Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring

Tom Sühr
Technische Universität Berlin
tom.suehr@googlemail.com

Sophie Hilgard
Harvard University
ash798@g.harvard.edu

Himabindu Lakkaraju
Harvard University
hlakkaraju@hbs.edu

ABSTRACT

Ranking algorithms are being widely employed in various online hiring platforms including LinkedIn, TaskRabbit, and Fiverr. Since these platforms impact the livelihood of millions of people, it is important to ensure that the underlying algorithms are not adversely affecting minority groups. However, prior research has demonstrated that ranking algorithms employed by these platforms are prone to a variety of undesirable biases. To address this problem, fair ranking algorithms (e.g., *Det-Greedy*) which increase exposure of underrepresented candidates have been proposed in recent literature. However, there is little to no work that explores if these proposed fair ranking algorithms actually improve real world outcomes (e.g., hiring decisions) for minority groups. Furthermore, there is no clear understanding as to how other factors (e.g., job context, inherent biases of the employers) play a role in impacting the real world outcomes of minority groups.

In this work, we study how gender biases manifest in online hiring platforms and how they impact real world hiring decisions. More specifically, we analyze various sources of gender biases including the nature of the ranking algorithm, the job context, and inherent biases of employers, and establish how these factors interact and affect real world hiring decisions. To this end, we experiment with three different ranking algorithms on three different job contexts using real world data from TaskRabbit. We simulate the hiring scenarios on TaskRabbit by carrying out a large-scale user study with Amazon Mechanical Turk. We then leverage the responses from this study to understand the effect of each of the aforementioned factors. Our results demonstrate that fair ranking algorithms can be an effective tool at increasing hiring of underrepresented gender candidates but induces inconsistent outcomes across candidate features and job contexts.

ACM Reference Format:
Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2020. Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/tba

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/tba

1 INTRODUCTION

Over the past decade, there has been a dramatic increase in online hiring platforms and marketplaces such as LinkedIn, TaskRabbit, and Fiverr. These platforms play a pivotal role in providing employmentto millions of job seekers. For instance, TaskRabbit provides employment opportunities to freelance labor suppliers by connecting them with consumers who are looking for help with everyday tasks such as cleaning, moving, and delivery. Several of these platforms are powered by automated tools and algorithms that determine how job seekers are presented to potential employers–e.g., TaskRabbit leverages ranking algorithms to sort through available candidates and generate a ranked list of candidates suitable for any given task. Since such platforms impact job seekers' livelihood, it is critical to ensure that the underlying algorithms are not adversely affecting underrepresented groups. However, recent research has demonstrated that ranking algorithms employed by various online platforms tend to amplify undesirable biases in the training data [8].

Recent literature on algorithmic fairness tackled the aforementioned challenges by proposing fair ranking algorithms which optimize for different notions of fairness. For example, Zehlike et al. [25] et. al. optimize for a *group fairness* criterion and propose a post processing approach which ensures that the representation of the underrepresented group does not fall below a minimum proportion p at any point in the list. On the other hand, Biega et al. [2] formalize an *individual equity-of-attention* notion of fairness and propose algorithms for fair division of attention between equally relevant candidates. The main idea behind all these fair ranking algorithms is to redistribute user attention across groups or individuals in an equitable fashion [14, 20, 21, 26]. While fair ranking algorithms seem to be a useful first step towards mitigating undesirable biases induced by ranked lists, it is unclear if these algorithms actually improve real world outcomes (e.g., hiring decisions in online portals) for underrepresented groups. Furthermore, there is little to no research that systematically explores how other factors (e.g., inherent biases of employers) interact with ranking algorithms and influence real world outcomes.

In this work, we address the aforementioned gaps in existing literature by studying how gender biases percolate in online hiring platforms and how they impact real world hiring decisions. More specifically, we analyze how various sources of gender biases in online hiring platforms such as the type of the ranking algorithm, the job context, and inherent biases of employers interact with each other and affect hiring decisions. By doing so, we provide answers to some very critical and fundamental questions which have not been systematically explored in existing literature: 1) Do employers exhibit gender bias uniformly across all job contexts and candidates? Or do certain kinds of job contexts promote gender biases more than others? 2) What kinds of ranking algorithms are effective in mitigating gender biases in hiring decisions? Can fair ranking algorithms lead to disparate outcomes when employers exhibit bias? 3) Do certain kinds of employers induce more gender biases into hiring decisions than others? If so, how do we characterize the employers who perpetrate gender biases the most? To the best of our knowledge, this work makes the first attempt at studying the interactions between various factors such as ranking algorithms, job contexts, employer profiles, and analyzing how they impact real world hiring decisions.

To answer the aforementioned questions, we carried out a large-scale user study with 1,079 participants on Amazon Mechanical Turk using real world data from Task Rabbit. Each participant served as a proxy employer and was required to select candidates to help him/her with three different tasks, namely, shopping, event staffing, and moving assistance. To this end, each participant was shown a list of 10 ranked candidates for each task and was asked to select top 4 candidates in each case. We also experimented with three different ranking algorithms, namely, RandomRanking, RabbitRanking, and FairDet-Greedy where candidates are ranked randomly, based on their TaskRabbit relevance scores, and using a fair ranking algorithm called *Det-Greedy* [6] respectively. We additionally created gender swapped versions of each ranking where genders of candidates were swapped from male to female and vice versa with all other information remaining unchanged. Each participant was shown ranked lists generated by one of the three aforementioned ranking algorithms or their gender swapped versions. . We then used the responses collected from this study to carry out our analysis and answer critical questions about percolation of gender biases in online hiring.

Our analysis revealed several critical and surprising insights about gender biases in online hiring. More specifically, we found that fair ranking algorithms can be helpful in increasing the number of underrepresented candidates selected, even after controlling for visibility. However, the effectiveness of fair ranking is mitigated by job contexts in which employers have a persistent gender preference. We find that fair ranking is more effective when underrepresented candidate features are similar to those of the overrepresented class. Further, we find evidence that fair ranking is ineffective at increasing representation when employer selections already represent demographic parity.

2 RELATED WORK

Our work spans multiple topics under the broad umbrella of research on fairness and bias detection. More specifically, our work lies at the intersection of: 1) empirical evidence of gender bias in online portals, 2) fair ranking algorithms and their effectiveness, and 3) user-algorithm interaction. We discuss related work on each of these topics in detail below.

Empirical Evidence of Gender Bias The existence of gender bias in evaluation and hiring settings has been well documented both in online settings and in the real world. For instance, Hannák et al. [8] empirically established the presence of gender and racial biases in reviews and ratings on online marketplaces such as TaskRabbit and Fiverr. They found that female candidates receive fewer reviews on TaskRabbit compared to their male counterparts with equivalent experience. They also found evidence that Black candidates receive worse ratings on TaskRabbit, and both worse ratings and fewer reviews on Fiverr. Nieva and Gutek [16] studied gender biases in evaluations and found strong evidence for pro-male bias.

More recently, Jahanbakhsh et al. [9] investigated the interaction of gender and performance on worker ratings in a simulated teamwork task on Amazon Mechanical Turk. They found that when male and female coworkers were equally low performing, the female worker received worse evaluations. Furthermore, Peng et al. [18] found that increasing the representation of underrepresented candidates can sometimes correct for biases caused by a skewed candidate distribution, but human biases in certain job contexts persist even after increasing representation of the underrepresented group. Other works have studied gender bias in platform contexts other than hiring[5, 11, 13, 19] and different biases other than gender within a hiring setting [1, 24]. Our work is the first to simultaneously manipulate objective quality measures, job contexts, and representation, using an explicit ranking environment.

Fair Ranking Algorithms Our work most closely resembles Geyik et al. [6], which seeks to understand the empirical effects of satisfying a *ranked group fairness criterion* in a deterministic ranking. The ranked group fairness criterion as developed in Zehlike et al. [25] satisfies the properties that at any position in the ranking: 1) all groups are proportionally represented, 2) the relevance of the ranking is maximal subject to this constraint, and 3) within any group, candidates are of decreasing relevance. Geyik et al. [6] conduct an A/B test on LinkedIn data using a post-hoc fairness re-ranking algorithm (*Det-Greedy*) that ensures a desired proportional representation in top-ranked positions by greedily selecting the most relevant candidate available at each position in the ranking while maintaining maximum and minimum representation constraints for each group. In this way, *Det-Greedy* generalizes the FA*IR algorithm developed in [25], allowing for multiple protected groups and arbitrary distribution requirements. *Det-Greedy* is unique in that it is empirically evaluated. However, the authors analyze the effectiveness of the re-ranking with respect to business metrics rather than distribution of outcomes. They observe no decrease in messages sent or accepted on their platform while increasing gender diversity of rankings, but it is not studied whether or not the message recipients are diverse. Celis et al. [3] further study the theoretical guarantees of such ranking constraints.

Other works have focused on non-static rankings which optimize more detailed fairness criteria over a series of rankings. Biega et al. [2] optimize individual-level equity of attention, a measure of whether or not cumulative attention is proportional to cumulative relevance, amortized over successive rankings in which a candidate does not always appear at the same position. Singh and Joachims [20] optimize group fairness of exposure over a probabilistic distribution of rankings. Fair algorithms for learning to rank demonstrate how to satisfy fairness constraints throughout the policy learning process, when relevance is not known a priori [14, 21].

User-Algorithm Interaction Research on manipulated rankings finds that users have a strong bias toward the top items in a list. Joachims et al. [10] attribute this effect partially to trust in the system generating the rankings, although they also find that item relevance mediates the effect of ranking. Keane et al. [12] find similar results but attribute the ranking preference to satisficing rather than trust in ranking systems, as they find in [17] that this effect persists even with a simple text list of items. A study of Amazon Mechanical Turk workers finds that algorithm users have a strong preference for *demographic parity* as a measure of fairness and are likely to prioritize accuracy over fairness in high stakes situations[22], potentially reducing the effectiveness of fairness-promoting recommendations. Further, it has been demonstrated that algorithms intended to increase objectivity can result in disparate outcomes when biased users have agency to accept or reject the algorithmic recommendations [7].

3 PROBLEM FORMULATION

Previous work has documented the existence of gender biases in hiring settings and shown how these biases can be affected by ranking algorithms, either aggravating existing bias with uncontrolled feedback loops or mitigating bias by increasing exposure of disadvantaged candidates. Unfortunately, current work in understanding bias in fair rankings falls into one of two camps: Empirical studies of platform rankings use real human selection data but suffer from confounders; If candidate features (e.g. star ratings) correlate with gender in some settings, bias cannot be fully disentangled from selection for apparently objective qualities. Simulation studies guarantee the satisfaction of fairness properties under specific assumptions, but it is not known whether the human choice models they assume are realistic in practice. Our approach uses real human decisions to identify where gender biases exist in online hiring, how bias effects interact with traditional gender roles and candidate qualifications, and under what circumstances fair ranking can be expected to be effective. We further study which groups of employers may drive online hiring biases.

- **RQ 1:** Do employers show gender bias which persists across rankings and candidate features? Is gender bias universal or tied to traditional gender roles?
- **RQ 2:** Are algorithmic rankings effective at enhancing or reducing bias when controlling for visibility? If so, are they equally effective in all settings, or do candidate and job features affect human interaction with ranking models?
- **RQ 3:** Can fair ranking lead to disparate outcomes when employers exhibit bias?
- **RQ 4:** Which employer demographic groups perpetrate hiring bias in our setting?

4 STUDY DESIGN

In order to study how bias in *human decisions* varies by job context, worker features, and ranking algorithm, we conduct large scale experiments on Amazon Mechanical Turk. Each of the 1079 participants is exposed to three different ranking tasks, in which they view 10 ranked candidates and are asked to select, in ranked order, their top four candidates (see Figure 1).

4.1 Job Contexts

Candidate features reflect real TaskRabbit data acquired by performing queries for **Shopping, Event Staffing**, and **Moving Assistance** in different US cities. These categories represent those found by Hannák et al. [8] to have varying levels of bias in favor of male workers, with Shopping exhibiting large biases in favor of male workers and Moving the smallest bias. [1]

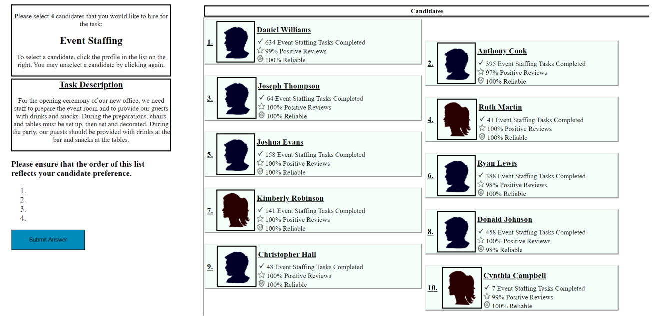


Figure 1: User interface for a ranking in our experiment

Rank	1	2	3	4	5	6	7	8	9	10
D1	m	m	m	f	m	m	f	m	m	f
D2	m	m	m	m	m	f	m	f	m	f
D3	m	m	m	m	m	m	m	f	f	f

Table 1: Rank and gender of the retrieved Task Rabbit data

Query regions varied to avoid repeating individual users' data across tasks, and specific locations were refined until the returned ranking contained 3 female candidates among the top 10, most of whom were ranked in the bottom 5. We identified the perceived gender of the candidates manually through profile pictures and pronouns in their description and user reviews. Excluding rankings with more than one female candidate in the top 5 positions ensured rankings for which fair ranking according to a *ranked group fairness constraint* would be meaningful, as women were underrepresented in the top 5 positions relative to the overall distribution of TaskRabbit [8]. Excluding rankings in which three women did not appear in the top 10 allowed us to limit our candidate lists to control for scrolling effects: because only 10 candidates are simultaneously visible in our UI, if 3 women did not already appear among the top 10, fair ranking would require that new candidates be upranked to those positions. This would cause the initial set of visible candidates to change across conditions. Table 1 shows the exact rank-gender distribution for each data set.

We note that the goal of this process is not to retrieve a dataset representative of TaskRabbit worker rankings, but to sample data for real scenarios in which fair ranking may be expected to benefit underrepresented workers.

4.2 Worker Features

This process leaves us with three datasets, each corresponding to one of three tasks. We refer to these datasets as **D1, D2,** and **D3**. We extracted the features *number of completed tasks*, *% positive reviews* and *% reliable* from the datasets and manually added a new attribute *gender* as determined manually through profile pictures, reviews and profile description.

[1]No studied contexts favored female workers.

	Avg. # tasks		S.d.		Avg. % pos reviews		S.d.	
	m	**f**	**m**	**f**	**m**	**f**	**m**	**f**
D1	306.43	63.0	220.74	69.66	99.14	99.67	1.21	0.58
D2	1.57	5.0	2.15	4.36	98.14	97.0	1.77	3.61
D3	0.57	0.0	0.79	0.0	97.43	100.0	4.43	0.0
	Avg. % reliability		**S.d.**		**Avg. Rabbit Score**		**S.d.**	
	m	**f**	**m**	**f**	**m**	**f**	**m**	**f**
D1	99.71	100.0	0.76	0.0	0.85	0.84	0.01	0.01
D2	96.29	100.0	8.2	0.0	0.76	0.72	0.03	0.02
D3	99.43	92.67	1.51	8.74	0.73	0.68	0.03	0.01

Table 2: Feature distributions for data sets

Candidates' names were removed and replaced with a set of 30 different first-last name combinations generated from the most common white first names [23] and last names [15]. We do this in order to exclude any racial or ethnic biases.

TaskRabbit shows additional features such as price per hour, "Elite Tasker" tags, "Great Value" tags, recent reviews and profile descriptions. We chose to exclude these features as they may be expected to affect selection in a way that confounds worker relevance. Table 2 shows that there are large differences for the features *number of tasks completed* between data sets and between genders. The disparities in TaskRabbit-assigned relevance score are smallest for **D1** and smallest for **D3** with little within-group variation. We consider the primary characteristics of the datasets to be:

- **D1**: In this dataset, the overrepresented candidates (men, in the unswapped rankings) have substantially more tasks completed than the disadvantaged candidates, while the percentage of positive reviews and reliability approximately equal but slightly favor the disadvantaged candidates.
- **D2**: In this dataset, candidates have only completed a few tasks, with one each from the overrepresented group and the disadvantaged group having > 5 tasks completed. The disadvantaged candidates all score high on percentage of positive reviews and reliability, while there is more variance for the overrepresented group.
- **D3**: In this dataset, none of the disadvantaged candidates have completed any tasks at all, and the overrepresented candidates have completed between 0 and 2 tasks each. Several candidates with no tasks completed (3 overrepresented, 1 disadvantaged) have maximum scores for the other features. The candidate who has the most tasks completed (2) also has the lowest percentage of positive reviews. One of the disadvantaged candidates has 100% positive reviews but a low reliability score.

We note that because this is real data, we did not attempt to engineer the features to capture specific hiring patterns among our employers. To decouple the effects of worker data and task context, in our experiment each dataset appears with each task an equal number of times.

4.3 Ranking Algorithms

For our first algorithmic ranking, we additionally extract from the TaskRabbit data a TaskRabbit-generated relevance score. The second algorithm produces a random ordering of the candidates for each study participant. As our third algorithm, we implemented LinkedIn's *DetGreedy* algorithm [6], as a post-processing method to the TaskRabbit ranking. We will refer to the used algorithms as follows:

- RabbitRanking: Candidates ranked by their Task Rabbit relevance score
- RandomRanking: Candidates ranked in a random order
- FairDet-Greedy: Candidates ranked by Det-Greedy [6] applied to the Task Rabbit relevance scores

We chose $p_{male} = 0.58$ and $p_{female} = 0.42$ for our experiments. According to Hannák et al. [8], this is the actual gender distribution on Task Rabbit. For the resulting ranking order see table 3.

To study whether there are disparate effects of fair ranking across genders, we additionally create versions of each ranking in which the data remains the same but genders are swapped from female to male and vice versa. We will refer to the algorithms with swapped genders as RabbitRanking(F↔M), RandomRanking(F↔M) and FairDet-Greedy(F↔M).

4.4 HIT Design

We recruited the participants for our study on Amazon Mechanical Turk. [2] We limited the worker selection to the U.S. and to workers who had at least 5,000 approved tasks with 95% approval rate. We compensated all workers with a base rate of \$0.70 and a bonus of \$0.15. Every approved worker received the bonus. We disapproved workers only if they did not attempt to reasonably answer the text field of the survey. With an average completion time of 6 minutes, we paid an average hourly wage of \$8.50. A total of 1079 participated in our study of which 55% identified themselves as male and 45% as female. As seen in Figure 2, 61% of the participants were between 25 and 44 years old, while 15% were older than 54 years and 5% between 18 and 24 years old. The highest level of education was a high school diploma for 25%, a college diploma for 56% and a master or PhD diploma for 19% of the participants. The median household income was between $\$30,000 - \$59,000$.

Each participant first viewed a short briefing in which they read the instructions of the task and then answered two comprehension questions and one attention check. Workers were not allowed to proceed until they answered these questions correctly. Workers then interacted with the simulated job candidate selection process seen in Figure 1. In each of 3 different job contexts, the workers were asked to select 4 candidates, in order of preference, to recommend to a company. Participants were told that if the company hired at least one candidate whom the participant had recommended, they would receive a bonus of \$0.15 (every participant received this bonus). Participants were also told that candidates provide a full resume to the platform and a computer algorithm analyzes these information and ranks them *"according to criteria including how likely they are to be hired and successfully complete the task."* A task

[2] All experiments were approved by our university's IRB.

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Det-Greedy(0.42, 0.58)	m	f	m	f	m	f	m	m	m	m

Table 3: Fair ordering for all data sets using FairDet-Greedy with $p_{male} = 0.58$ and $p_{female} = 0.42$.

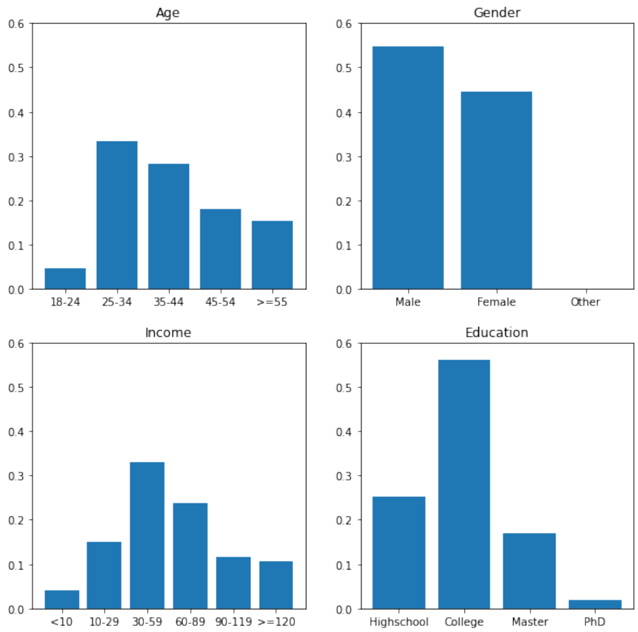


Figure 2: Distribution of age, gender, education and household income among employers in our sample

description was displayed to limit the possible interpretations of the tasks.

Participants were randomly assigned to one of the ranking algorithms in either the original gender of the dataset or with swapped genders. Participants completed one selection task for each job context, and datasets were cycled between participants such that each dataset appeared with each job context an equal number of times. The order of the job contexts was randomized.

After the selection process, we asked participants to rate the importance of each the displayed features for each job context on a 5-point Likert scale. We also asked them how much they trusted the computer system's assessment of the candidates on a 5-point Likert scale. We then asked the participants to describe their decision-making process with at least 40 characters in a free text field. Participants optionally self-reported gender, age, education level and household income.

5 RESULTS

We find that:

- Our employers do exhibit pro-male gender bias after controlling for candidate features and ranking, however this bias is only significant for our traditionally male-dominated job context (moving).

	Probability of Selection (w/o Interactions)		**Probability of Selection (w/ Interactions)**	
	4 Selections	1 Selection	4 Selections	1 Selection
(Intercept)	**-0.670*****	**-2.316*****	**-0.761*****	**-2.358*****
Job Moving Assistance	-0.005	0.025	**0.115*****	0.113
Job Shopping	-0.007	0.014	-0.001	-0.019
Female in Moving A.			**-0.426*****	-0.374
Female in Shopping			-0.019	0.127
Positive Reviews	**0.244*****	**0.155*****	**0.143*****	**0.143***
Reliability	**1.230*****	**1.570*****	**1.368*****	**1.571*****
Completed Tasks	**0.215*****	**0.281*****	**0.235*****	**0.295*****
Rank	**-0.226*****	**-0.716*****	**-0.215*****	**-0.728*****
Female	**0.171*****	-0.071	0.028	**-0.548***
Positive Reviews for Females			**0.543*****	**0.222*****
Reliability for Females			-0.028	0.366
Completed Tasks for Females			**-0.744*****	**-1.501*****
Rank for Females			0.007	0.072

Table 4: Logistic Regression with clustered standard errors to predict the selection of candidates. Note that the Intercept represents baseline probability of a male candidate being selected in the event staffing job context. $*** = p < \frac{0.001}{4}$; $** = p < \frac{0.01}{4}$; $* = \frac{0.05}{4}$

- Controlling for visibility (all candidates displayed in a single page), upranking underrepresented gender candidates increases the probability of selection. The size of this effect is mediated by candidate and job features.
- Underrepresented men are selected at a higher rate than equivalent underrepresented women. Most of this effect comes specifically from the traditionally male-dominated job context (moving). We do not find statistically significant evidence that fair ranking induces additional disparate effects for underrepresented candidates based on their gender.

In analyzing employer bias and the effectiveness of fair ranking algorithms, we focus only on the data in which male candidates tend to be ranked higher than female candidates. Because the relevance scores are provided by TaskRabbit, we cannot know whether flipping the gender of the candidates breaks correlations that are relevant to the ranking algorithm, e.g. through implicit feedback. For this reason, we reserve the data with flipped genders for examining disparate effects of underrepresentation and fair ranking.

5.1 Employer Bias

To study whether employers exhibit a gender bias after controlling for ranking and candidate features, we use only the data where male candidates tend to be ranked higher than female candidates and fit a logistic regression predicting whether a candidate will be selected among the top n candidates on the candidate features, candidate ranking, and candidate gender. We use employer rankings to study the effect on underrepresented women at each of 1, 2, 3, and 4 selections, assuming that the employer ranks first the candidate they would have selected if they were only allowed to select one, and so on. Because we repeat this test 4 times, we apply a Bonferroni correction to resulting p-values. Data is standardized such that all variables have a mean of 0 and standard deviation of 1, allowing for comparison of the coefficients. We additionally cluster the standard errors on mTurk WorkerID to account for dependencies in our

TaskRabbit vs. Fair	**Underrepresented candidates per selection (w/o Interactions)**				**Underrepresented candidates per selection (w/ Interactions)**			
	4 Selections	3 Selections	2 Selections	1 Selection	4 Selections	3 Selections	2 Selections	1 Selection
(Intercept)	**-0.780*****	**-0.632*****	**-0.891*****	**-1.020*****	0.029	**0.463*****	0.270	0.373
Moving	-0.075	-0.075	-0.035	0.000	**-1.496*****	**-1.977*****	**-2.118*****	**-2.485*****
Moving D2					**1.326*****	**1.497*****	**1.389*****	**2.275*****
Moving D3					**2.495*****	**3.559*****	**3.841*****	**4.729*****
Moving and FLIP					**0.396****	**0.478***	**0.733***	0.504
Shopping	0.010	-0.007	0.130	0.144	**-1.281*****	**-1.825*****	**-1.760*****	**-2.140*****
Shopping D2					**2.652*****	**3.607*****	**3.458*****	**4.138*****
Shopping D3					**1.247*****	**2.191*****	**2.261*****	**2.233*****
Shopping and FLIP					0.030	-0.053	0.191	0.324
RabbitRanking	**-0.183*****	**-0.283*****	**-0.373*****	**-0.633*****	-0.097	-0.195	**0.443***	**-0.964****
FLIP	**-0.091***	-0.080	-0.129	-0.008	-0.228	-0.203	-0.410	-0.396
D2	**-0.007**	-0.081	0.042	0.032	**-1.335*****	**-1.801*****	**-1.636*****	**-2.209*****
D3	0.032	-0.081	-0.025	-0.104	**-1.105*****	**-1.827*****	**-1.876*****	**-2.246*****

Table 5: Logistic regression with clustered standard errors for the number of underrepresented candidates per selected candidate comparing RabbitRanking, RabbitRanking(F↔M) **and** FairDet-Greedy, FairDet-Greedy(F↔M). **Note:** $*** = p < \frac{0.001}{4}$; $** = p < \frac{0.01}{4}$; $* = \frac{0.05}{4}$

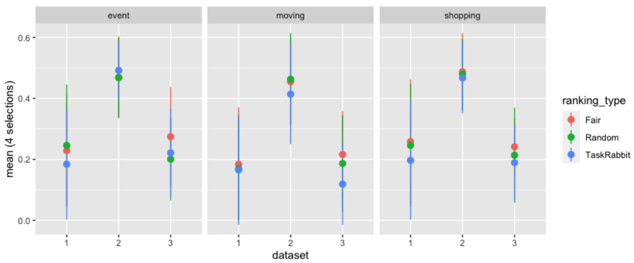
data and avoid overreporting significance. Results appear in Table 4. We find that when we do not include interaction terms, the dummy variable on female gender has a positive coefficient ($p < .001$), suggesting that female candidates are selected more often than might be expected given their features. This may be due to employers attempting to exhibit some demographic parity across all of their selections, despite men having higher relevance - in particular higher tasks completed. For more discussion, see Section 5.6. When interaction terms for all features are included, we find that female gender is not significant across all tasks, but there is a significant negative interaction term between female gender and a moving job context ($p < .001$), suggesting that females are less likely to be chosen for this job relative to event staffing jobs, as well as a significant positive term for moving job context for male candidates ($p < .001$), suggesting that men are more likely to be hired in this context than event staffing. We also find that females tend to benefit more from positive reviews compared to male candidates ($p < .001$). .

5.2 Effectiveness of Fair Rankings

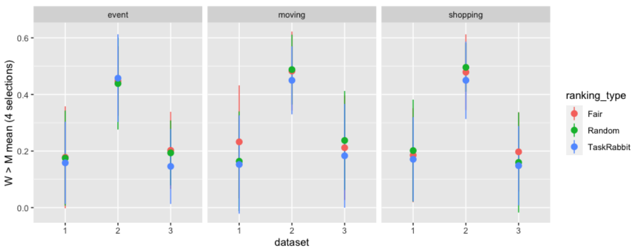
To study whether applying a post-hoc fair re-ranking to the data has a significant effect, we conduct a 3-level ANOVA on a linear model predicting the % of candidates selected who are female (we call this variable *selected*) on instrumental variables of ranking type, dataset, and job context. We again study the effect on underrepresented women at each of $n = \{1, 2, 3, 4\}$ and apply a Bonferroni correction to resulting p-values. Without interaction terms, the F-statistic confirms that ranking type and dataset are statistically significant ($p < .001$) for all numbers of selections, and task is significant except for the first selection ($p < .001$). This remains true when interaction terms are added to the model. We follow by analyzing both Wald tests with clustered standard errors and Tukey's HSD statistics on the pairwise effects and find that RabbitRanking and FairDet-Greedy differ significantly across all numbers of choices when interactions are not included ($p < .001$), with underrepresented candidates doing better under FairDet-Greedy. When we do include interaction terms, we find that the significance of RabbitRanking vs FairDet-Greedy continues to be significant ($p < .001$) at all selection levels by Tukey's HSD and but is only significant at choices 1 and 2 ($p < .05$ before correction) by the Wald test. We find that interaction terms are insignificant by Wald Test with robust standard errors due to small sample sizes, but exploratory analysis (see Figure 3a) reveals that dataset and task both affect the efficacy of the FairDet-Greedy. In particular, FairDet-Greedy appears to be most successful at increasing the proportion of underrepresented candidates selected in **D3**, in which the underrepresented candidates receive lower rankings but are not strongly differentiated by their features. In **D1**, underrepresented candidates have significantly fewer jobs completed relative to the overrepresented candidates, which appears to mitigate the effectiveness of FairDet-Greedy relative to RabbitRanking. In **D2** underrepresented candidates have, on average, better features than their overrepresented counterparts and thus already are selected approximately equally often despite their low rankings. This, too, appears to mitigate the effect of FairDet-Greedy, as employers potentially target demographic parity.

5.3 Disparate Impact of Fair Rankings

To study whether fair ranking algorithms may exhibit a disparate impact on gender groups, we conduct a linear regression predicting percentage of underrepresented candidates chosen (female in the original data, male in the swapped data) on categorical variables representing the dataset, job context, and ranking type, FairDet-Greedy vs RabbitRanking, as well as a variable **FLIP** that indicates whether the data is from the counterfactual world in which men are underrepresented. If identical underrepresented female candidates and underrepresented male candidates are treated equally, we expect to see no significant coefficients on **FLIP** or any of its interactions. We find instead that without interactions, **FLIP** is significant with a negative coefficient for 4 selections, suggesting that underrepresented men are *less* likely to be selected than their female



(a) Underrepresented selection, men ranked higher than women



(b) Underrepresented selection, women ranked higher than men

counterparts. With interactions, **FLIP** is not significant across all conditions, but has a significant positive interaction with the moving job context, again confirming a persistent human preference. We find no significant interactions between **FLIP** and RabbitRanking, suggesting that fair ranking does not impose consistent disparate effects on men and women. Three-way interactions are limited by sample size, however, exploratory analysis (See Figure 3b) suggests that the fair ranking has an outsize effect on men for the moving job context exactly when they appear to be underqualified relative to women, in **D1**. This potentially aligns with findings of disparate impact in recidivism prediction settings, in which we find that judges are more likely to use algorithmic recommendations when they confirm existing biases [7].

5.4 Impact of Fair Rankings Over Sequential Selection

In this section we highlight patterns in the impact of FairDet-Greedy across sequential selections in various job contexts and data sets. Because of the large number of pairwise comparisons and small sample sizes, we do not focus on statistical significance here but rather directional effects.

Table 6 shows the percentage of female candidates of all selections made. For example, in the case of data set **D1**, Moving job context and algorithm RabbitRanking, 16.53% of all selected candidates were female while 4.84% of all first selections were female. The row *all* compares the selections of female candidates for all three data sets.

Our results reveal that FairDet-Greedy increases number of selected female candidates in almost all cells compared to RabbitRanking. FairDet-Greedy is particularly effective in increasing female representation at employers’ first selection in all job contexts . We find the highest increase of 17.35 percentage points in the first selection for the Moving job context followed by Event staffing with an increase of 13.02 percentage points. Across all job contexts and datasets, we observe that the difference between FairDet-Greedy and RabbitRanking decreases as the number of selections increases, suggesting that FairDet-Greedy mainly pushes female candidates higher in the priority list of users but has less effect on the overall number of selected female candidates.

Moving assistance												
	Task Rabbit				Random				Fair			
#Choices	1	2	3	4	1	2	3	4	1	2	3	4
D1	4.84%	9.68%	15.59%	16.53%	11.29%	16.13%	16.13%	16.94%	12.28%	16.67%	18.71%	18.42%
D2	24.59%	39.34%	44.26%	41.39%	44.26%	48.36%	53.01%	46.31%	60.66%	46.72%	51.91%	45.49%
D3	1.64%	5.74%	9.84%	11.89%	10.17%	11.86%	14.12%	18.68%	8.47%	14.41%	16.95%	21.61%
all	10.33%	18.25%	23.23%	23.27%	21.98%	25.45%	27.75%	27.30%	27.68%	25.93%	29.19%	28.51%
Event staffing												
	Task Rabbit				Random				Fair			
#Choices	1	2	3	4	1	2	3	4	1	2	3	4
D1	4.92%	11.48%	16.94%	18.44%	10.17%	22.88%	24.29%	24.58%	16.95%	22.03%	22.60%	22.88%
D2	37.10%	46.77%	58.06%	49.19%	48.39%	48.39%	54.84%	46.77%	52.63%	50.88%	56.14%	46.93%
D3	4.92%	12.30%	14.75%	22.13%	8.20%	10.66%	13.11%	20.08%	16.39%	20.49%	23.50%	27.46%
all	15.64%	23.51%	29.92%	29.92%	22.25%	27.31%	30.75%	30.48%	28.66%	31.13%	34.08%	32.42%
Shopping												
	Task Rabbit				Random				Fair			
#Choices	1	2	3	4	1	2	3	4	1	2	3	4
D1	6.56%	14.75%	15.30%	19.67%	16.39%	22.95%	23.50%	24.59%	11.48%	26.23%	28.42%	25.82%
D2	47.54%	49.18%	52.46%	46.72%	49.15%	53.39%	57.06%	47.88%	54.24%	55.08%	57.06%	48.73%
D3	8.06%	11.29%	13.44%	18.95%	8.06%	12.1%	17.2%	21.37%	17.54%	20.18%	22.81%	24.12%
all	20.72%	25.07%	27.07%	28.45%	24.54%	29.48%	32.59%	31.28%	27.75%	33.83%	36.09%	32.89%

Table 6: Female selection rates for the cumulative selections 1-4. Colors indicate the change in ppt compared to RabbitRanking:

[-10,-5] (-5,0) |0| (0,+5) | [+5,+10) | [+10,+15) | [+15,∞)

Moving assistance												
	Task Rabbit_swapped				Random_swapped				Fair_swapped			
#Choices	1	2	3	4	1	2	3	4	1	2	3	4
D1	5.08%	9.32%	14.69%	15.25%	14.75%	16.39%	15.85%	16.39%	10.53%	18.42%	24.56%	23.25%
D2	38.33%	48.33%	52.22%	45.00%	48.39%	51.61%	54.84%	48.79%	48.28%	51.72%	57.47%	48.28%
D3	8.33%	10.00%	14.44%	18.33%	15.00%	16.67%	21.67%	23.75%	18.64%	15.25%	16.38%	21.19%
all	17.25%	22.55%	27.12%	26.20%	26.05%	28.22%	30.78%	29.64%	25.82%	28.47%	32.81%	30.90%
Event Staffing												
	Task Rabbit_swapped				Random_swapped				Fair_swapped			
#Choices	1	2	3	4	1	2	3	4	1	2	3	4
D1	6.67%	12.50%	14.44%	15.84%	13.33%	15.00%	16.67%	17.50%	8.47%	16.95%	17.51%	17.80%
D2	37.29%	46.61%	55.37%	45.76%	42.62%	45.08%	50.27%	43.85%	52.63%	45.61%	55.56%	44.74%
D3	1.67%	1.67%	6.67%	14.58%	6.45%	12.90%	17.74%	19.35%	5.17%	7.76%	16.09%	20.26%
all	15.21%	20.26%	25.49%	25.39%	20.80%	24.33%	28.23%	26.90%	22.09%	23.44%	29.72%	27.60%
Shopping												
	Task Rabbit_swapped				Random_swapped				Fair_swapped			
#Choices	1	2	3	4	1	2	3	4	1	2	3	4
D1	5.00%	6.67%	13.89%	17.08%	4.84%	13.71%	18.82%	20.16%	13.79%	18.10%	18.39%	18.53%
D2	38.33%	45.83%	49.44%	45.00%	53.33%	53.33%	60.00%	49.58%	49.15%	50.85%	54.80%	47.88%
D3	5.08%	7.63%	10.17%	14.83%	14.75%	12.30%	12.02%	15.98%	14.04%	16.67%	17.54%	19.74%
all	16.14%	20.04%	24.50%	25.64%	24.31%	26.45%	30.28%	28.58%	25.66%	28.54%	30.25%	28.72%

Table 7: Male selection rates for the cumulative selections 1-4. Colors indicate the change in ppt compared to RabbitRanking(F↔M)**:**

[-10,-5] (-5,0) |0| (0,+5) | [+5,+10) | [+10,+15) | [+15,∞)

3 female candidates 7 male candidates				
	Mean selected female candidates per user	Moving assistance	Event staffing	Shopping
Moving assistance	1.06	-	0.18***	0.17***
Event staffing	1.24	-	-	0.01
Shopping	1.23	-	-	-

Table 8: Average selected female candidates per user compared between job contexts. $*** = p < 0.01$; $** = p < 0.05$; $* = p < 0.1$; $. = p < 0.15$

5.5 Gender bias of different employer groups

In order to investigate if consistent human gender preferences we observer are limited to specific demographic groups of employers, we limit our analysis to the moving job context, in which we see the largest effects of gender bias across all other conditions. We perform a logistic regression to predict the number of selected

3 male candidates 7 female candidates				
	Mean selected male candidates per user	Moving assistance	Event staffing	Shopping
Moving assistance	1.16	-	0.10*	0.04
Event staffing	1.06	-	-	0.05
Shopping	1.11	-	-	-

Table 9: Average selected male candidates per user compared between job contexts. For the scenarios with swapped genders $*** = p < 0.01; ** = p < 0.05; * = p < 0.1; . = p < 0.15$

female candidates at each selection level in this job context, on the gender and age group of the employer. We include both the original data conditions and the gender flipped data conditions due to small sample size. In the first two priority selections, we find negative coefficients for male employers who are between 25 and 34 years old ($p = 0.053$) and older than 54 years. All other age categories have negative coefficients for male employers, but they are not significant, suggesting that male employers are less likely to hire women for this job context. On the other hand, all coefficients for female employers are positive, significantly for women between 25 and 34 years of age ($p = 0.055$). Other job contexts show no such evidence of demographic-specific biases.

5.6 Employer Understandings of Fairness

We hypothesized that at least some employers try to select candidates with the goal of demographic parity. In order to confirm this, we took a random sample of 100 employer responses of at least 200 characters and counted the free text responses in which they described a preference for gender parity. We find that 12.4% of employers actively tried to make a diverse choice and stated that in the text field question of our survey. During this analysis, we also found comments of employers who considered gender in more specific ways. Interestingly, instead of trying to make it fair *within* a job context, some employers tried to achieve gender parity *between* the tasks, a behavior which is likely to exacerbate gender bias when hiring across roles with opposing gender biases.

Conversely, 35% of respondents instead display a preference for *individual fairness*, treating similar individuals similarly without consideration for gender [4], in their text responses. Likert scale responses for the importance of gender in their decision making reveal that the majority of respondents (59%, 67%, and 66% in moving, event staffing, and shopping, respectively) report that they did not consider gender at all in their selections. Future work could further investigate how employers' mental models of fairness interact with their practical use of fair rankings.

6 CONCLUSIONS

We conduct the first user study to explore the effectiveness of fair ranking algorithms on candidate outcomes, including manipulations to understand the interplay between ranking, candidate features, job context, and human biases. Our study reveals that fair ranking can successfully increase the opportunities available to underrepresented candidates, particularly when candidate features are similar between the underrepresented and overrepresented groups. However, we find that the effectiveness of fair ranking is inconsistent across job contexts and candidate features, suggesting that it may not be sufficient to increase representation outcomes in all settings. We hope that this work represents a step toward better understanding how algorithmic tools can (or cannot) reduce gender bias in hiring settings.

REFERENCES

[1] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[2] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

[3] L Elisa Celis, Anay Mehrotra, and Nisheeth K Vishnoi. 2020. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 369–380.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[5] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM conference on recommender systems*. 242–250.

[6] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2221–2231.

[7] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 90–99.

[8] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1914–1933.

[9] Farnaz Jahanbakhsh, Justin Cranshaw, Scott Counts, Walter S Lasecki, and Kori Inkpen. 2020. An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[10] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[11] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.

[12] Mark T Keane, Maeve O'Brien, and Barry Smyth. 2008. Are people biased in their use of search engines? *Commun. ACM* 51, 2 (2008), 49–52.

[13] Anna May, Johannes Wachs, and Anikó Hannák. 2019. Gender differences in participation and reward on Stack Overflow. *Empirical Software Engineering* 24, 4 (2019), 1997–2019.

[14] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. *arXiv preprint arXiv:2005.14713* (2020).

[15] namecensus.com. 2000. Most common last names for Whites in the U.S. https://namecensus.com/data/white.html. Accessed: 2020-07-15.

[16] Veronica F Nieva and Barbara A Gutek. 1980. Sex effects on evaluation. *Academy of management Review* 5, 2 (1980), 267–276.

[17] Maeve O'Brien and Mark T Keane. 2006. Modeling result-list searching in the World Wide Web: The role of relevance topologies and trust bias. In *Proceedings of the 28th annual conference of the cognitive science society*, Vol. 28. Citeseer, 1881–1886.

[18] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? The impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 125–134.

[19] Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2020. Exploring Artist Gender Bias in Music Recommendation. *arXiv preprint arXiv:2009.01715* (2020).

[20] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.

[21] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems*. 5426–5436.

[22] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International*

Conference on Knowledge Discovery & Data Mining. 2459–2468.
[23] Social Security Administration (SSA). 2019. Top Names Over the Last 100 Years. https://www.ssa.gov/oact/babynames/decades/century.html. Accessed: 2020-07-15.
[24] Jacob Thebault-Spieker, Loren G Terveen, and Brent Hecht. 2015. 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 & Social Computing*. 265–275.
[25] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.
[26] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.