O'Reilly Media, Inc., 2009.
10. Bolton, G., Greiner, B., and Ockenfels, A. Engineering trust: Reciprocity in the production of reputation information. *Management Science 59*, 2 (2013), 265–285.
11. Brashears, M. E. Sex, society, and association: A cross-national examination of status construction theory. *Social Psychology Quarterly 71*, 1 (2008), 7–85.
12. Carlsson, M., and Rooth, D.-O. Evidence of ethnic discrimination in the swedish labor market using experimental data. *Labour Economics 14*, 4 (2007), 716 – 729. European Association of Labour Economists 18th annual conference CERGE-EI, Prague, Czech Republic 21-23 September 2006.

13. Che, Y. K., and Horner, J. Optimal Design for Social Learning. Levine's Bibliography 786969000000001075, UCLA Department of Economics, Sept. 2015.

14. Chen, L., Mislove, A., and Wilson, C. Peeking beneath the hood of uber. In *Proceedings of the 2015 ACM Conference on Internet Measurement* (2015).

15. Chen, L., Mislove, A., and Wilson, C. An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the 25th International World Wide Web Conference* (May 2016).

16. Cohen, E. G. Expectation states and interracial interaction in school settings. *Annual Review of Sociology 8*, 1 (1982), 209–235.

17. Correll, S. J. Gender and the career choice process: The role of biased self-assessments1. *American journal of Sociology 106*, 6 (2001), 1691–1730.

18. Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (2008).

19. Dai, W., Jin, G. Z., Lee, J., and Luca, M. Optimal aggregation of consumer ratings: An application to yelp.com. Working Paper 18567, National Bureau of Economic Research, November 2012.

20. Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. In *In Proceedings of the 15th Privacy Enhancing Technologies Symposium* (2015).

21. David G. Wagner, J. B. Gender and interpersonal task behaviors: Status expectation accounts. *Sociological Perspectives 40*, 1 (1997), 1–32.

22. Dovidio, J. F., and Gaertner, S. L. Aversive racism and selection decisions: 1989 and 1999. *Psychological Science 11*, 4 (2000), 315–319.

23. Edelman, B. G., Luca, M., and Svirsky, D. Racial discrimination in the sharing economy: Evidence from a field experiment, 2015. `http://ssrn.com/abstract=2701902`.

24. Einav, L., Farronato, C., Levin, J. D., and Sundaresan, N. Sales mechanisms in online markets: What happened to internet auctions? Working Paper 19021, National Bureau of Economic Research, May 2013.

25. Eslami, M., Aleyasen, A., Karahalios, K., Hamilton, K., and Sandvig, C. Feedvis: A path for exploring news feed curation algorithms. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* (2015).

26. Fradkin, A., Grewal, E., Holtz, D., and Pearson, M. Bias and reciprocity in online reviews: Evidence from field experiments on airbnb. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* (2015).

27. Gorham, B. W. News media's relationship with stereotyping: The linguistic intergroup bias in response to crime news. *Journal of communication 56*, 2 (2006), 289–308.

28. Guha, S., Cheng, B., and Francis, P. Challenges in measuring online advertising systems. In *Proceedings of the 2010 ACM Conference on Internet Measurement* (2010).

29. Hannak, A., Sapieżyński, P., Kakhki, A. M., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. Measuring Personalization of Web Search. In *Proceedings of the 22nd International World Wide Web Conference* (May 2013).

30. Hannak, A., Soeller, G., Lazer, D., Mislove, A., and Wilson, C. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 ACM Conference on Internet Measurement* (2014).

31. Hopcroft, R. L. Is gender still a status characteristic? *Current Research in Social Psychology 7*, 20 (2002), 339–346.

32. Hoque, F. How the rising gig economy is reshaping businesses. Fast Company, Sept. 2015. `https://www.fastcompany.com/3051315/the-future-of-work/the-gig-economy-is-going-global-heres-why-and-what-it-means`.

33. Intuit. Intuit 2020 report. `http://http-download.intuit.com/http.intuit/CMO/intuit/futureofsmallbusiness/intuit_2020_report.pdf`.

34. Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., and Mislove, A. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 ACM Conference on Internet Measurement* (2015).

35. Kricheli-Katz, T., and Regev, T. How many cents on the dollar? women and men in product markets. *Science Advances 2*, 2 (2016).

36. Laumeister, G. *The Next Big Thing In E-Commerce: Online Labor Marketplaces*. Forbes, 2014. `http://www.forbes.com/sites/groupthink/2014/10/21/the-next-big-thing-in-e-commerce-online-labor-marketplaces/`.

37. Lécuyer, M., Ducoffe, G., Lan, F., Papancea, A., Petsios, T., Spahn, R., Chaintreau, A., and Geambasu, R. Xray: Enhancing the web's transparency with differential correlation. In *Proceedings of the 23rd USENIX Conference on Security Symposium* (2014).

38. Lee, M. K., Kusbit, D., Metsky, E., and Dabbish, L. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015).

39. Lemkau, J. P. Personality and background characteristics of women in male-dominated occupations: A review. *Psychology of Women Quarterly 4*, 2 (1979), 221–240.

40. Maass, A., Salvi, D., Arcuri, L., and Semin, G. R. Language use in intergroup contexts: the linguistic intergroup bias. *Journal of personality and social psychology 57*, 6 (1989), 981.

41. Martha Foschi, Larissa Lai, K. S. Gender and double standards in the assessment of job applicants. *Social Psychology Quarterly 57*, 4 (1994), 326–339.

42. Mikians, J., Gyarmati, L., Erramilli, V., and Laoutaris, N. Detecting price and search discrimination on the internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks* (2012).

43. Mikians, J., Gyarmati, L., Erramilli, V., and Laoutaris, N. Crowd-assisted search for price discrimination in e-commerce: First results. In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies* (2013).

44. Morton, F., Zettelmeyer, F., and Silva-Risso, J. Consumer information and discrimination: Does the internet affect the pricing of new cars to women and minorities? *Quantitative Marketing and Economics 1*, 1 (2003), 65–92.

45. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., and Handelsman, J. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences 109*, 41 (2012), 16474–16479.

46. Musthag, M., and Ganesan, D. Labor dynamics in a mobile micro-task market. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013).

47. Otterbacher, J. Linguistic bias in collaboratively produced biographies: Crowdsourcing social stereotypes? In *In Proceedings of the 9th International AAAI Conference on Web and Social Media* (2015).

48. Richardson, M. Predicting clicks: Estimating the click-through rate for new ads. In *In Proceedings of the 16th International World Wide Web Conference* (2007).

49. Ridgeway, C. L. Framed by gender - how gender inequality persists in the modern world. *Oxford University Press* (2011).

50. Ridgeway, C. L. Why status matters for inequality. *American Sociological Review 79*, 1 (2014), 1–16.

51. Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Proceedings of "Data and Discrimination: Converting Critical Concerns into Productive Inquiry", a preconference at the 64th Annual Meeting of the International Communication Association* (2014).

52. Schmader, T. Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology 38*, 2 (2002), 194–201.

53. Soeller, G., Karahalios, K., Sandvig, C., and Wilson, C. MapWatch: Detecting and Monitoring International Border Personalization on Online Maps. In *Proceedings of the 25th International World Wide Web Conference* (May 2016).

54. Steele, Claude M.; Aronson, J. Stereotype threat and the intellectual test performance of african americans. *Journal of Personality and Social Psychology 69*, 5 (1995), 797–811.

55. Sundararajan, A. The 'gig economy' is coming. what will it mean for work? The Guardian, July 2015. `https://www.theguardian.com/commentisfree/2015/jul/26/will-we-get-by-gig-economy`.

56. Support, T. Are your taskers screened and background checked? TaskRabbit, May 2016. `https://support.taskrabbit.com/hc/en-us/articles/204411630-Are-your-Taskers-screened-and-background-checked-`.

57. Sweeney, L. Discrimination in online ad delivery, 2013. `http://ssrn.com/abstract=2208240`.

58. Teodoro, R., Ozturk, P., Naaman, M., Mason, W., and Lindqvist, J. The motivations and experiences of the on-demand mobile workforce. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and; Social Computing* (2014).

59. Thebault-Spieker, J., Terveen, L. G., and Hecht, B. Avoiding the south side and the suburbs: The geography of mobile crowdsourcing markets. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing* (2015).

60. Todisco, M. Share and share alike? considering racial discrimination in the nascent room-sharing economy. *Stanford Law Review Online 67* (2015), 121–129.

61. Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *CoRR abs/1501.06307* (2015).

62. Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *In Proceedings of the 9th Annual International AAAI Conference on Web and Social Media* (2015).

63. Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., and Zhao, B. Y. Social turing tests: Crowdsourcing sybil detection. In *In Proceedings of the 20th Annual Network & Distributed System Security Symposium* (2013).

64. Wilson, T., Wiebe, J., and Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (2005).

65. Zyskowski, K., Morris, M. R., Bigham, J. P., Gray, M. L., and Kane, S. K. Accessible crowdwork?: Understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing* (2015).

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.

	NYC		SF		LA		Chicago	
	w/o Int.	w/ Int.	w/o Int.	w/ Int.	w/o Int.	w/ Int.	w/o Int.	w/ Int.
Intercept	**-2.892*****	**-2.888*****	**-2.033*****	**-0.041*****	**-2.599*****	**-2.596*****	**-3.475*****	**-3.404*****
Completed Tasks	**0.01*****	**0.01*****	**0.006*****	**0.006*****	**0.012*****	**0.012*****	**0.016*****	**0.016*****
Elite	**0.372****	**0.375****	**0.438*****	**0.436*****	0.232	0.222	0.384	0.405
Member Since	**-0.321*****	**-0.322*****	**-0.303*****	**-0.303*****	**-0.286*****	**-0.28*****	**-0.277****	**-0.287****
Recent Activity	**0.008***	**0.009***	0.003	0.003	0.001	0.001	0.004	0.002
Rating Score	**0.051*****	**0.05*****	**0.047*****	**0.047*****	**0.047*****	**0.047*****	**0.055*****	**0.055*****
Female	-0.073	-0.069	**-0.127***	-0.109	-0.017	-0.049	-0.186	**-0.31***
Asian	0.126	0.004	**-0.245****	-0.201	-0.105	-0.043	**-0.632****	**-1.379*****
Black	**0.137***	**0.166***	0.01	0.04	0.057	-0.042	0.159	0.082
Asian Female		0.256		-0.1		-0.199		**1.189****
Black Female		-0.074		-0.065		0.204		0.163
Observations	1194	1194	845	845	582	582	211	211
Log Likelihood	-3587.8	-3587	-3375	-3374.8	-1777.1	-1776.6	-609.56	-608.08

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:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

	NYC		SF		LA		Chicago	
	w/o Int.	w/ Int.	w/o Int.	w/ Int.	w/o Int.	w/ Int.	w/o Int.	w/ Int.
Completed Tasks	-0.005	-0.005	0	0	-0.006	-0.006	-0.017	-0.017
Elite	**0.683***	**0.683***	0.464	0.46	0.64	0.477	0.318	0.32
Member Since	-0.148	-0.147	0.107		-0.134	-0.142	-0.532	-0.536
Number of Reviews	0.006	0.006	0	0	0.007	0.008	0.02	0.02
Recent Activity	**0.033*****	**0.033*****	-0.002	-0.002	**0.019***	**0.019***	**0.074*****	**0.074*****
Female	-0.069	-0.189	-0.004	-0.01	-0.132	-0.163	0.331	0.312
Asian	-0.211	-0.314	0.111	-0.013	-0.468	-0.631	**2.395****	**2.719***
Black	**-0.292***	**-0.41****	-0.301	-0.0164	-0.07	-0.062	-0.561	-0.621
Asian Female		0.237		0.371		0.495		-0.663
Black Female		0.284		-0.289		-0.006		0.118
Observations	1194	1194	845	845	611	611	211	211
Log Likelihood	-1858.36	-1858.61	-1448.24	-1447.58	-934.73	-934.44	-293.24	-293.12

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

	NYC		SF		LA		Chicago	
	w/o Int.	w/ Int.	w/o Int.	w/ Int.	w/o Int.	w/ Int.	w/o Int.	w/ Int.
Avg. Rating	**-0.011*****	**-0.012*****	0.003	0.004	**0.008*****	**0.01*****	**-0.013*****	**-0.013*****
Completed Tasks	0	0.001	**-0.008*****	**-0.008*****	**-0.018*****	**-0.017*****	-0.01	-0.01
Member Since	**-0.887*****	**-0.85*****	**-0.38*****	**-0.391*****	**-0.24*****	**-0.306*****	**-0.815*****	**-0.788*****
Number of Reviews	**0.004****	**-0.005*****	**0.009*****	**0.009*****	**0.017*****	**0.016*****	0.007	0.008
Recent Activity	0.128	0.127	**-0.092****	**-0.121*****	**-0.209*****	**-0.16****	**-0.41*****	**-0.4*****
Female	**-1.462*****	**-0.595*****	**0.898*****	**0.89*****	0.023	**0.628*****	**0.521*****	**0.716*****
Asian	-0.064	**-1.639*****	0.087	0.148	**-0.867*****	**1.883*****	-0.415	-0.38
Black	**-0.777*****	-0.001	0.158	0.124	**0.83*****	**1.155*****	0.266	**0.386***
Asian Female		**1.669****		-0.68		**-3.754*****		
Black Female		**-0.556***		0.289		**-1.465*****		-0.416
Observations	2257	2257	2801	2801	2299	2299	860	860
Log Likelihood	-6209.79	-6199.6	-8445.9	-8444.01	-6792.3	-6743.3	-3009.02	-3007.88

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:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

	Databases w/ Int.	Animation w/ Int.	Financial w/ Int.	Dieting w/ Int.	Web Analytics w/ Int.	Banner Ads w/ Int.	Songwriters w/ Int.	T-shirts w/ Int.
Intercept	**-2.276*****	**-2.122****	**-2.669*****	**-2.67*****	**-1.814*****	**-1.648****	**-3.022*****	**-3.611*****
About Length	**0.013***	-0.007	**0.02****	0.003	**0.014***	-0.001	**0.021*****	**0.021*****
Avg. Response Time	**0.001*****	**0.001*****	**0.002*****	0	**0.002*****	**0.01*****	**0*****	**0.002*****
Facebook Profile	-0.015	**0.464***	**0.689****	0.09	0.118	0.38	0.274	-0.096
Google+ Profile	0.25	0.303	0.184	-0.072	-0.074	0.087	0.25	0.125
Member Since	**0.866*****	**0.81*****	**0.525*****	**0.726*****	**0.836*****	**0.749*****	**0.898*****	**1.055*****
Rating Score	**1.016*****	**1.138*****	**0.885*****	**1.002*****	**0.88*****	**1.018*****	**0.992*****	**1.198*****
Spoken Languages	**-0.221***	0	-0.116	0.153	**-0.107***	**-0.314***	0.004	-0.006
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	-580.38	-1155.9	-1074.8	-1541.5	-1772.4	-1684.4

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:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

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

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