

# OK Computer: Worker Perceptions of Algorithmic Recruitment\*

Elena Fumagalli<sup>†</sup>  
*Utrecht University*

Sarah Rezaei<sup>‡</sup>  
*Innsbruck University*

Anna Salomons<sup>§</sup>  
*Utrecht University*

July 4, 2021

## Abstract

We provide evidence on how workers on an online platform perceive algorithmic versus human recruitment, through two incentivized experiments designed to elicit willingness to pay for human or algorithmic evaluation. In particular, we test how information on workers' performance affects their recruiter choice and whether the algorithmic recruiter is perceived as more or less gender-biased than the human one. We find that workers do perceive human and algorithmic evaluation differently, even though both recruiters are given the same inputs in our controlled setting. Specifically, human recruiters are perceived to be more error-prone evaluators, and place more weight on personal characteristics, whereas algorithmic recruiters are seen as placing more weight on task performance. Consistent with these perceptions, workers with good task performance relative to others prefer algorithmic evaluation, whereas those with lower task performance prefer human evaluation. However, we find only limited evidence that perceived differences in gender bias are substantial enough to drive preferences for human versus algorithmic recruitment.

**Keywords:** Algorithmic evaluation; Technological change; Online labor market; Online experiment

**JEL:** C9, J24, M51, O33

---

\*Helpful input from Dominik Prugger on an earlier version of this project is gratefully acknowledged. We are grateful to Maria Savona and two anonymous reviewers for helpful comments, and to Lucky Belder and Nina Bontje for helping to interpret legal provisions concerning algorithmic versus human evaluation. This work was supported by a Utrecht University Institutions seed money grant. Salomons thanks Instituut Gak for financial support.

<sup>†</sup>Utrecht University School of Economics, e.fumagalli@uu.nl

<sup>‡</sup>Corresponding author. Innsbruck University, Department of Economics, sarah.rezaei@uibk.ac.at

<sup>§</sup>Utrecht University School of Economics, a.m.salomons@uu.nl

# 1 Introduction

Algorithmic predictions are increasingly used in decision-making as a result of ongoing advances in artificial intelligence (AI) and data infrastructure, and labor markets are an important domain for such applications. Specifically, algorithmic input is widely used when hiring or otherwise evaluating workers (Ajunwa and Greene, 2019), for example through automated CV-screening and other algorithmic assessment services (WSJ, 2012; Carey and Smith, 2016; Raghavan et al., 2020).<sup>1</sup> Broadly put, regression-based or machine-learning models are used to predict workers’ future job performance, leveraging data on current and previous job applicants and/or existing workers at the firm.

Despite a large body of work evaluating the performance of algorithms in predicting outcomes across a wide range of settings, much less is known about how people interact with these technologies. For labor markets in particular, studies on the use of algorithms have focused on the employer perspective<sup>2</sup>: as yet, there is only limited evidence on how *workers* perceive algorithmic versus human evaluation, as we discuss below. However, this question is critical for understanding how the increasing use of algorithms in the labor market affects worker welfare and organizational commitment. Worker welfare also matters for governments aiming to regulate the use of algorithms in labor markets.

In this paper, we study workers’ perceptions of and preferences for algorithmic evaluation, as compared to human evaluation. In particular, we mimic a hiring setting, using workers from Amazon’s online labor market Mechanical Turk (MTurk) as subjects. We design experiments to uncover causal evidence about the factors driving preferences for human versus algorithmic prediction, allowing us to abstract from considerations that

---

<sup>1</sup>Algorithmic recommendations are also being used to guide workers’ job search (e.g. see Belot et al. 2018; Goos et al. 2019); and for predicting worker turnover to maximize worker retention by for example IBM’s Watson Analytics (IBM, 2016).

<sup>2</sup>This literature suggests that algorithmic screening is a valuable hiring tool for firms, because it saves on hiring costs and can lead to a better prediction of applicants’ job performance, a higher match quality, and increased worker retention. Specifically, firms receiving algorithmically recommended applicants have higher fill rates for vacancies without crowding out non-recommended job applicants (Horton, 2017); and algorithmically recommended candidates are more likely to pass interviews and receive a job offer, more likely to accept job offers when extended, and also more productive once hired as employees (Cowgill, 2020). Combined with the low (marginal) cost of automated versus human evaluation, these findings help explain the broad appeal of these technologies to employers.

are unrelated to differences in evaluation. To our knowledge, we provide the first incentivized analyses of workers’ preferences for algorithmic versus human evaluation, what these preferences are based on, and how they differ across different worker subgroups.

Workers are registered in the U.S. or Canada and selected on quality.<sup>3</sup> After performing a job task, we elicit workers’ willingness to pay for their preferred recruiter, human or algorithmic. Importantly, we inform workers that both recruiters have access to the same information: performance on a job test as well as a set of pre-determined personal characteristics (age, education, gender, and ethnicity) that are usually available on (or can be inferred from) a CV. Further, all recruitment decisions are communicated online, at the same time, and without in-person contact with human recruiters. This allows us to isolate as much as possible any preferences based on recruiter decision-making, rather than other factors such as time preferences, (dis)tastes for interpersonal interaction, or concerns about data privacy.

We go beyond describing incentivized preferences by implementing two experiments in this hiring setting. These interventions are aimed at understanding the causal impacts on recruiter choice of the two major inputs in recruitment decisions: workers’ observed task performance and their personal characteristics (specifically gender). In the first experiment (*information experiment*), we study the effect of task performance on recruiter preference by randomly providing workers with information about their task performance compared to others. In the second experiment (*gender bias experiment*), we focus on gender bias by having workers compete against a fictitious worker whose gender is randomized. This allows us to obtain worker perceptions of recruiters’ relative gender bias.

Our work is related to two strands of literature. The first considers how users, consumers, and other decision-makers perceive and use algorithmic input. Some studies document aversion to using algorithmic prediction (Dietvorst et al., 2015, 2018; Lee, 2018; Yeomans et al., 2019; Newman et al., 2020), particularly when algorithms are observed making mistakes (Dietvorst et al., 2015). However, acceptance of algorithmic input is increased when decision-makers have the option to modify the algorithm’s forecast (Di-

---

<sup>3</sup>Specifically, having completed at least 100 MTurk tasks with an approval rating of at least 95%.

etvorst et al., 2018) or are given more information about the functioning of the algorithm (Yeomans et al., 2019). Other studies find algorithmic input is valued more than recommendations from an external human adviser in a range of settings, although decision-makers still prefer their own judgment over an algorithm's, to the detriment of prediction accuracy (Logg et al., 2019; Hoffman et al., 2018), and particularly in the hiring context, workers prefer being judged by a human rather than an algorithm (Dineen et al., 2004). However, this preference may depend on the nature of the task: individuals perceive algorithms as less fair and reliable than humans when human skills are required, and equally reliable when mechanical skills are required (Lee, 2018). Our paper contributes to this literature by studying algorithmic versus human evaluation from the perspective of workers rather than decision-makers in the hiring process. This distinction is important because recruitment is different from other scenarios where algorithmic or human input is chosen: while in previously studied settings, decision-makers were asked to maximize prediction accuracy, it is not necessarily always in the worker's interest to choose the recruiter with the most accurate prediction of their ability. To our knowledge, ours is the first study to elicit incentivized preferences of workers for algorithmic or human evaluation.

The second strand of related literature studies bias in algorithms, either in isolation or compared to human bias (see Cowgill and Tucker, 2020; Köchling and Wehner, 2020, for an overview). This literature is in part motivated by the concern that algorithms could perpetuate or worsen existing biases, through having biased objectives or biased training data (see Obermeyer et al., 2019) but also through financial incentives. For example, female job seekers are found to be presented with fewer ads for high-paying jobs (Datta et al., 2015), and fewer jobs in science, technology, engineering, and math fields (Lambrecht and Tucker, 2019) as a result of advertising algorithms maximizing cost-effectiveness. While this literature investigates the existence of bias in algorithmic (and human) decision-making outcomes, we study to what extent perceived differences in bias between human and algorithmic evaluation determine recruiter preferences for those subject to these decisions. This is related to a recent set of studies which document bias perceptions in different settings (Saxena et al., 2019; Newman et al., 2020; Wang

et al., 2020). Distinct from these studies, we consider a specific hiring scenario and also investigate how perceived differences in bias affect worker choices.

Our work also has relevance for current policy discussions on the use of algorithmic decision-making. The European Commission’s new proposal on regulating the use of AI (European Commission, 2021) is a clear example of this, and the proposal specifically classifies AI for job screening and evaluation as high-risk: “*AI systems used in employment, workers management and access to self-employment, notably for the recruitment and selection of persons, for making decisions on promotion and termination and for task allocation, monitoring or evaluation of persons in work-related contractual relationships, should also be classified as high-risk, since those systems may appreciably impact future career prospects and livelihoods of these persons. [...] Throughout the recruitment process and in the evaluation, promotion, or retention of persons in work-related contractual relationships, such systems may perpetuate historical patterns of discrimination, for example against women, certain age groups, persons with disabilities, or persons of certain racial or ethnic origins or sexual orientation.*” Similarly, various states in the U.S. have (pending) legislation requiring data disclosure for automated decision-making in hiring (State of Illinois, 2021), and for using acquisition methods that minimize the risk of adverse and discriminatory impacts resulting from the design and application of automated decision systems (State of California, 2021; State of Vermont, 2021). Further, the European Union has already included provisions aimed at protecting people from being subjected to fully-automated decision-making under certain conditions in 2016 General Data Protection Regulation (GDPR) legislation.<sup>4</sup> These legislative efforts targeted at protecting those who are algorithmically evaluated underscore that not only the decision-makers but

---

<sup>4</sup>Article 22(1) of the GDPR states that “*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.*” Profiling is defined as “*Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.*” A 2021 lawsuit of workers against the ride-sharing firm Uber (Rechtbank Amsterdam, 2021) cites these GDPR protections, arguing they had been removed from the platform based on automated decisions made by fraud detection software— in this case, however, Uber was able to show humans were meaningfully involved such that there was no fully-automated decision-making.

also those subjected to algorithmic decisions should be studied.

We find that workers do perceive human and algorithmic evaluation differently, even though both recruiters are given the same inputs for evaluation in our controlled setting. Specifically, human recruiters are perceived to process the information available to them in a more biased and error-prone way, and to place more weight on personal characteristics, whereas algorithmic recruiters are seen as more transparent and as placing more weight on task performance. Consistent with these perceptions, workers with good task performance relative to others prefer algorithmic recruiters, whereas those with lower task performance prefer human recruiters. Furthermore, workers with low task performance are willing to pay to obtain their preferred recruiter. In terms of the relative gender bias in recruitment, workers are found to perceive the human recruiter as more strongly favoring men with worse task performance than their competitor. However, unlike for task performance, this perception does not translate in a higher willingness to pay for the preferred recruiter, suggesting that perceived differences in gender bias are not as critical for recruiter choice.

This paper is organized as follows. Section 2 outlines our experimental set-up. Section 3 describes the data and Section 4 presents our results. We discuss our findings in the context of governance of artificial intelligence in Section 5, and Section 6 concludes.

## 2 Recruiter choice experiment

We run two experiments mimicking a hiring setting in which a firm tries to predict job applicants' ability at performing a task and where workers have the possibility to choose their favorite recruiter: algorithmic or human (recruiter choice experiments).<sup>5</sup> Experiments are conducted using oTree (Chen et al., 2016). Specifically, we recruit workers to perform a real-effort number-finding task. In this task, developed by Buser et al. (2020), workers have to find the two unique numbers which add up to 100 out of a  $3 \times 3$  matrix containing nine numbers between 1 and 99. We chose this task because how fast a worker

---

<sup>5</sup>Our experiments have been given clearance by the Ethical Review Committee of Utrecht University's Faculty of Law, Economics, and Governance. Further experimental details can be found in Appendix A.1.

gives the correct answer is a well-defined measure of performance. Further, speed varies across workers but the task is not prohibitively difficult nor time-consuming. Importantly, workers cannot cheat on this task even in a non-monitored environment.<sup>6</sup>

We recruit workers from Amazon’s Mechanical Turk (henceforth MTurk), an online labor market where employers offer tasks to a large pool of potential workers. MTurk was founded in 2005 for crowd-sourcing small labor tasks but has been widely used for social experiments since then (see Paolacci et al., 2010). Compared to laboratory experiments, online labor markets have the important advantage that a large number of diverse workers can be recruited at a reasonable cost. Past research has shown that the outcomes of standard experimental games (ranging from prisoner’s dilemma to interactive multi-period public goods games) do not differ substantially between workers on MTurk and in economic lab settings (Horton et al., 2011; Arechar et al., 2018). However, since MTurk workers (henceforth, workers) can leave the experiment at any time, the literature calls for participation restrictions both before and after the experiment: we outline ours in Section 3.

We run two recruiter choice experiments: an information experiment and a gender bias experiment. Choices in both experiments are incentivized. The former is a piece-rate experiment aimed at uncovering the effect of task performance. By contrast, the latter is a tournament experiment where the worker competes with a fictitious competitor, aimed at uncovering the effect of gender bias.

Each experiment has a different treatment and a different set of incentives. In what follows, we outline the two experimental set-ups, explain how recruiters are designed, and describe the resulting data.

## 2.1 Experimental set-up

The information experiment consists of six phases. In phase one, workers enter demographic information: age, gender, ethnicity, and education level.

---

<sup>6</sup>Niederle and Vesterlund (2007) use a similar task but with adding numbers rather than finding the correct pair: however, without monitoring, workers could cheat by using a calculator.

In phase two, workers are shown the number-finding task, and given an example to familiarize themselves: this is a non-incentivized practice round. We inform workers that there are 10 subsequent rounds of the same task (with different numbers) and that the time in one of these rounds will be selected and sent to the recruiter. Workers are given 90 seconds to play each round, and are informed when their chosen answer is incorrect: they can use any remaining time to find the correct pair of numbers. After completing the 10 rounds, workers are informed that their time in round 9 is selected. We call it recruiter-observed task time (henceforth observed task time). The workers are shown their observed task time (in seconds) as well as their average task time across all rounds.

Our information experiment consists of two related treatments where treated workers receive additional information. In treatment 1a, we inform workers about their observed task time relative to that of past workers. Specifically, we show the median observed time of past workers and tell workers whether they were faster or slower than that. In treatment 1b, in addition to the information provided in treatment 1a, we inform workers about their hypothetical pay-off if it were only based on their observed task time (i.e. ignoring personal characteristics). In both treatments, workers in the control group do not receive any additional information apart from their observed task time and their average task time across all rounds. By comparing treated workers' recruiter choice to that of control group workers, treatments 1a and 1b help us understand to what extent workers choose a particular recruiter based on their task performance.

In phase three, workers are told that a firm wants to hire workers who are good at the number-finding task, and will use recruiters to predict workers' average task performance based on their characteristics (age, gender, ethnicity, and education level) as well as their observed task time. This set of recruiter inputs is chosen to mimic characteristics usually contained on (or inferred from) real-world applicant CVs, combined with a recruitment task which is designed to inform on workers' ability for the job they are being considered for. Two types of recruiters are described: an algorithmic recruiter, and a human recruiter.<sup>7</sup> Workers are also informed that both recruiters are given the same set of inputs,

---

<sup>7</sup>See Appendix A.1 for details on the recruiter descriptions.

but may give different predictions of the worker’s average task time. After this explanation, workers are asked to choose which recruiter they would prefer to be evaluated by. To incentivize this choice, workers are informed that they will receive a payment (up to \$2, piece rate) based on the recruiter’s prediction of their average task time.

In phase four, we assess workers’ willingness to pay for their preferred recruiters. This is achieved by first assigning each worker to the recruiter they did *not* select. Next, workers are given \$0.40 (corresponding to 13%-20% of the maximum task payment) and asked to state their willingness to pay (between \$0 and \$0.40) to obtain their preferred recruiter. Following Becker et al. (1964), workers are informed that an amount between \$0 and \$0.40 is randomly drawn: if the stated willingness to pay is equal to or greater than the drawn amount, workers are able to get their preferred recruiter and must pay the drawn amount, while any amount not spent is theirs to keep. Conversely, if the stated willingness to pay is lower than the drawn amount, the recruiter is unchanged (and workers retain the \$0.40).

In phase five, workers are asked to estimate the average task time each recruiter (i.e. algorithmic and human) predicts for them. To incentivize this phase, one of the estimates is randomly chosen for payment. The more accurate the estimate, the higher the pay-off: payment declines linearly in the mean squared prediction error.<sup>8</sup> This payment is up to \$0.5.

In the sixth and final phase, workers answer several survey questions about their beliefs concerning human and algorithmic recruiters, about their risk preferences (see also Dohmen et al., 2011), and a short version of big five personality traits (Gerlitz and Schupp, 2005). At the end of phase six, workers receive a \$0.1 participation fee.

Phases one, two, four, and six of the gender bias experiment are identical to those of the information experiment; phases three and five differ. For the gender bias experiment, we change our experiment to a tournament set-up. Specifically, in phase three we assign each worker to compete against a fictitious worker, whose education, ethnicity, and age are the same as the worker’s. However, we randomly vary the gender of the competitor as well

---

<sup>8</sup>See Appendix Table A.4 for a detailed payment table.

as their observed task time, in a 2-by-2 design. To make these tournaments close races, we set the observed task time of the competitor to either 1.18 seconds faster or slower than the worker’s. Treated workers are those assigned to a different-gender competitor. After completing the number-finding task, workers are informed about their observed task time as well as the observed task time and personal characteristics of their competitor and ask to choose which recruiter will evaluate both competitors. To incentivize this choice, workers receive up to \$1 based on their predicted average time in the number-finding task and \$2 bonus if they are predicted to be faster than the competitor.<sup>9</sup>

In phase five, workers are asked to estimate the average task time each recruiter (i.e. algorithmic and human) predicts not only for them but also for their competitor. The payment for this phase is up to \$1.

The gender bias experiment allows us to study the effect of competitor gender on recruiter choice. Specifically, by comparing the recruiter choice of workers matched with a different-gender competitor to those with a same-gender competitor, we can study whether workers perceive one type of recruiter to be more gender biased than the other. Note that the difference relative to same-gender competitors nets out any overall impact of the competitive setting on recruiter preference.

The structure of the experiments and their incentive schemes are described in Figure 1. The distribution of the total earnings in the full sample and by experiment is shown in Figure A.8.

## 2.2 Recruiter training

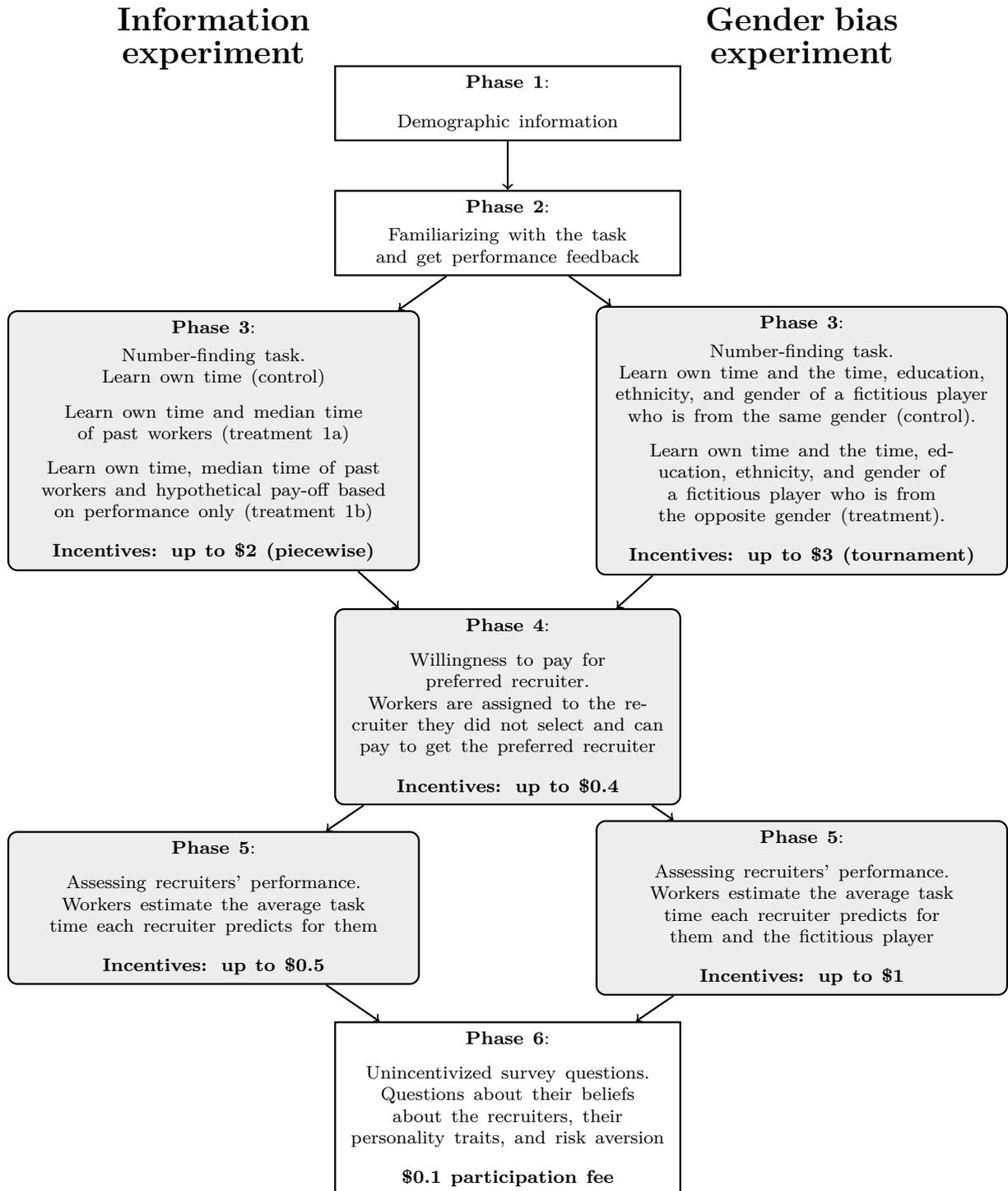
To avoid deception in our experiments, we use actual human and algorithmic recruiters trained on past data. To achieve this, we collected data from two pilot experiments: an MTurk experiment with 345 MTurk workers on April 17 and August 5, 2019, and a lab experiment with 22 participants on June 13, 2019.

The algorithmic recruiter is designed as coefficients from an OLS regression of average

---

<sup>9</sup>This design is adjusted from Niederle and Vesterlund (2007), where the worker who loses the tournament receives no payment. We pay a baseline pay of up to \$1 to ensure a minimum payment to all workers.

Figure 1: Experimental design



Note: gray boxes represent the incentivized phases.

task time on workers' age, gender, ethnicity, education level, and observed task time, across all 345 workers of the MTurk pilot experiment. These coefficients are then used to predict the average task time of workers who obtain the algorithmic recruiter in the information and gender bias experiments.

For the human recruiter, we conducted a lab experiment in the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University. Here, 22 participants were paid to act as recruiters and evaluate the average task time of 83 of the 345 MTurk workers. The human recruiters first performed two rounds of the number-finding task to understand it. Next, each of them individually evaluated either 41 or 42 MTurk workers. All recruiters received 3 euros as a show-up fee and up to 8 euros for their recruitment performance. This performance is defined as the squared distance of their prediction from the actual average task time of the workers they evaluated: after completing the prediction task, one of these predictions was randomly selected for payment. We use an OLS regression with interaction terms for each recruiter to compute 22 sets of coefficients for predicting workers' average task time. One of these sets of coefficients is randomly assigned and used to predict the average task time of workers who obtain the human recruiter in the information and the gender bias experiments.

Both recruiter types assign positive weights to observed task time, and also give non-zero weights to personal characteristics. Both algorithmic and human recruiters tend to predict faster average task time for those with fast observed task time, and to male and younger workers. Specifically, out of 22 human recruiters, 13 place more weight on task performance than does the algorithm, 17 predict a faster average task time for male workers, and 16 predict a faster average task time for younger workers. Human recruiters (20 out of 22) also tend to predict faster average task times for more highly educated workers, while the algorithm does not. Non-whites are predicted to have a faster average task time by the algorithm, while predictions across human recruiters vary widely. In particular, 10 out of 22 human recruiters predict a slower average task time for non-white workers than white workers. Coefficients used for both the algorithmic and human recruiters are reported in Appendix A.1.5. Note that these weights are not our

object of study: we only employ real algorithmic and human recruiters to avoid deceiving participants.

Figure A.5 shows the kernel density of prediction errors for both human and algorithmic recruiters, defined as the worker-level difference between time predicted by the recruiter and the observed average time of the worker in the main experiments. The graph shows that the algorithmic recruiter has a lower variance. This is not surprising since the human recruiter is randomly assigned from a set of 22 recruiters. Both distributions have a mean close to zero but on average, the human scores the workers slightly more favorably than does the algorithm.

### 3 Data description

We impose three ex-ante restrictions on experiment participation (see Mason and Suri, 2012; Goodman et al., 2013; Paolacci and Chandler, 2014): workers have to be registered in the U.S. or Canada, have an internal MTurk approval rating for past tasks of over 95%, and have at least 100 completed tasks.<sup>10</sup> We allow every worker to participate at most once, and exclude workers who did not complete the entire experiment (609 workers in total). We also eliminate a small number of workers who on average take more than 55 out of the allotted 90 seconds to complete the tasks (21 workers in total): these workers likely just let time run out on most tasks, which makes their data uninformative.

In addition to the ex-ante restrictions, we ask two attention check questions in phase three of the information and gender bias experiments: 1) “The algorithm and the human recruiter make the hiring decision based on the same information” (correct answer=yes); 2) “The algorithmic recruiter may assign a different score to you than the human recruiter” (correct answer=yes). All results in Sections 4 and 5 are complemented with a robustness check on the subsample of workers who gave the correct answer to both questions.

Experimental data was collected between April and July 2020, in 17 separate sessions. In total, we have 1,725 valid workers across both recruiter choice experiments. Out of

---

<sup>10</sup>These three measures are also recommended for avoiding bots as participants.

these workers, 1,056 participated in the information experiment (with 681 treated), and 669 participated in the gender bias experiments (with 348 treated). Table 1 shows descriptive statistics on each of these four sets of workers. Male is a dummy variable indicating whether the worker self-identifies as male. Age is a continuous variable constructed using a variable collected in bands by taking the median point of each band (less than 15 years old; 15–19; 20–24; 25–29; 30–34; 35–39; 40–44; 45–49; 50–54; 55–59; 60–64; 65 years or older). Non-white is a dummy variable indicating the worker self-identifies as either: Hispanic or Latino; black or African American; Asian; American Indian or Alaska native; Middle Eastern or North African; Native Hawaiian or other Pacific Island; or Some race or ethnicity other than white. Higher educated is a dummy variable indicating the worker completed at least some college at the undergraduate level (possibilities were: some secondary education; completed secondary education; trade/technical/vocational training; some undergraduate education (college or university); completed undergraduate education; some postgraduate education; completed postgraduate education). Risk propensity is a categorical variable measuring how much the respondent is willing to take risk: this variable ranges from 0 (not at all willing to take risks) to 10 (very willing to take risks). Finally, Conscientiousness, Agreeableness, Neuroticism, Openness, and Extroversion are indices capturing the big-five personality traits, constructed from a set of 15 questions:<sup>11</sup> the minimum score for each trait is 1 and the maximum 15.

Table 1 shows that on average, workers take around 20 to 25 seconds to solve the round-9 task – this performance will be observed by recruiters. Across all ten rounds, they take 18 to 21 seconds on average. Slightly more than half of the workers are male, and around 30% are non-white. Workers’ average age is around 37, and two-thirds have at least some undergraduate education. It is reassuring that the large majority of workers answer each of the attention questions correctly: however, our Appendix contains robustness checks for all specifications where we only retain the subsample of workers who answered both questions correctly. This sample contains 747 out of 1,056 = 71% of all workers for the information experiment and 407 out of 669 = 61% of all workers for the gender bias

---

<sup>11</sup>These questions are outlined in Appendix A.1.6.

experiment.

Overall, the observable characteristics are well-balanced between treated and control groups. However, since there are some minor differences, we also report specifications where we control for all observables, including risk propensity and personality traits. Lastly, our data were collected over a number of sessions taking place on different dates: since worker characteristics can vary by date, we include fixed effects for the date of data collection in all our specifications (even the ones without additional controls). Our conclusions are not affected by this.

Table 1: Worker descriptives

	<i>Sample:</i> Information experiment		Gender bias experiment	
	Treated (1)	Control (2)	Treated (3)	Control (4)
Observed task time (round 9)	21.04 (13.87)	20.27 (14.05)	25.64 (17.66)	25.49 (18.40)
Average task time (all rounds)	18.42 (7.75)	17.43 (8.28 )	21.11 (9.19)	20.76 (9.07)
Male	0.55 (0.50)	0.51 (0.50)	0.57 (0.50)	0.57 (0.50)
Age	37.01 (12.27)	37.17 (11.32)	36.40 (10.45)	36.66 (11.12)
Non-white	0.26 (0.44)	0.29 (0.45)	0.32 (0.47)	0.28 (0.45)
Highly educated	0.64 (0.48)	0.65 (0.48)	0.67 (0.47)	0.67 (0.47)
Risk propensity	5.34 (2.68)	4.99 (2.57)	6.26 (2.69)	6.29 (2.55)
Conscientiousness	11.70 (2.17)	11.87 (2.20)	11.32 (2.08)	11.27 (2.10)
Agreeableness	11.39 (2.22)	11.37 (2.23)	11.00 (2.06)	10.93 (2.13)
Neuroticism	8.63 (3.26)	8.40 (3.30)	8.61 (2.82)	8.60 (2.88)
Openness	11.39 (2.42)	11.31 (2.44)	11.67 (2.17)	11.37 (2.21)
Extraversion	8.46 (3.04)	8.10 (3.23)	9.17 (2.62)	8.89 (2.66)
First attention question correct	0.80 (0.40)	0.86 (0.34)	0.85 (0.36)	0.82 (0.38)
Second attention question correct	0.87 (0.34)	0.88 (0.32)	0.74 (0.44)	0.73 (0.44)
N	681	375	348	321

*Notes:* Means reported, standard deviations in parentheses. Risk propensity can range from 0 (most risk-averse) to 10 (most risk-loving). Big-five personality traits can range from 1 (lowest) to 15 (highest).

## 4 Results

To study how workers perceive algorithmic versus human recruitment, we start by documenting qualitative survey evidence of recruiter perceptions (Section 4.1). Next, we study workers’ recruitment preferences by analyzing recruiter choice and willingness to pay for

one’s preferred recruiter (Section 4.2). Lastly, Section 4.3 presents causal evidence on what drives recruiter choice, based on our two experimental interventions.

## 4.1 Recruiter perceptions

Figure 2 shows how the algorithmic and human recruiter are perceived across nine different dimensions: fairness, discrimination, prediction accuracy (i.e. not being error-prone), transparency, simplicity, familiarity, speed, and as giving importance to workers’ characteristics or as giving importance to workers’ observed task performance. The figure presents results where the five-point Likert scale is collapsed into three categories: disagree, neutral, and agree. This highlights that compared to the human recruiter, the algorithmic recruiter is typically perceived as more fair, more transparent, simpler, faster, and as placing a higher weight on workers’ task performance. On the contrary, the human recruiter is perceived as more discriminatory, more error-prone, more familiar, and as placing a higher weight on workers’ personal characteristics.<sup>12</sup>

These basic survey results suggest both recruiter types are perceived differently by MTurk workers. We now turn to consider to what extent these perceptions are accompanied by different recruiter preferences and the correlates of these choices.

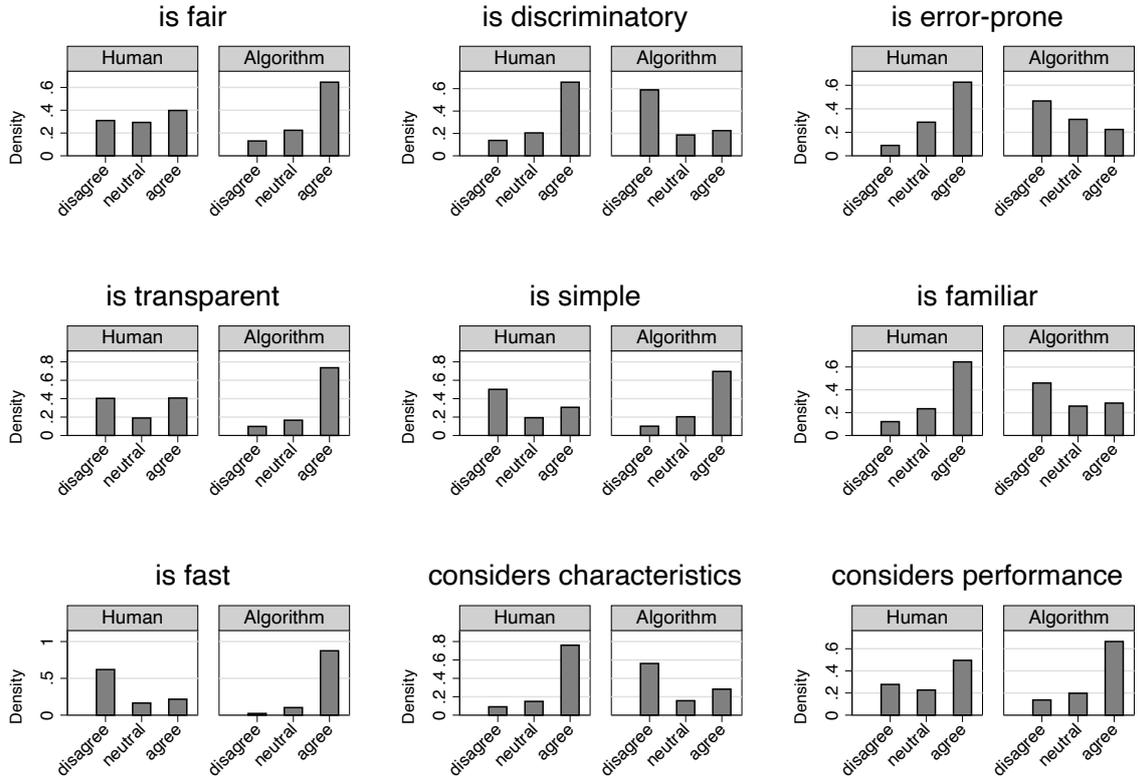
## 4.2 Recruiter preferences

We measure workers’ recruiter preferences in two main ways: their incentivized choice of recruiter, as well as their willingness to pay for this preferred recruiter. Our first finding, shown in the first panel of Figure 3, is that workers do not unanimously prefer one recruiter over the other: around 50% of workers prefer the algorithm, and around 50% prefer the human. Moreover, the second panel shows that close to 60% of all workers are willing to pay to have their preferred recruiters evaluate them. This means that,

---

<sup>12</sup>To avoid framing, we randomize the order in which the statements are presented: this means that some workers are asked to agree or disagree with statements presented as “The human recruiter will give more weight to my personal characteristics than the algorithmic recruiter” (left panel) and others with statements presented as “The algorithmic recruiter will give more weight to my personal characteristics than the human recruiter” (right panel). Although groups who see the human mentioned first tend to choose the neutral option more frequently, results are qualitatively consistent in both groups.

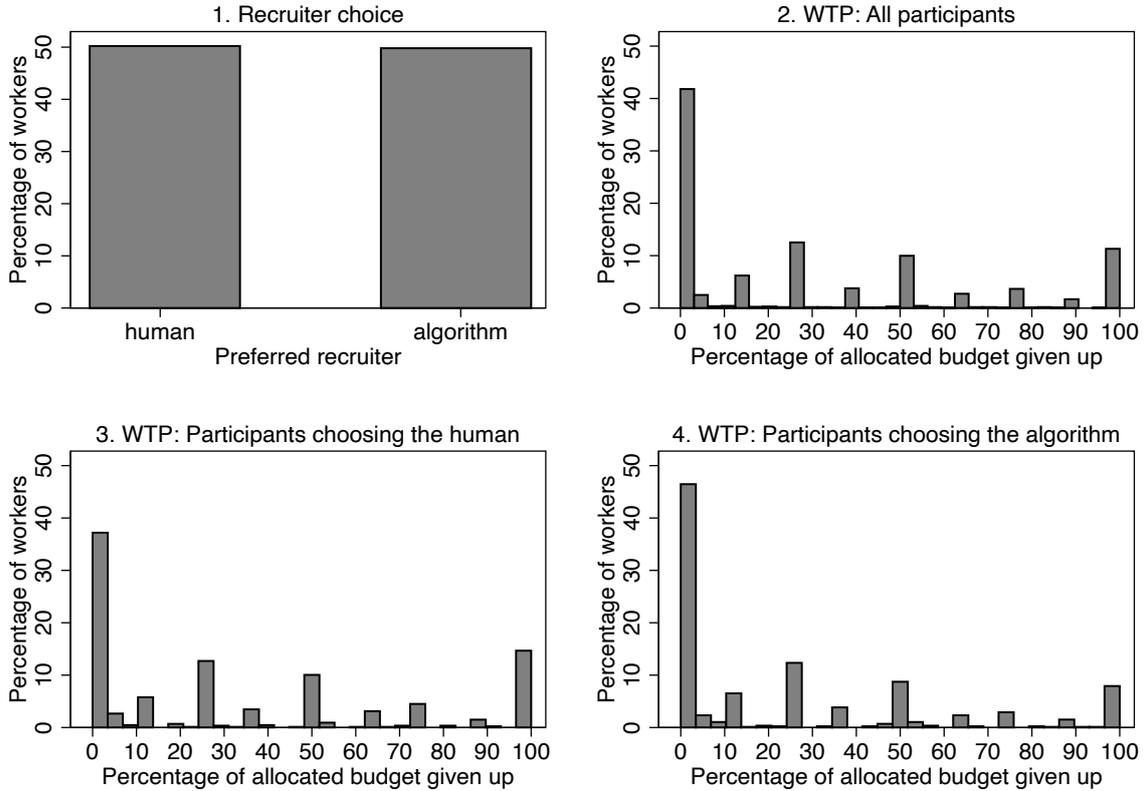
Figure 2: Perception of the recruiter



while some workers are indifferent as measured by their willingness to pay, most are not. On average, workers give up 29% of the allocated recruiter choice budget to obtain their favorite recruiter (this rises to 47% of the budget among those with non-zero willingness to pay). Further, willingness to pay is higher for those who prefer the human recruiter as compared to those who prefer the algorithmic recruiter, as shown in the third and fourth panels of Figure 3. All in all, this highlights that workers have different recruiter preferences, but that those who are assigned an algorithmic recruiter when they prefer the human have larger welfare losses than those who are assigned a human recruiter when they prefer the algorithm. Taken at face value, this suggests that GDPR protections have a place in providing the right to human evaluation for those who prefer it.

Our set-up also allows descriptively studying what factors underlie these preferences. First, we consider to what extent workers choose the recruiter that they expect will rate their performance more favorably: although our design is set up to minimize other factors such as preferences for in-person interaction, we cannot rule out that workers still prefer

Figure 3: Recruiter choice and willingness to pay (WTP)



a specific type of recruiter for other reasons, such as familiarity. Second, we can study the role of observed task performance: to the extent that workers perceive the algorithm to give more weight to performance, and be less error-prone as a predictor, we would expect those with better performance to prefer algorithmic recruitment. Lastly, we can study whether worker characteristics such as gender, age, ethnicity, education level, propensity to take risks, or big-five personality type are correlated with recruiter preferences.

For the descriptive analysis of this section, we focus on the information experiment and consider the full sample of valid workers.<sup>13</sup> Results are presented in Table 2. Columns (1), (2), and (3) show the correlates of choosing the human recruiter, whereas columns (4), (5), and (6) show the correlates of willingness to pay for one's preferred recruiter (as a percentage of the total allocated budget), whether human or algorithm. For each de-

<sup>13</sup>In Appendix Tables A.7 through A.9, we present the same analysis for the gender bias experiment, as well as a robustness check on the subsample of workers who gave correct answers to the two attention checks. Results are very similar.

pendent variable we present three specifications: in the first, reported in columns (1) and (4), the independent variables are two categorical variables equal to one if, respectively, the worker expects the algorithm or the human to rate them more favorably (the omitted category being equal expected ratings); in the second, reported in columns (2) and (5), we add the recruiter-observed task performance (henceforth “observed task performance”); and in the third, reported in columns (3) and (6), we also control for individual worker characteristics.

Table 2 highlights several findings. First, workers do choose the recruiter that they believe will rate them most favorably: when they expect the human to be more favorable, they prefer the human recruiter; and when they expect the algorithm to be more favorable, they prefer the algorithmic recruiter. This suggests our incentive design is successful in eliciting preferences based on expected evaluation scores. This is also evidenced by workers’ willingness to pay for the recruiter that they believe rates them more favorably.<sup>14</sup> However, preferences are asymmetric: our third specification shows that workers who expect the human to be in their favor are 15.6 percentage points more likely to choose the human recruiter, whereas workers who expect the algorithm to be in their favor are only 7.8 percentage points more likely to choose the algorithm. Similarly, preferring the human recruiter is a significant predictor of willingness to pay, even when controlling for all other covariates.

The second key finding is that the observed task performance is a significant predictor of recruiter preference. We define observed task performance as the negative of time spent solving round 9, standardized across all workers of both the information and the gender bias experiment. As such, a one-unit increase in observed task performance indicates the worker solved the round-9 task one standard deviation faster. Table 2, column (3) shows that workers who perform one standard deviation better in the observed task are 7.2 percentage points less likely to choose the human recruiter. Further, those with better observed task performance have a lower willingness to pay for their preferred recruiter

---

<sup>14</sup>Note that we have designed the incentives such that rational workers will choose the recruiter which they *believe* will give them a higher rating. In the absence of any preferences or beliefs, the rational thing to do would be not to pay for one’s chosen recruiter.

(see column (6)). Taken together, this means it is particularly workers with poor observed task performances who prefer being evaluated by a human recruiter. This is consistent with a belief, documented in the previous section, that human recruiters will give less weight to task performance, and are more error-prone predictors. Specifically, less weight on task performance will benefit those with poor performance, and the competitive element inherent in recruiting may suggest to workers that adding noise to the prediction is beneficial to how they compare to others.

The third set of findings concerns worker characteristics: overall, these do not seem particularly important. Their inclusion does not substantially change the magnitude of the coefficients of the categorical variables indicating whether the algorithm or the human is favorable, and only slightly decreases the effect of the observed task performance. Including these characteristics does increase the overall precision and explanatory power of the model. All else equal, male and non-white workers are somewhat less likely to choose the human recruiter, while extroverted workers are more likely to choose the human recruiter. There are no differences in recruiter choice by worker age or education level. Further, risk-loving workers have a higher willingness to pay for their preferred recruiters. This is likely due to the uncertainty inherent in the auction which assigns the final recruiter.<sup>15</sup>

---

<sup>15</sup>Appendix Table A.8 presents results on the subsample of workers who answered the attention check questions correctly: results are very similar to those obtained for the full sample of workers.

Table 2: Correlates of recruiter choice and willingness to pay (WTP)

	<i>Dependent variable:</i>					
	Preference for human recruiter			WTP for preferred recruiter		
	(1)	(2)	(3)	(4)	(5)	(6)
Human favorable	0.183*** (4.09)	0.165*** (3.71)	0.156*** (3.52)	11.82*** (4.11)	10.95*** (3.83)	11.16*** (3.97)
Algorithm favorable	-0.0710* (-1.71)	-0.0795* (-1.93)	-0.0780* (-1.90)	11.60*** (4.36)	10.98*** (4.16)	11.72*** (4.51)
Observed task performance		-0.0816*** (-4.81)	-0.0723*** (-4.19)		-5.053*** (-4.62)	-3.825*** (-3.48)
Male			-0.0740** (-2.36)			-1.019 (-0.51)
Age			0.0000925 (0.07)			0.124 (1.49)
Non-white			-0.0658* (-1.92)			4.032* (1.86)
Highly educated			0.0447 (1.45)			0.0192 (0.01)
Risk propensity			-0.00211 (-0.34)			2.408*** (6.09)
Conscientiousness			-0.00456 (-0.58)			-0.622 (-1.26)
Agreeableness			0.00216 (0.29)			-0.292 (-0.63)
Neuroticism			0.00887* (1.69)			0.737** (2.22)
Openness			0.00972 (1.48)			-0.124 (-0.30)
Extroversion			0.0202*** (3.85)			0.215 (0.64)
Prefer human				5.187*** (2.63)	3.839* (1.94)	3.647* (1.85)
Constant	0.458*** (12.77)	0.477*** (13.37)	0.189 (1.17)	12.69*** (5.16)	14.51*** (5.88)	0.254 (0.02)
N	1,056	1,056	1,056	1,056	1,056	1,056
R <sup>2</sup>	0.065	0.086	0.117	0.053	0.072	0.121
adj. R <sup>2</sup>	0.055	0.074	0.097	0.041	0.059	0.100

*Notes:* Estimates for the information experiment. Willingness to pay is measured as a percentage of the total allocated budget. The variable *Human (/ Algorithm) favorable* is defined as a dummy variable that is equal to one when the worker expects the human (/ algorithm) to rate them more favorably: the omitted category is workers who believe both recruiters will rate them equally. Observed task performance is measured as the negative of the number of seconds the worker took to solve the observed task (i.e. round 9), standardized to have a zero mean and unit standard deviation. All models are estimated with OLS and control for the date of data collection. *t* statistics in parentheses; \* p<0.1 \*\* p<0.05, \*\*\* p<0.01.

### 4.3 Recruiter choice: causal evidence

So far we have shown descriptive, albeit incentivized, evidence of workers’ preferences over human and algorithmic recruitment: we now turn to the results of our two experimental interventions. The treatments are designed to identify the causal effects of the two sets of performance predictors: observed performance in the selected task, and observable characteristics, specifically gender.

#### 4.3.1 Information experiment

In the first experiment, we study the impact of randomly making workers aware of their round-9 task performance compared to the median round-9 performance of past workers: this information treatment allows us to uncover the impact of observed task performance on recruiter preference. Specifically, if workers who are made aware of having a low (high) observed task performance are more likely to select the human (algorithmic) recruiter, the association we uncovered earlier is truly caused by workers’ expectations of recruiter evaluations, and not driven by the performance difference per se (or by omitted variables).

To study this, we compare treated workers to control group workers of the same observed performance level: the only difference between these groups is that the former has been made aware of their observed performance relative to the median. Our estimating equation is:

$$y_i = \alpha + \beta \times treat_i + \psi_i + \varepsilon_i \quad (1)$$

where  $i$  indexes individual workers. The dependent variable  $y_i$  is either a dummy for preferring the human recruiter or willingness to pay for one’s preferred recruiter as a percentage of the total budget allocated to this part.  $treat_i$  is a dummy equal to one when workers have been informed about their observed performance relative to the median of past workers. We estimate equation 1 separately for those with observed performance below and above the median since the information signal differs for them: the former know they are low observed task performers whereas the latter know they are high observed task performers. We also estimate this model with a set of controls contained in  $\psi_i$ . These are the observed task performance (as before, measured as the negative of the number of

seconds needed to solve the round-9 task), dummies for whether the worker expects the human or algorithmic recruiter to assign them a better score (the omitted category being equal expected scores across recruiters), worker characteristics (gender, age, ethnicity, and education level), big-five personality traits and risk propensity, as well as fixed effects for the date of data collection. In the willingness to pay equation, we also control for a dummy variable equal to one if the worker prefers the human recruiter.

Panel A of Table 3 confirms that workers who are made aware that their observed task performance is low are indeed more likely to select the human recruiter compared to a control group of low performers not made aware of this. Panel B shows those being made aware of being a low performer are also more willing to pay for their preferred recruiter. This confirms a causal link between low observed performance and preference for human recruitment. However, we again find some asymmetry: workers who are made aware of having a high observed task performance do not substantially alter their recruiter choice, and have a smaller increase in their willingness to pay as a result of this information. These results are consistent with workers perceiving the human recruiter to place less weight on task performance since this is more likely to benefit those not able to send a positive performance signal.<sup>16</sup>

---

<sup>16</sup>In our Appendix, we present two robustness checks on these findings: first we estimate the model on a subsample of workers who gave the correct answer to both attention check questions (see Table A.10); second, we estimate the model for information treatments 1a and 1b separately (see Table A.11). The robustness checks confirm the results from our main specification.

Table 3: The impact of information on task performance on recruiter preference and willingness to pay

<i>A. Dependent variable: Preference for human recruiter</i>		
	(1)	(2)
Treated, low observed task performance	0.178*** (3.30)	0.140*** (2.63)
N	595	595
<i>B. Dependent variable: Willingness to pay for preferred recruiter</i>		
	(5)	(6)
Treated, low observed task performance	9.866*** (2.70)	7.768** (2.16)
N	595	595
	(7)	(8)
Treated, high observed task performance	3.141 (0.91)	2.854 (0.84)
N	461	461
Additional controls	No	Yes

*Notes:* Estimates for the information experiment. Treated workers who were made aware they had a low observed task performance are compared to control group workers with low observed task performance. Treated workers who were made aware they had a high observed task performance are compared to control group workers with high observed task performance. All models are estimated with OLS and control for the date of data collection. Additional controls are  $\psi_i$  from equation 1.  $t$  statistics in parentheses; \*  $p < 0.1$  \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

### 4.3.2 Gender bias experiment

In the second experiment, we assign workers to compete with a fictitious competitor: the worker who obtains the best predicted performance wins the tournament. The fictitious competitor is randomly presented as having either the same gender as the worker (control group) or not (treatment group); and differs only slightly in terms of observed task performance (1.18 seconds faster or slower). All the other characteristics are the same as the worker's. By comparing recruiter choice between men and women for workers paired with same-gender competitors to that of workers paired with different-gender competitors,

we can uncover any perceived differences in gender discrimination across the recruitment methods.

Specifically, we use a Differences-in-Differences framework, allowing us to study whether being paired with a different-gender competitor has a different effect for men and women and the sign of the difference. We modify equation (1) as follows:

$$y_i = \alpha + \beta_1 \times \text{treat}_i + \beta_2 \times \text{male}_i + \beta_3 \times \text{treat}_i \times \text{male}_i + \psi_i + \varepsilon_i \quad (2)$$

where  $\text{treat}_i$  is a dummy variable equal to one if worker  $i$  has been treated, i.e., if  $i$  has been paired with a different-gender competitor, and  $\text{male}_i$  is a dummy variable equal to one if  $i$  is a male. Therefore,  $\beta_3$  captures whether the treatment differently affects men and women: it is equal to the effect that being paired with a different-gender competitor has on men minus the effect that being paired with a different-gender competitor has on women. We also estimate this model with a set of controls contained in  $\psi_i$ . These are dummies for whether the worker expects the human or algorithmic recruiter to assign them a better score than the competitor (the omitted category being equal expected scores across recruiters), worker characteristics (gender, age, ethnicity, and education level), big-five personality traits and risk propensity, as well as fixed effects for the date of data collection.

We estimate equation (2) separately for workers paired with a (slightly) slower or a (slightly) faster competitor. Even though these performance differences are only 1.18 seconds in either direction, our previous analysis suggests that observed task performance is a strong predictor of recruiter choice. When using preference for human recruitment as the dependent variable, we would expect  $\beta_3 > 0$  if human recruiters are perceived to be more biased in favor of men than algorithmic recruiters. Conversely, we would expect  $\beta_3 < 0$  if human recruiters are perceived to be less biased in favor of men than algorithmic recruiters. For differences in perceived gender bias to be important for recruiter choice, we would expect  $\beta_3 \neq 0$  whenever we use willingness to pay for one’s favored recruiter is the dependent variable.

Results are reported in Table 4. This highlights only one possible dimension of per-

ceived relative recruiter gender bias: workers perceive human recruiters to be more favorable to men who are slower than their competitors. Specifically, the coefficients in columns (3) and (4) suggest that male workers who are slower than their female competitors are 21 to 24 percentage points more likely to choose human recruitment as compared to female workers who are slower than their male competitors. However, perceived bias is not particularly pervasive. For one, no differences are found for those competing against faster competitors (columns (1) and (2)), suggesting there is no overall perceived bias against women in human relative to algorithmic recruitment. More importantly, columns (5) through (8) show workers are not differentially willing to pay to obtain their favored recruiter.<sup>17</sup> The lack of willingness to pay suggests differences in gender bias are not perceived as a first-order concern for recruiter choice—this is in stark contrast to the results for observed task performance uncovered previously. Besides being statistically insignificant, the sign on willingness to pay is actually negative rather than positive for the subsample where perceived gender bias was found: those slower at the task than their different-gender competitors. All in all, we interpret the results from this experiment as providing only weak evidence of perceived differences in gender bias across human and algorithmic recruitment in our setting.<sup>18</sup>

---

<sup>17</sup>We find the same when looking at willingness to pay for both men and women when paired with a different-gender competitor, i.e. only using a single difference rather than differences-in-differences estimator.

<sup>18</sup>Results of the robustness check on the subsample of workers with correct answers to both attention checks are presented in Appendix Table A.13 and lead to the same conclusions.

Table 4: Perceived relative recruiter gender bias, by task performance

<i>A. Dependent variable: Preference for human recruiter</i>				
	Faster than competitor		Slower than competitor	
	(1)	(2)	(3)	(4)
Treated	-0.0123 (-0.11)	-0.0601 (-0.55)	0.208* (1.94)	0.244** (2.27)
N	329	329	340	340
<i>B. Dependent variable: Willingness to pay</i>				
	Faster than competitor		Slower than competitor	
	(5)	(6)	(7)	(8)
Treated	5.346 (0.63)	2.295 (0.28)	-5.147 (-0.64)	-3.789 (-0.49)
N	329	329	340	340
Additional controls	No	Yes	No	Yes

*Notes:* Estimates for the gender bias experiment. Treated workers are those paired with a different-gender competitor. All models are estimated with OLS and control for the date of data collection. Additional controls are  $\psi_i$  from equation 2.  $t$  statistics in parentheses; \*  $p < 0.1$  \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

## 5 Discussion

Although our results do not necessarily generalize to a full hiring process, the setting we mimic is closely related to a first stage of CV screening combined with some standardized job or aptitude test. Further, our controlled set-up allows us to uncover causal evidence and largely abstract from factors that are unrelated to differences in evaluation (such as general preferences for human interaction).

Our evidence suggests a number of recommendations for the governance of Artificial Intelligence (AI). For one, the current policy discussion with respect to evaluation is largely focused on bias: however, our findings suggest prediction accuracy is of foremost importance from the perspective of workers being evaluated.<sup>19</sup> While increased prediction accuracy could improve the match quality between workers and jobs, our results suggest it could also be perceived to give applicants less of a chance based on their previous

<sup>19</sup>If anything, our results suggest human recruiters are perceived as more biased. This is consistent with a literature finding that algorithms can attenuate biases found in human decision-making (e.g., see Kleinberg et al., 2017).

performance (e.g. as reflected on their CV), or their performance in a job test. In our experiments, all workers can signal their performance equally (albeit noisily), but in the real world, this need not be the case. For example, workers with recent unemployment spells or skills that are not formally credentialed may not be given a chance, potentially leading to further entrenchment of labor market disadvantages. Further, there are many cases where job performance is difficult to predict *ex ante*, or depends on factors not easily measured in individual tests, such as social skills (Deming and Weidmann, 2020).

Within the context of our experiments, we can provide preliminary evidence for this mechanism by studying workers’ recruiter choice as a function of the quality of their performance signal. Specifically, we estimate the following model:

$$\text{human}_i = \alpha + \beta_1 \times \text{bad signal}_i + \zeta_i + u_i \quad (3)$$

where  $i$  indexes individual workers. The dependent variable  $\text{human}_i$  is a dummy for preferring the human recruiter. The variable  $\text{bad signal}$  is the standardized difference between the number of seconds the worker took to solve the round-9 task and the number of seconds the worker took to solve all observed tasks on average. Average task time here captures workers’ actual performance in this task, which is unobserved to recruiters but which they try to predict; whereas performance in round 9 is the performance signal observed by the recruiter. Workers who have a higher value for this difference are sending a bad signal, as the time they spent to solve the observed task is higher than their actual average task time. We also estimate this model with a set of controls contained in  $\zeta_i$ . These are worker characteristics (gender, age, ethnicity, and education level), big-five personality traits and risk propensity, as well as fixed effects for the date of data collection.<sup>20</sup> Results in Table 5 highlight that workers who send a worse performance signal have a stronger preference for human recruitment.<sup>21</sup> This suggests the ability to

---

<sup>20</sup>We limit our analysis to the information experiment, where the pay-offs from recruiter choice depend only on the worker and not also on the fictitious competitor, for whom we do not have an average task ability.

<sup>21</sup>Appendix table A.15 reports the results for the subsample of workers who passed the attention checks. Results are qualitatively identical. In a separate analysis of the monetary pay-off of recruiter choice

signal job performance could be a determinant of worker preferences for algorithmic versus human recruitment.

These findings are an argument in favor of using algorithms not only to select successful candidates but also to increase exploration in hiring, as has recently been proposed (Li et al., 2020). Specifically, Li et al. (2020) show that exploration can be used in hiring settings by designing an algorithm which gives a premium to learning about underrepresented groups. Our results suggest that it may be important to encourage algorithmic exploration for groups with different abilities to signal task performance. These considerations could become even more important in a setting where most training data is generated by algorithmic decisions: given the lower prediction noise in such data, fewer candidates with poor observed task performance will be hired, resulting in further entrenchment. Our findings suggest that more research is needed to design algorithms that are not only focused on selecting the best candidates based on a pool of previous successes, but which gives applicants a chance to prove themselves in the job. To our knowledge, these considerations have not yet featured in recent legislative proposals.

A final policy-relevant finding is that while the GDPR is focused on avoiding fully automated decision-making, particularly in high-risk cases such as hiring, our results suggest many workers would prefer to be evaluated by an algorithm when given the choice, at least in an initial screening phase. This highlights that for the governance of AI in labor markets, any issues with AI decision-making should be compared to the counterfactual of human decision-making, rather than be considered in isolation. However, our finding that workers are more willing to pay for human than algorithmic recruitment does provide a basis for placing emphasis on the right to be evaluated by a human, as the GDPR currently does.

As with any experiment, our findings are limited to the specific setting we examine. Several dimensions seem most relevant here. First, the MTurk population is one very

---

reported in Appendix Table A.12, we also find that workers who send a worse performance signal and choose human recruitment gain from their choice, whereas those who send a worse signal but choose algorithmic recruitment lose. On average, workers make the right choice: that is, they choose human recruitment when their signal is bad. Appendix Table A.14 shows these findings are robust to restricting the sample to workers who answered both attention check questions correctly.

familiar with online environments, and therefore perhaps also more amenable to algorithmic evaluation than the public at large. If this is the case, we may understate overall preferences and willingness to pay for human evaluation: however, even in this particular population, we find significant willingness to pay based on workers' expectations of the recruiter's evaluation score.

Second, our finding that perceived relative gender bias is not a first-order concern for recruiter choice could be context-specific for a number of reasons. Firstly, our number-finding task and the target of finding average ability at number-finding may be relatively gender-neutral as compared to the real-world labor market. On the one hand, our math-based task is likely more male- than female-stereotyped, and both algorithmic and human recruiters predict a lower ability for women. However, a setting more high-powered to detect gender bias could be to directly contrast female- to male-stereotyped tasks, as in Sarsons et al. (2020). Comparing outcomes for different tasks in our hiring setting could be an interesting extension of our work. Further, expected gender bias in human recruitment may play out more through in-person interactions than in the remote setting provided here. This means our setting could artificially lower perceived differences in gender bias between recruiters.<sup>22</sup> And lastly, our setting has not considered other potential dimensions of bias, such as by ethnicity: however, the gender bias experiment we designed can be easily modified to study this.

Third, preferences for the use of algorithms in hiring settings may depend on the nature of the job: evidence from non-labor market settings suggests that tasks where human qualities are important are perceived to be better judged by humans than machines (Lee, 2018; Gruber et al., 2020). Our numerical task could clearly be performed well by machines: this implies the preference and willingness to pay for human recruitment we find could be higher for tasks requiring soft and/or social skills. This could be studied in our setting by replacing the numerical task with one involving such skills or allowing big-five personality characteristics to enter into the recruitment decision. However, our

---

<sup>22</sup>It should also be noted that our experiment informs only on relative bias: it does not rule out that both recruiters are perceived to be equally gender biased. However, relative bias is the appropriate metric for recruiter choice.

Table 5: Performance signals and recruiter choice

*Dependent variable: Preference for human recruiter*

	(1)	(2)
Bad signal	0.0507*** (2.97)	0.0442*** (2.61)
N	1,056	1,056
R <sup>2</sup>	0.024	0.064
adj. R <sup>2</sup>	0.013	0.045
Additional controls	No	Yes

*Notes:* Estimates for the information experiment. Bad signal is the difference between the number of seconds the worker took to solve round 9 and the number of seconds the worker took to solve all observed tasks on average; standardized to have a zero mean and unit standard deviation. All models are estimated with OLS and control for the date of data collection. Additional controls are gender, age, race, education, personality characteristics and risk aversion.  $t$  statistics in parentheses; \*  $p < 0.1$  \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

recruitment process does already give a role to human judgment by allowing characteristics other than task performance as an evaluation input.

## 6 Conclusion

We mimic a hiring setting on the MTurk platform to elicit workers' preferences for algorithmic or human evaluation in recruitment. We find that individuals do perceive human and algorithmic evaluation differently, and are willing to pay for their preferred recruiter even though both recruiters are given the same inputs in our controlled setting. Specifically, human recruiters are perceived to be more error-prone and biased evaluators, and place more weight on personal characteristics, whereas algorithmic recruiters are seen as more transparent and placing more weight on task performance. Consistent with these perceptions, workers with good observed task performance relative to others prefer algorithmic evaluation, whereas those with lower observed task performance prefer human evaluation. After all, when observed task performance is low, less weight on performance is beneficial, and noisier prediction improves the chance of being ranked above one's labor market competitor(s). However, we uncover only limited evidence that perceived differences in gender bias between both recruiters matter for preferences over human and

algorithmic recruitment.

Overall, these results suggests that widespread adoption of algorithmic recruitment would reduce some workers' welfare, even while improving prediction accuracy. Our findings also highlight that more emphasis should be placed on prediction accuracy as opposed to bias for understanding workers' perception of algorithmic evaluation. Further, workers who send a worse signal relative to their actual task ability have the strongest preference for human recruitment. Therefore, the quality of performance measurement, including but not limited to signal quality, can have an important effect on workers' preference for human recruitment. This finding suggests a role for designing algorithms which purposefully explore candidates with noisier performance signals (Li et al., 2020).

Since algorithmic evaluation from the perspective of workers is only beginning to be studied, we see many opportunities for follow-up research. An important next step would be to consider these issues in a field experiment. Further, while we have focused on evaluation differences for driving preferences over human and algorithmic recruitment, future studies could consider other dimensions of interest such as data privacy concerns or (dis)amenities from human interaction or judgment.

## References

- Ajunwa, I. and Greene, D. (2019). Platforms at Work: Automated Hiring Platforms and Other New Intermediaries in the Organization of Work. In *Work and Labor in the Digital Age*, pages 61–91. Emerald Publishing Limited.
- Arechar, A. A., Gächter, S., and Molleman, L. (2018). Conducting Interactive Experiments Online. *Experimental Economics*, 21(1):99–131.
- Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring Utility by a Single-Response Sequential Method. *Behavioral science*, 9(3):226–232.
- Belot, M., Kircher, P., and Muller, P. (2018). Providing Advice to Jobseekers at Low Cost: An Experimental Study on Online Advice. *The Review of Economic Studies*, 86(4):1411–1447.
- Buser, T., Niederle, M., and Oosterbeek, H. (2020). *Can Competitiveness Predict Education and Labor Market Outcomes? Evidence from Incentivized Choice and Survey Measures*. Working Paper.
- Carey, D. and Smith, M. (2016). *How companies are using simulations, competitions, and analytics to hire*. <https://hbr.org/2016/04/how-companies-are-using-simulations-competitions-and-analytics-to-hire>.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.
- Cowgill, B. (2020). Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening. *Working Paper, Columbia Business School*.
- Cowgill, B. and Tucker, C. (2020). *Algorithmic Fairness and Economics*. <https://ssrn.com/abstract=3361280>.

- Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on privacy enhancing technologies*, 2015(1):92–112.
- Deming, D. and Weidmann, B. (2020). *Team Players: How Social Skills Improve Group Performance*. Working Paper.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.
- Dineen, B. R., Noe, R. A., and Wang, C. (2004). Perceived fairness of web-based applicant screening procedures: Weighing the rules of justice and the role of individual differences. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, 43(2-3):127–145.
- Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.
- European Commission (2021). Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act). <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence>.
- Gerlitz, J.-Y. and Schupp, J. (2005). Zur erhebung der big-five-basierten persönlichkeitsmerkmale im soep. *DIW Research Notes*, 4:2005.
- Goodman, J. K., Cryder, C. E., and Cheema, A. (2013). Data Collection in a Flat World:

- The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, 26(3):213–224.
- Goos, M., Rademakers, E., Salomons, A., and Willekens, B. (2019). Markets for Jobs and Their Task Overlap. *Labour Economics*, 61.
- Gruber, J., Handel, B. R., Kina, S. H., and Kolstad, J. T. (2020). *Managing Intelligence: Skilled Experts and AI in Markets for Complex Products*. NBER Working Paper 27038.
- Hoffman, M., Kahn, L. B., and Li, D. (2018). Discretion in Hiring. *The Quarterly Journal of Economics*, 133(2):765–800.
- Horton, J. J. (2017). The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment. *Journal of Labor Economics*, 35(2):345–385.
- Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The Online Laboratory: Conducting Experiments in a Real Labor Market. *Experimental Economics*, 14(3):399–425.
- IBM (2016). *Watson Analytics Use Case for HR: Retaining Valuable Employees*. <https://www.ibm.com/blogs/business-analytics/watson-analytics-use-case-for-hr-retaining-valuable-employees/>.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.
- Köchling, A. and Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development. *Business Research*, pages 1–54.
- Lambrecht, A. and Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, 65(7):2966–2981.

- Lee, M. K. (2018). Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society*, 5(1):2053951718756684.
- Li, D., Raymond, L., and Bergman, P. (2020). *Hiring as Exploration*. SSRN Working Paper 3630630.
- Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.
- Mason, W. and Suri, S. (2012). Conducting Behavioral Research on Amazon’s Mechanical Turk. *Behavior Research Methods*, 44(1):1–23.
- Newman, D. T., Fast, N. J., and Harmon, D. J. (2020). When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160:149–167.
- Niederle, M. and Vesterlund, L. (2007). Do Women Shy away from Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics*, 122(3):1067–1101.
- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464):447–453.
- Paolacci, G. and Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23(3):184–188.
- Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision making*, 5(5):411–419.
- Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481.
- Rechtbank Amsterdam (2021). Case C/13/692003 / HA RK 20-302. <http://deelink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBAMS:2021:1018>.

- Sarsons, H., Gerxhani, K., Reuben, E., and Schram, A. (2020). Gender Differences in Recognition for Group Work. *Journal of Political Economy*, Forthcoming.
- Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., and Liu, Y. (2019). How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106.
- State of California (2021). Act 2021 CA A 13. [https://custom.statenet.com/public/resources.cgi?id=ID:bill:CA2021000A13&ciq=ncsl&client\\_md=d0b9abc5423ba59c83de196a4321a141&mode=current\\_text](https://custom.statenet.com/public/resources.cgi?id=ID:bill:CA2021000A13&ciq=ncsl&client_md=d0b9abc5423ba59c83de196a4321a141&mode=current_text).
- State of Illinois (2021). Act 2021 IL H 53. [https://custom.statenet.com/public/resources.cgi?id=ID:bill:IL2021000H53&ciq=ncsl&client\\_md=cf812e17e7ae023eba694938c9628eea&mode=current\\_text](https://custom.statenet.com/public/resources.cgi?id=ID:bill:IL2021000H53&ciq=ncsl&client_md=cf812e17e7ae023eba694938c9628eea&mode=current_text).
- State of Vermont (2021). Act 2021 VT H 263. [https://custom.statenet.com/public/resources.cgi?id=ID:bill:VT2021000H263&ciq=ncsl&client\\_md=7fcb043c609beb468b49a53ca7e6dee1&mode=current\\_text](https://custom.statenet.com/public/resources.cgi?id=ID:bill:VT2021000H263&ciq=ncsl&client_md=7fcb043c609beb468b49a53ca7e6dee1&mode=current_text).
- Wang, R., Harper, F. M., and Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.
- WSJ (2012). *Your Resume Vs. Oblivion*. The Wall Street Journal. <https://www.wsj.com/articles/SB10001424052970204624204577178941034941330>.
- Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. (2019). Making Sense of Recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414.

# A Appendices

For online publication only.

## A.1 Recruiter choice experiment: experimental instructions

### A.1.1 Information experiment-control

#### Instructions

This Human Intelligence Task (HIT) is an academic experiment on decision-making. It will require you to perform a few simple tasks. Your payment will be based on your performance. We also ask some survey questions, both before and after the tasks. All data will be stored anonymously. The entire HIT takes around 20 minutes to complete, and payment is conditional on full completion of the tasks and survey: please only accept this HIT if you can commit to completing it.

After completing this HIT, you will earn 0.1 US Dollar for participating, and up to 3.0 US Dollars in total depending on your performance.

After you have accepted this HIT, the URL to the study will appear here: [link](#).

On the last page, you will be given a completion code. Please copy/paste that code below.

#### Number-finding task

Next, you will see a matrix with 9 numbers. You need to find the two numbers that sum to 100. You have a maximum time of 90 seconds. Your payment depends on how fast you find the two correct numbers. If you submit the wrong numbers, an error message will appear and you have to try again until time runs out and the next exercise appears. You repeat this task for 10 rounds. In the end, one of the 10 rounds will be chosen for payment.

Your payment is determined by how fast you find the two correct numbers. Every second is one point and every 90 points are equivalent to 1 US Dollar. On the next page, you are given one training round, which does not affect your payment in any way.

## Training round

Figure A.1: Training round

Time left to complete this page: 1:08

Write in the fields the two numbers that sum to 100. Do it as fast as possible. If you don't write down the correct two numbers, an error message will appear and you have to try again.

87	69	71
4	66	79
21	68	97

First number:

Second number:

[Next](#)

## Number-finding task result

Round 9 was chosen for your payment.

Here is your time in round 9:

Round	Your time
9	3.85

## Hiring

A firm needs people to perform this number-finding task and wants to hire those who are fastest. The firm doesn't know your skill at this task and needs to predict it in one of the two following ways: 1) using a **human recruiter**; 2) using an **algorithmic recruiter**.

Think of the **algorithmic recruiter** as a tool that can poll hundreds of people and determine how many seconds each person needs to solve the task on average. This way, the algorithm can learn which characteristics lead to a specific average time. The algorithm

was trained on past data and will compare your characteristics to the characteristics of past participants. Thus, it will calculate your average time based on how similar participants performed in the past.

Think of the **human recruiter** as a person from a group of 20 people who have evaluated several similar participants in the past. We randomly assign one of these human recruiters to you.

Summarizing, both the human recruiter and the algorithmic recruiter are given the same information about you. This information is:

1. your time in round 9, and
2. your personal characteristics: gender, age, education, and ethnicity.

You can earn up to \$2 in this part.

Please answer the following questions to make sure you understood the explanation above:

1. The algorithm and human recruiter make the hiring decision based on the same information.

*False*    *True*

2. The algorithmic recruiter may assign a different score to you than the human recruiter.

*False*    *True*

### Recruiter choice

Now, choose the recruiter that you would prefer to be evaluated by.

Human recruiter	Algorithmic recruiter
-----------------	-----------------------

### Willingness to pay for the recruiter

You indicated a preference for a **human recruiter** but you were assigned an **algorithmic recruiter**, instead.

Now you are endowed with 40 cents and asked to choose the portion of this amount (between 0 and 40 cents, inclusive) that you wish to pay to change your assigned recruiter to your preferred option. Then the computer randomly chooses a number between 0

and 40. If your number is higher than the random number, your assigned recruiter will change. Those cents not invested are yours to keep. If your number is lower than the random number, you assigned recruiter will not change and all 40 cents are yours to keep. For your convenience, these instructions will remain available to you on the next page.

**Prediction task instruction**

On the next page, you will be asked to guess what average time you think each recruiter (i.e. algorithmic and human) predicted for you. After you have answered all questions, one of the answers will be randomly chosen for payment. Your payoff is determined by how close you are to the actual predictions of either the algorithm or the human recruiter. You will be paid according to the table below:

Table A.1: Prediction task payoff

Difference in seconds*	0	1	2	3	4	5	6	7	8	9	$\geq$ 10
Payoff in \$	0.5	0.49	0.48	0.45	0.42	0.37	0.32	0.25	0.18	0.09	0

\* Absolute difference between your guess and the actual average time assigned to you by each recruiter.

Please answer the following questions to make sure you understood the explanation above:

1. You always earn the same amount of money, no matter how close you are to the actual prediction.

*oFalse oTrue*

2. If you are very close to the actual predictions, you can earn up to 1\$.

*oFalse oTrue*

**Prediction task**

Now, please state how many seconds do you think each recruiter will predict for your average task performance. One of the two answers will be randomly chosen for payment.

1. What average time (in seconds) do you think the algorithmic recruiter predicts for you?

2. What average time (in seconds) do you think the human recruiter predicts for you?

### **A.1.2 Information experiment-treatment 1a**

#### **Instructions**

This Human Intelligence Task (HIT) is an academic experiment on decision-making. It will require you to perform a few simple tasks. Your payment will be based on your performance. We also ask some survey questions, both before and after the tasks. All data will be stored anonymously. The entire HIT takes around 20 minutes to complete, and payment is conditional on full completion of the tasks and survey: please only accept this HIT if you can commit to completing it.

After completing this HIT, you will earn 0.1 US Dollar for participating, and up to 3.0 US Dollars in total depending on your performance.

After you have accepted this HIT, the URL to the study will appear here: [link](#).

On the last page, you will be given a completion code. Please copy/paste that code below.

#### **Number-finding task**

Next, you will see a matrix with 9 numbers. You need to find the two numbers that sum to 100. You have a maximum time of 90 seconds. Your payment depends on how fast you find the two correct numbers. If you submit the wrong numbers, an error message will appear and you have to try again until time runs out and the next exercise appears. You repeat this task for 10 rounds. In the end, one of the 10 rounds will be chosen for payment.

Your payment is determined by how fast you find the two correct numbers. Every second is one point and every 90 points are equivalent to 1 US Dollar. On the next page, you are given one training round, which does not affect your payment in any way.

#### **Training round**

## Figure A.2: Training round

Time left to complete this page: 1:08

Write in the fields the two numbers that sum to 100. Do it as fast as possible. If you don't write down the correct two numbers, an error message will appear and you have to try again.

87	69	71
4	66	79
21	68	97

First number:

Second number:

Next

## Number-finding task result

Round 9 was chosen for your payment.

Here is your time in round 9:

Round	Your time
9	3.85

Here is the time of an average participant in round 9:

Round	Average participant's time
9	15.4

This means you were faster than an average participant in round 9.

## Hiring

A firm needs people to perform this number-finding task and wants to hire those who are fastest. The firm doesn't know your skill and needs to predict them in one of the two following ways: 1) using a **human recruiter**; 2) using an **algorithmic recruiter**.

Think of the **algorithmic recruiter** as a tool that can poll hundreds of people and determine how many seconds each person needs to solve the task on average. This way, the algorithm can learn which characteristics lead to a specific average time. The algorithm was trained on past data and will compare your characteristics to the characteristics of past participants. Thus, it will calculate your average time based on how similar participants performed in the past.

Think of the **human recruiter** as a person from a group of 20 people who have evaluated several similar participants in the past. We randomly assign one of these human recruiters to you.

Summarizing, both the human recruiter and the algorithmic recruiter are given the same information about you. This information is:

1. your time in round 9, and
2. your personal characteristics: gender, age, education, and ethnicity.

You can earn up to \$2 in this part.

Please answer the following questions to make sure you understood the explanation above:

1. The algorithm and human recruiter make the hiring decision based on the same information.

*False*    *True*

2. The algorithmic recruiter may assign a different score to you than the human recruiter.

*False*    *True*

### Recruiter choice

Now, choose the recruiter that you would prefer to be evaluated by.

Human recruiter

Algorithmic recruiter

### Willingness to pay for the recruiter

You indicated a preference for a **human recruiter** but you were assigned an **algorithmic recruiter**, instead.

Now you are endowed with 40 cents and asked to choose the portion of this amount (between 0 and 40 cents, inclusive) that you wish to pay to change your assigned recruiter to your preferred option. Then the computer randomly chooses a number between 0 and 40. If your number is higher than the random number, your assigned recruiter will change. Those cents not invested are yours to keep. If your number is lower than the random number, you assigned recruiter will not change and all 40 cents are yours to keep. For your convenience, these instructions will remain available to you on the next page.

**Prediction task instruction**

On the next page, you will be asked to guess what average time you think each recruiter (i.e. algorithmic and human) predicted for you and the other participant. After you have answered all questions, one of the answers will be randomly chosen for payment. Your payoff is determined by how close you are to the actual predictions of either the algorithm or the human recruiter. You will be paid according to the table below:

Table A.2: Prediction task payoff

Difference in seconds*	0	1	2	3	4	5	6	7	8	9	$\geq 10$
Payoff in \$	0.5	0.49	0.48	0.45	0.42	0.37	0.32	0.25	0.18	0.09	0

\* Absolute difference between your guess and the actual average time assigned to you by each recruiter.

**Prediction task**

Now, please state how many seconds do you think each recruiter will predict for your average task performance. One of the two answers will be randomly chosen for payment.

1. What average time (in seconds) do you think the algorithmic recruiter predicts for you?

2. What average time (in seconds) do you think the human recruiter predicts for you?

### **A.1.3 Information experiment-treatment 1b**

#### **Instructions**

This Human Intelligence Task (HIT) is an academic experiment on decision-making. It will require you to perform a few simple tasks. Your payment will be based on your performance. We also ask some survey questions, both before and after the tasks. All data will be stored anonymously. The entire HIT takes around 20 minutes to complete, and payment is conditional on full completion of the tasks and survey: please only accept this HIT if you can commit to completing it.

After completing this HIT, you will earn 0.1 US Dollar for participating, and up to 3.0 US Dollars in total depending on your performance.

After you have accepted this HIT, the URL to the study will appear here: [link](#).

On the last page, you will be given a completion code. Please copy/paste that code below.

#### **Number-finding task**

Next, you will see a matrix with 9 numbers. You need to find the two numbers that sum to 100. You have a maximum time of 90 seconds. Your payment depends on how fast you find the two correct numbers. If you submit the wrong numbers, an error message will appear and you have to try again until time runs out and the next exercise appears. You repeat this task for 10 rounds. In the end, one of the 10 rounds will be chosen for payment.

Your payment is determined by how fast you find the two correct numbers. Every second is one point and every 90 points are equivalent to 1 US Dollar. On the next page, you are given one training round, which does not affect your payment in any way.

#### **Training round**

Figure A.3: Training round

Time left to complete this page: 1:08

Write in the fields the two numbers that sum to 100. Do it as fast as possible. If you don't write down the correct two numbers, an error message will appear and you have to try again.

87	69	71
4	66	79
21	68	97

First number:

Second number:

### Number-finding task result

Round 9 was chosen for your payment.

Here is your time in round 9:

Round	Your time
9	3.85

Here is the time of an average participant in round 9:

Round	Average participant's time
9	15.4

This means you were faster than an average participant in round 9.

### Hiring

A firm needs people to perform this number-finding task and wants to hire those who are fastest. The firm doesn't know your skill and needs to predict them in one of the two following ways: 1) using a **human recruiter**; 2) using an **algorithmic recruiter**.

Think of the **algorithmic recruiter** as a tool that can poll hundreds of people and determine how many seconds each person needs to solve the task on average. This way, the algorithm can learn which characteristics lead to a specific average time. The algorithm was trained on past data and will compare your characteristics to the characteristics of past participants. Thus, it will calculate your average time based on how similar participants performed in the past.

Think of the **human recruiter** as a person from a group of 20 people who have evaluated several similar participants in the past. We randomly assign one of these human recruiters to you.

Based on only your performance in round 9, your payoff would be \$1.93. However, your other characteristics (gender, age, education, and ethnicity) will be shared with both recruiters, too. This may affect your predicted skill at the number-finding task, and therefore your payoff.

Summarizing, both the human recruiter and the algorithmic recruiter are given the same information about you. This information is:

1. your time in round 9, and
2. your personal characteristics: gender, age, education, and ethnicity.

You can earn up to \$2 in this part.

Please answer the following questions to make sure you understood the explanation above:

1. The algorithm and human recruiter make the hiring decision based on the same information.

*False*    *True*

2. The algorithmic recruiter may assign a different score to you than the human recruiter.

*False*    *True*

### Recruiter choice

Now, choose the recruiter that you would prefer to be evaluated by.

Human recruiter

Algorithmic recruiter

## Willingness to pay for the recruiter

You indicated a preference for a **human recruiter** but you were assigned an **algorithmic recruiter**, instead.

Now you are endowed with 40 cents and asked to choose the portion of this amount (between 0 and 40 cents, inclusive) that you wish to pay to change your assigned recruiter to your preferred option. Then the computer randomly chooses a number between 0 and 40. If your number is higher than the random number, your assigned recruiter will change. Those cents not invested are yours to keep. If your number is lower than the random number, your assigned recruiter will not change and all 40 cents are yours to keep. For your convenience, these instructions will remain available to you on the next page.

## Prediction task instruction

On the next page, you will be asked to guess what average time you think each recruiter (i.e. algorithmic and human) predicted for you and the other participant.

After you have answered all questions, one of the answers will be randomly chosen for payment. Your payoff is determined by how close you are to the actual predictions of either the algorithm or the human recruiter. You will be paid according to the table below:

Table A.3: Prediction task payoff

Difference in seconds*	0	1	2	3	4	5	6	7	8	9	$\geq 10$
Payoff in \$	0.5	0.49	0.48	0.45	0.42	0.37	0.32	0.25	0.18	0.09	0

\* Absolute difference between your guess and the actual average time assigned to you by each recruiter.

## Prediction task

Now, please state how many seconds do you think each recruiter will predict for your average task performance. One of the two answers will be randomly chosen for payment.

1. What average time (in seconds) do you think the algorithmic recruiter predicts for you?

2. What average time (in seconds) do you think the human recruiter predicts for you?

#### **A.1.4 Gender bias experiment**

##### **Instructions**

This Human Intelligence Task (HIT) is an academic experiment on decision-making. It will require you to perform a few simple tasks. Your payment will be based on your performance. We also ask some survey questions, both before and after the tasks. All data will be stored anonymously. The entire HIT takes around 20 minutes to complete, and payment is conditional on full completion of the tasks and survey: please only accept this HIT if you can commit to completing it.

After completing this HIT, you will earn 0.1 US Dollar for participating, and up to 4.5 US Dollars in total depending on your performance.

After you have accepted this HIT, the URL to the study will appear here: [link](#).

On the last page, you will be given a completion code. Please copy/paste that code below.

##### **Number-finding task**

Next, you will see a matrix with 9 numbers. You need to find the two numbers that sum to 100. You have a maximum time of 90 seconds. Your payment depends on how fast you find the two correct numbers. If you submit the wrong numbers, an error message will appear and you have to try again until time runs out and the next exercise appears. You repeat this task for 10 rounds. In the end, one of the 10 rounds will be chosen for payment.

Your payment is determined by how fast you find the two correct numbers. Every second is one point and every 90 points are equivalent to 1 US Dollar. On the next page, you are given one training round, which does not affect your payment in any way.

##### **Training round**

Figure A.4: Training round

Time left to complete this page: 1:08

Write in the fields the two numbers that sum to 100. Do it as fast as possible. If you don't write down the correct two numbers, an error message will appear and you have to try again.

87	69	71
4	66	79
21	68	97

First number:

Second number:

Next

### Number-finding task result

Round 9 was chosen for your payment.

Here is your time in round 9:

Round	Your time
9	3.85

The other participant is female, with the same education, ethnicity, and age as you.

Here is the time of the other participant in round 9:

Round	Other participant's time
9	2.67

### Hiring

A firm needs people to perform this number-finding task and wants to hire those who are fastest. The firm doesn't know your skill and the skill of the other participant at this task and needs to predict them in one of the two following ways: 1) using a **human recruiter**; 2) using an **algorithmic recruiter**.

The participant who is the fastest receives up to \$1 effort payoff and a \$2 bonus for this number-finding task, while the other participant only receives the effort payoff.

Think of the **algorithmic recruiter** as a tool that can poll hundreds of people and determine how many seconds each person needs to solve the task on average. This way, the algorithm can learn which characteristics lead to a specific average time. The algorithm was trained on past data and will compare your characteristics to the characteristics of past participants. Thus, it will calculate your average time based on how similar participants performed in the past.

Think of the **human recruiter** as a person from a group of 20 people who have evaluated several similar participants in the past. We randomly assign one of these human recruiters to you.

Summarizing, both the human recruiter and the algorithmic recruiter are given the same information about you and the other participant. This information is:

1. your time in round 9, and
2. your personal characteristics: gender, age, education, and ethnicity.

Please answer the following questions to make sure you understood the explanation above:

1. The algorithm and human recruiter make the hiring decision based on the same information.

*False*    *True*

2. The algorithmic recruiter may assign a different score to you than the human recruiter.

*False*    *True*

### Recruiter choice

Now, choose your preferred recruiter that is going to evaluate you and the other participant in your group.

Human recruiter	Algorithmic recruiter
-----------------	-----------------------

### Willingness to pay for the recruiter

You indicated a preference for a **human recruiter** but you were assigned an **algorithmic recruiter**, instead.

Now you are endowed with 40 cents and asked to choose the portion of this amount (between 0 and 40 cents, inclusive) that you wish to pay to change your assigned recruiter to your preferred option. Then the computer randomly chooses a number between 0 and 40. If your number is higher than the random number, your assigned recruiter will change. Those cents not invested are yours to keep. If your number is lower than the random number, you assigned recruiter will not change and all 40 cents are yours to keep. For your convenience, these instructions will remain available to you on the next page.

**Prediction task**

On the next page, you will be asked to guess what average time you think each recruiter (i.e. algorithmic and human) predicted for you and the other participant. After you have answered all questions, one of the answers will be randomly chosen for payment. Your payoff is determined by how close you are to the actual predictions of either the algorithm or the human recruiter. You will be paid according to the table below:

Table A.4: Prediction task payoff

Difference in seconds*	0	1	2	3	4	5	6	7	8	9	$\geq 10$
Payoff in \$	1	0.99	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19	0

\* Absolute difference between your guess and the actual average time assigned to you by each recruiter.

**Prediction task**

Now, please state how many seconds do you think each recruiter will predict for your average task performance. One of the two answers will be randomly chosen for payment.

1. What average time (in seconds) do you think the algorithmic recruiter predicts for you?

2. What average time (in seconds) do you think the human recruiter predicts for you?

Below enter your choices for the predicted time by both recruiters for the **other participant's** average task performance.

1. What average time (in seconds) do you think the algorithmic recruiter predicts for the other participant?

2. What average time (in seconds) do you think the human recruiter predicts for the other participant?

### A.1.5 Recruiter weights

The variables we used to construct the weights the training phase are: i) Male: a dummy variable equal to 1 if the respondent is a male; Non-white: a dummy variable equal to 1 if the respondent is not white; Education: a categorical variable taking values from 1 to 7; Age: a categorical variable taking values from 1 to 12; Round 9 time: a continuous variable equal to the time in round 9.

Figure A.5: Human vs Algorithm precision

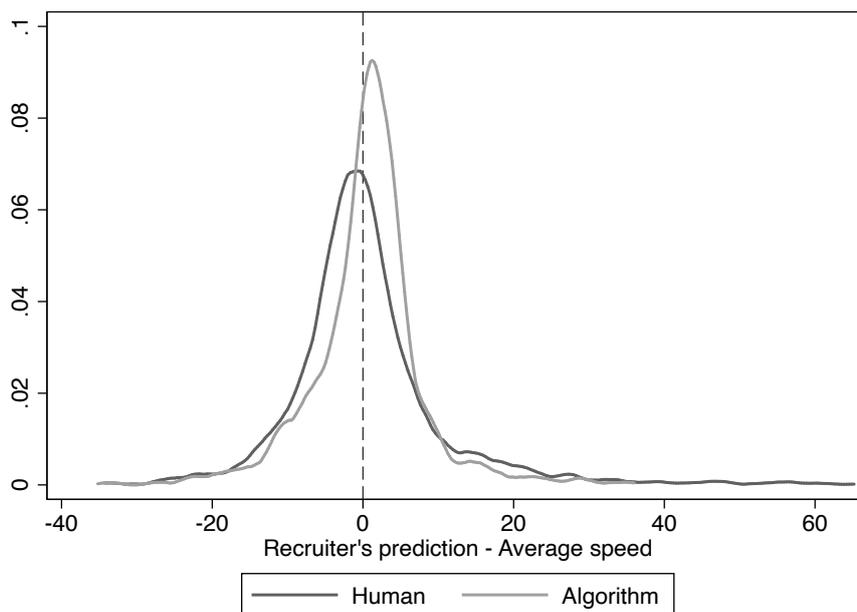


Table A.5: Recruiter coefficients for predicted average time

	Male	Ethnicity	Education	Age	Round 9 time
Algorithm	-1.02	-1.40	0.17	0.36	0.56
Human 1	-1.41	-0.96	-0.35	-0.05	1.04
Human 2	-0.22	-0.52	-0.10	0.09	0.61
Human 3	1.26	2.41	-0.70	1.45	0.47
Human 4	-0.77	-0.22	-0.45	-0.35	0.98
Human 5	-1.48	-1.77	-0.17	-0.27	0.84
Human 6	-0.54	1.23	-0.06	0.45	0.53
Human 7	-0.53	-0.36	-0.28	0.59	0.47
Human 8	-0.35	0.00	-0.23	0.36	0.57
Human 9	-1.83	-0.60	0.01	1.33	0.13
Human 10	-0.15	0.99	-0.76	0.00	0.91
Human 11	-1.67	0.00	-0.12	-0.18	0.87
Human 12	-0.76	-0.36	-0.43	-0.25	0.95
Human 13	-1.33	-0.31	-0.60	0.56	0.70
Human 14	-2.25	-1.96	0.06	0.23	0.77
Human 15	-1.03	1.34	-0.08	0.53	0.50
Human 16	-3.24	-2.62	-1.88	2.72	0.45
Human 17	-1.45	0.79	-1.38	1.30	0.78
Human 18	1.19	-0.36	-0.24	0.20	0.61
Human 19	0.51	0.95	-1.16	1.16	0.53
Human 20	1.68	0.69	-0.45	0.38	0.28
Human 21	0.00	1.12	-0.11	1.45	0.13
Human 22	-0.22	-2.36	-0.97	-0.29	0.98

*Notes:* The dependent variable is the average time to solve the task: a negative coefficient therefore implies a faster time and a better performance.

### A.1.6 Big-five survey

Respondents indicate to what extent they agree or disagree with the following statements:

1. I see myself as someone who worries a lot. (Neuroticism)
2. I see myself as someone who gets nervous easily. (Neuroticism)
3. I see myself as someone who is relaxed, handles stress well. (Neuroticism, reversed scale)

4. I see myself as someone who is talkative. (Extraversion)
5. I see myself as someone who is outgoing, sociable. (Extraversion)
6. I see myself as someone who is reserved. (Extraversion, reversed scale)
7. I see myself as someone who is original, comes up with new ideas. (Openness)
8. I see myself as someone who values artistic, aesthetic experiences. (Openness)
9. I see myself as someone who has an active imagination. (Openness)
10. I see myself as someone who is sometimes rude to others. (Agreeableness, reversed scale)
11. I see myself as someone who has a forgiving nature. (Agreeableness)
12. I see myself as someone who is considerate and kind to almost everyone. (Agreeableness)
13. I see myself as someone who does a thorough job. (Conscientiousness)
14. I see myself as someone who tends to be lazy. (Conscientiousness, reversed scale)
15. I see myself as someone who does things efficiently. (Conscientiousness)

The relevant trait for each statement is shown in parentheses (but not visible to respondents). For each of these items, possible answers are: 1, Strongly disagree; 2, Disagree; 3, Neither agree nor disagree; 4, Agree; and 5, Strongly agree. We sum scores across items belonging to the same category (where items 3, 6, 10, and 14 enter with a reversed scale). This means that for each of the big-five personality traits, the minimum score is 1 and the maximum score is 15.

## A.2 Additional descriptive statistics

Table A.6 shows descriptive statistics for workers included in the pilot study: data from these workers were used to train the algorithmic and human recruiters as described in Section 2.2. Specifically, column (1) shows the data used for training the algorithm whereas column (2) shows the data used by the human recruiters – this is a subset of the data used for the algorithm. Note that we did not collect personality characteristics or risk propensity for this group as this does not enter into the recruitment decision.

Table A.6: Worker descriptives: recruiter training

<i>Recruiter training:</i>	Algorithm (1)	Human (2)
Observed task time (round 9)	20.61 (17.74 )	21.44 (20.14)
Average task time (all rounds)	18.67 (12.34 )	19.33 (12.92)
Male	0.50 (0.50)	0.59 (0.49)
Age	36.09 (11.35)	36.42 (12.80)
Non-white	0.24 (0.43)	0.25 (0.44)
Highly educated	0.58 (0.49)	0.61 (0.49)
N	345	83

*Notes:* Means reported, standard deviations in parentheses. The training data for human recruiters is a subset of the data used to train the algorithm.

Figures A.6 and A.7 show the entire distribution of observed and average task times across all workers. Lastly, Figure A.8 shows the distribution of earnings from the experiments.

Figure A.6: Distribution of recruiter-observed task time (round 9)

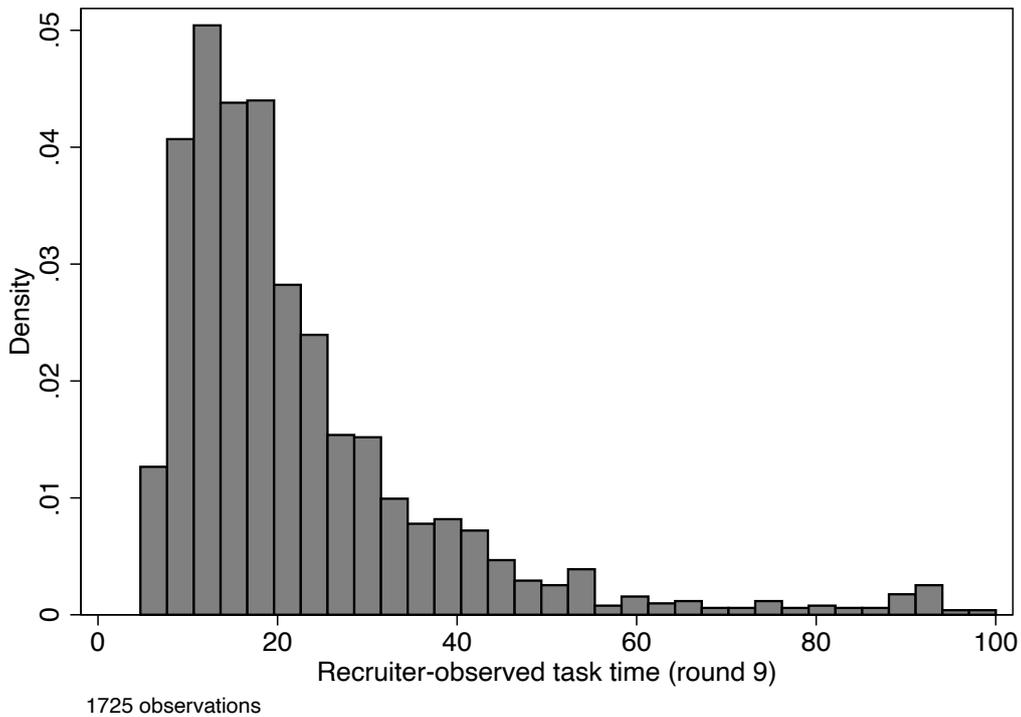
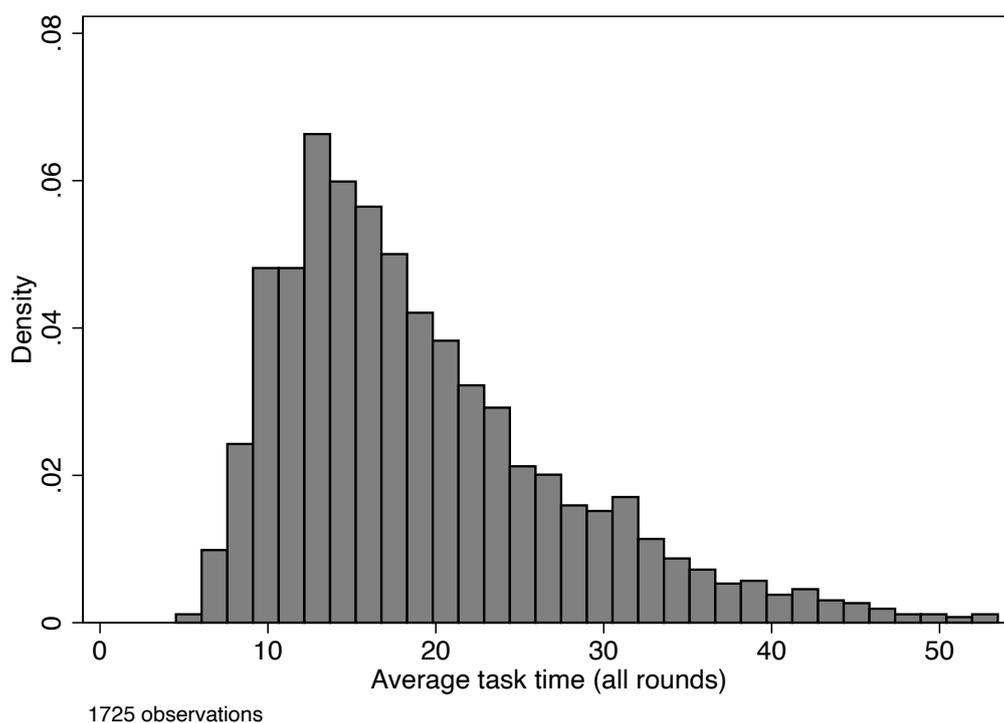


Figure A.7: Distribution of average task time (all rounds)



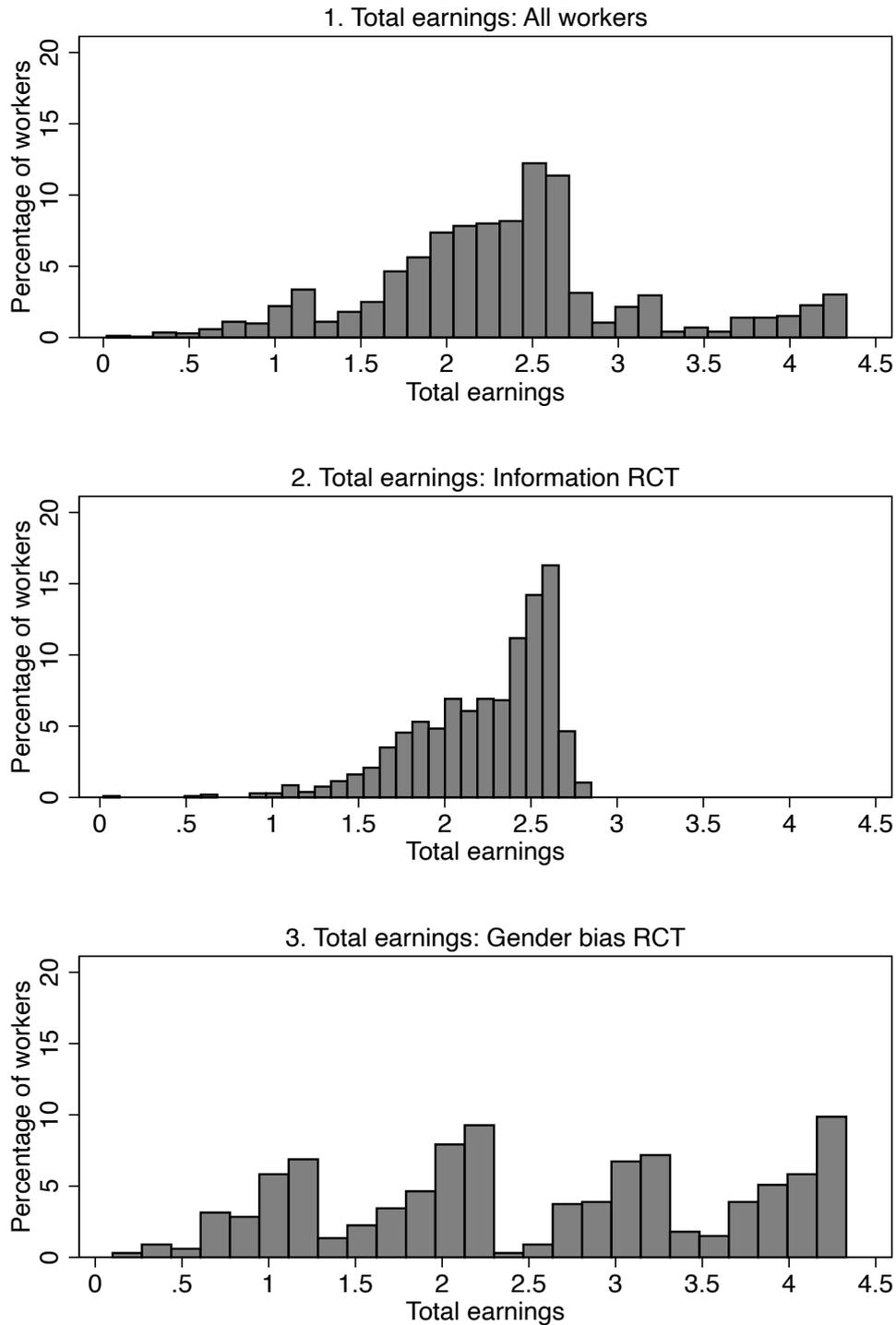
### A.3 Additional results and robustness checks

This appendix presents some additional results and robustness checks.

Figure A.9 shows recruiter perceptions on the original five-point Likert scales rather than collapsed to a three-point scale.

Table A.7 presents the correlates of recruiter choice and willingness to pay for the gender bias experiment. While results are similar overall, in the gender bias experiment, the evidence that workers choose the recruiter that rates them most favorably is weaker than in the information experiment (see Table 2). This may be due to the fact that they have to evaluate their performance in relation to the competitor's. Workers who expect the human (algorithm) to be in more their favor than the algorithm (human) are more (less) likely to choose the human recruiter, but the point estimate are small and imprecisely estimated. As in the information experiment, preferring the human recruiter is a significant predictor of willingness to pay, even when controlling for all other covariates. Observed task performance is still a strong and significant predictor of recruiter preference also in the gender bias experiment: workers who are faster than their competitor are less

Figure A.8: Distribution of total earnings

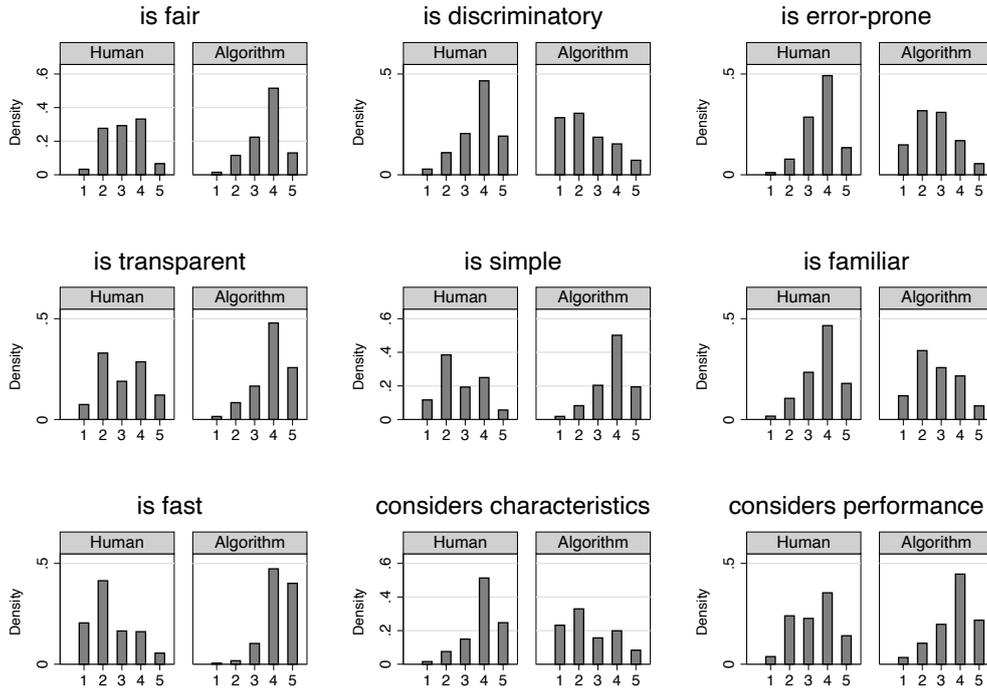


1725 observations

likely to choose the human, even if the difference in speed between the two players is, by construction, very small.

As before, worker characteristics are not the main determinant of recruiter choice.

Figure A.9: Perception of the recruiter



Notes: Likert scale. 1 is Strongly disagree, 2 is Disagree, 3 is Neither agree nor disagree, 4 is Agree, and 5 is Strongly agree.

While non-white workers are 7.4 percentage points more likely to choose the human recruiter, all other coefficients on worker characteristics are precisely estimated but small in magnitude. There is also some evidence that workers who are male, older, non-white, neurotic, open, extroverted, and risk-loving have a higher willingness to pay. As in the information experiment, willingness to pay is higher for workers who prefer the human recruiter. These higher differences in willingness to pay as compared to the information experiment are likely due to the monetary pay-off being bigger, if the competitor is bested.

Table A.8 shows robustness check on the main results presented in Table 2, by using the subsample of workers who gave the correct answer to both attention check questions. Similarly, Table A.9 presents the same robustness check for the gender bias experiment.

Further, Table A.11 presents results for the two separate information treatments (1a and 1b, as described in Section 2) rather than pooling them into a single treatment dummy (as in Table 3). Results are similar to those of the main specification: the information treatment makes workers with a low observed task performance more likely to choose the

Table A.7: Correlates of recruiter choice and willingness to pay, gender bias experiment

	<i>Dependent variable:</i>					
	Preference for human recruiter			WTP for preferred recruiter		
	(1)	(2)	(3)	(4)	(5)	(6)
Human favorable	0.0874*	0.0829*	0.0689	9.438***	9.444***	7.269**
	(1.81)	(1.73)	(1.46)	(2.64)	(2.64)	(2.13)
Algorithm favorable	-0.0645	-0.0569	-0.0674	7.326**	7.313**	5.392
	(-1.39)	(-1.23)	(-1.48)	(2.14)	(2.13)	(1.65)
Faster than competitor		-0.131***	-0.137***		0.253	-0.528
		(-3.41)	(-3.64)		(0.09)	(-0.19)
Male			-0.0187			7.170**
			(-0.49)			(2.58)
Age			0.00378**			0.319**
			(2.16)			(2.52)
Non-white			0.0741*			5.249*
			(1.77)			(1.74)
Highly educated			-0.0530			-0.438
			(-1.30)			(-0.15)
Risk			0.0172**			1.933***
			(2.20)			(3.43)
Conscientiousness			-0.0135			-2.202***
			(-1.30)			(-2.94)
Agreeableness			0.00241			-0.743
			(0.24)			(-1.01)
Neuroticism			0.0148**			1.084**
			(2.05)			(2.07)
Openness			-0.0168*			1.181*
			(-1.86)			(1.82)
Extroversion			0.0310***			2.272***
			(4.00)			(4.02)
Prefer human				11.15***	11.19***	5.907**
				(3.89)	(3.86)	(2.09)
Constant	0.536***	0.600***	0.298	26.45***	26.31***	-9.039
	(20.00)	(18.47)	(1.43)	(10.56)	(8.84)	(-0.60)
N	669	669	669	669	669	669
R <sup>2</sup>	0.019	0.036	0.098	0.055	0.055	0.170
adj. R <sup>2</sup>	0.009	0.025	0.073	0.043	0.042	0.146

*Notes:* Estimates for the gender bias experiment. Willingness to pay is measured as a percentage of the total allocated budget. The variable *Human (/ Algorithm) favorable* is defined as a dummy variable which is equal to one if the human (/ algorithm) rates the respondent more favorably than the competitor compared to the other recruiter: the omitted category are workers who expect no difference in how the two recruiters rate them relative to their competitor. *Faster than competitor* is a dummy equal to 1 if the worker has a faster observed task time than their competitor. All models are estimated with OLS and control for the date of data collection. *t* statistics in parentheses; \* p<0.1 \*\* p<0.05, \*\*\* p<0.01.

human recruiter, and more willing to pay for their preferred recruiter. This is true both for information treatment 1a and information treatment 1b, and the point estimates of the two treatments are very similar.

We also study workers' monetary pay-off from their recruiter choice. Specifically, we calculate workers' pay-off for both recruiters, and compare the one they actually obtained to the counterfactual one had they chosen a different recruiter. Specifically, we estimate the following model:

$$\text{payoff}_i = \alpha + \beta_1 \times \text{observed performance}_i + \beta_2 \times \text{average performance}_i + \phi_i + u_i \quad (4)$$

where  $i$  indexes individual workers. The dependent variable is the pay-off for the recruiter choice, in US Dollar cents: this depends exclusively on the recruiter's prediction of average task performance across all task rounds. The variable observed performance is task performance in round 9 (as before, measured as the negative of the number of seconds needed to solve the round 9 task, standardized): in other words, this is the performance signal observed by recruiters. Further, average performance is average task performance across all rounds (measured as the negative of the number of seconds needed to solve the task on average across all rounds, standardized): that is, it reflects workers' actual average ability at the number-finding task which recruiters try to predict. We also estimate this model with a set of controls contained in  $\phi_i$ . These are dummies for whether the worker expects the human or algorithmic recruiter to assign them a better score (the omitted category being equal expected scores across recruiters), worker characteristics (gender, age, ethnicity, and education level), big-five personality traits and risk propensity, as well as fixed effects for the date of data collection. We also control for workers' willingness to pay, and a dummy for being treated.

We limit our analysis to the information experiment, where the pay-offs from recruiter choice depend only on the worker and not also on the fictitious competitor, for whom we do not have an average task ability. Lastly, we additionally estimate equation 4 separately for those who choose the human or algorithmic recruiter.

Results in Appendix Table A.12 highlight that workers' performance signal and average

performance have *opposite* signs on pay-offs from recruiter choice. As seen in columns (1) and (2), workers who have a one standard deviation better average task ability earn more from their recruitment choice (the point estimate is 0.931): this suggests those who are better at the number-finding task are also better at choosing the recruiter that evaluates them more favorably. However, performing better on the observed task (round 9) has the opposite effect: when a worker performs better in round 9, holding constant their average task ability, they gain less from their recruiter choice. This means that workers who send a *worse* performance signal, holding constant their actual ability, are *more* likely to gain from their choice – and this is entirely driven by workers choosing human recruitment. This can be seen by comparing columns (3) and (4), the subsample of workers choosing algorithmic recruitment, to columns (5) and (6), the subsample of workers choosing human recruitment. Workers who send a worse performance signal (relative to their average ability) and choose human recruitment gain from their choice, whereas those who send a worse signal (relative to their average ability) but choose algorithmic recruitment lose. Since on average, workers make the right choice (i.e. they choose human recruitment when their signal is poor), workers with a poor signal gain more.<sup>23</sup>

Lastly, Tables A.13 and A.14 provide robustness checks on the results found in Tables 4 and A.12, respectively, by considering the subsample of workers who answered both attention check questions correctly.

---

<sup>23</sup>Appendix Table A.14 shows these findings are robust to restricting the sample to workers who answered both attention check questions correctly.

Table A.8: Correlates of recruiter choice and willingness to pay, information experiment (robustness check)

	<i>Dependent variable:</i>					
	Preference for human (1)	Preference for human recruiter (2)	WTP for preferred recruiter (3)	WTP for preferred recruiter (4)	WTP for preferred recruiter (5)	WTP for preferred recruiter (6)
Human favorable	0.243*** (4.70)	0.226*** (4.42)	0.208*** (4.10)	11.27*** (3.49)	11.00*** (3.41)	11.53*** (3.62)
Algorithm favorable	-0.0465 (-0.96)	-0.0575 (-1.20)	-0.0619 (-1.30)	9.851*** (3.30)	9.482*** (3.18)	10.78*** (3.65)
Observed task performance		-0.105*** (-5.03)	-0.0913*** (-4.31)		-3.082** (-2.34)	-2.043 (-1.54)
Male			-0.0537 (-1.45)			1.824 (0.79)
Age			-0.00036 (-0.23)			0.157* (1.66)
Non-white			-0.0740* (-1.84)			4.113 (1.65)
Highly educated			0.0406 (1.12)			-1.167 (-0.52)
Risk			-0.00017 (-0.02)			1.847*** (3.98)
Conscientiousness			-0.00493 (-0.55)			-0.612 (-1.11)
Agreeableness			0.00457 (0.53)			-0.393 (-0.74)
Neuroticism			0.0158*** (2.61)			0.922** (2.45)
Openness			0.0114 (1.51)			-0.237 (-0.51)
Extroversion			0.0181*** (2.88)			0.275 (0.70)
Prefer human				3.944* (1.74)	2.961 (1.29)	2.743 (1.19)
Constant	0.408*** (9.86)	0.438*** (10.64)	0.0756 (0.40)	11.46*** (4.24)	12.74*** (4.63)	-0.618 (-0.05)
N	747	747	747	747	747	747
R <sup>2</sup>	0.081	0.112	0.146	0.039	0.046	0.089
adj. R <sup>2</sup>	0.066	0.096	0.119	0.021	0.027	0.059

*Notes:* Estimates for the information experiment (subsample of workers who gave the correct answer to both check questions). Willingness to pay is measured as a percentage of the total allocated budget. The variable *Human (/ Algorithm) favorable* is defined as a dummy variable which is equal to one if the human (/ algorithm) rates the respondent more favorably than the competitor compared to the other recruiter: the omitted category are workers who expect no difference in how the two recruiters rate them relative to their competitor. Task performance is measured as the negative of the number of seconds the worker took to solve the observed task, standardized to have a zero mean and unit standard deviation. All models are estimated with OLS and control for the date of data collection. *t* statistics in parentheses; \* p<0.1 \*\* p<0.05, \*\*\* p<0.01.

Table A.9: Correlates of recruiter choice and willingness to pay, gender bias experiment (robustness check)

	<i>Dependent variable:</i>					
	Preference for human recruiter (1)	(2)	(3)	WTP for preferred recruiter (4)	(5)	(6)
Human favorable	0.119* (1.94)	0.118* (1.93)	0.0990 (1.63)	8.513* (1.86)	8.520* (1.86)	5.360 (1.27)
Algorithm favorable	-0.00981 (-0.16)	0.00423 (0.07)	-0.0193 (-0.32)	8.631* (1.91)	8.696* (1.91)	5.683 (1.35)
Faster than competitor		-0.159*** (-3.19)	-0.167*** (-3.35)		-0.746 (-0.20)	-0.850 (-0.24)
Male			-0.0165 (-0.32)			12.28*** (3.47)
Age			0.00453** (2.03)			0.424*** (2.74)
Non-white			0.0713 (1.30)			5.907 (1.56)
Highly educated			-0.0162 (-0.31)			-1.931 (-0.53)
Risk			0.0194* (1.79)			2.558*** (3.41)
Conscientiousness			-0.00061 (-0.05)			-2.428*** (-2.61)
Agreeableness			0.00572 (0.44)			-1.106 (-1.23)
Neuroticism			0.0193** (2.04)			1.455** (2.21)
Openness			-0.00700 (-0.62)			1.477* (1.89)
Extroversion			0.0230** (2.21)			1.787** (2.47)
Prefer human				12.54*** (3.39)	12.42*** (3.32)	6.741* (1.92)
Constant	0.489*** (13.98)	0.565*** (13.46)	-0.0548 (-0.20)	24.41*** (7.74)	24.82*** (6.56)	-15.83 (-0.83)
N	407	407	407	407	407	407
R <sup>2</sup>	0.014	0.039	0.091	0.062	0.062	0.243
adj. R <sup>2</sup>	-0.003	0.019	0.049	0.043	0.040	0.206

*Notes:* Estimates for the gender bias experiment (subsample of workers who gave the correct answer to both check questions). Willingness to pay is measured as a percentage of the total allocated budget. The variable *Human (/ Algorithm) favorable* is defined as a dummy variable which is equal to one if the human (/ algorithm) rates the respondent more favorably than the competitor compared to the other recruiter: the omitted category are workers who expect no difference in how the two recruiters rate them relative to their competitor. *Faster than competitor* is a dummy equal to 1 if the worker has a faster observed task time than their competitor. All models are estimated with OLS and control for the date of data collection. *t* statistics in parentheses; \* p<0.1 \*\* p<0.05, \*\*\* p<0.01.

Table A.10: The impact of task performance on recruiter preference (robustness check)

<i>A. Dependent variable: Preference for human recruiter</i>		
	(1)	(2)
Treated, low observed task performance	0.260*** (4.14)	0.212*** (3.42)
N [0.5em]	395 (3)	395 (4)
Treated, high observed task performance	0.0145 (0.23)	0.00501 (0.08)
N	352	352
<i>B. Dependent variable: Willingness to pay for preferred recruiter</i>		
	(5)	(6)
Treated, low observed task performance	10.02** (2.42)	8.273** (1.98)
N	395 (7)	395 (8)
Treated, high observed task performance	4.319 (1.20)	2.541 (0.70)
N	352	352
Additional controls	No	Yes

*Notes:* Estimates for the information experiment (subsample of workers who gave the correct answer to both check questions). Treated workers who were made aware they had a low observed task performance are compared to control group workers with low observed task performance. Treated workers who were made aware they had a high observed task performance are compared to control group workers with high observed task performance. All models are estimated with OLS and control for the date of data collection. Additional controls are  $\psi_i$  from equation 1. *t* statistics in parentheses; \*  $p < 0.1$  \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table A.11: The impact of task performance on recruiter preference (two information treatments)

<i>A. Dependent variable: Preference for human recruiter</i>		
	(1)	(2)
Treated (1a), low observed task performance	0.182*** (3.64)	0.145*** (2.93)
Treated (1b), low observed task performance	0.175*** (3.66)	0.135*** (2.82)
N	595	595
	(3)	(4)
Treated (1a), high observed task performance	0.00888 (0.17)	0.00327 (0.06)
Treated (1b), high task performance	0.0969* (1.77)	0.0676 (1.27)
N	461	461
<i>B. Dependent variable: Willingness to pay for preferred recruiter</i>		
	(5)	(6)
Treated (1a), low observed task performance	7.748** (2.27)	6.249* (1.87)
Treated (1b), low observed task performance	8.744*** (2.68)	5.868* (1.82)
N	595	595
	(7)	(8)
Treated (1a), high observed task performance	1.520 (0.48)	1.452 (0.46)
Treated (1b), high observed task performance	9.047*** (2.78)	7.453** (2.30)
N	461	461
Additional controls	No	Yes

*Notes:* Estimates for the information experiment. Treated workers who were made aware they had a low observed task performance are compared to control group workers with low observed task performance. Treated workers who were made aware they had a high observed task performance are compared to control group workers with high observed task performance. All models are estimated with OLS and control for the date of data collection. Additional controls are  $\psi_i$  from equation 1.  $t$  statistics in parentheses; \*  $p < 0.1$  \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table A.12: Performance signals and gain from recruiter choice

*Dependent variable: Pay-off from recruiter choice*

	Full sample		Algorithm chosen		Human chosen	
	(1)	(2)	(3)	(4)	(5)	(6)
Observed task performance	-1.461*** (-5.44)	-1.552*** (-5.71)	5.242*** (20.75)	5.285*** (24.50)	-6.176*** (-32.19)	-6.126*** (-37.34)
Average task performance	0.913*** (3.58)	0.912*** (3.49)	0.257 (1.15)	0.535*** (2.75)	0.371* (1.91)	0.112 (0.66)
N	1,056	1,056	551	551	505	505
Additional controls	No	Yes	No	Yes	No	Yes

*Notes:* Estimates for the information experiment. Pay-off from choice are measured in US Dollar cents. Observed task performance is the negative of the number of seconds the worker took to solve the observed task, standardized to have a zero mean and unit standard deviation. Average task performance is the negative of the number of seconds the worker took to solve all observed tasks on average, standardized to have a zero mean and unit standard deviation. All models are estimated with OLS and control for the date of data collection. Additional controls are  $\phi_i$  from equation 3. *t* statistics in parentheses; \* p<0.1 \*\* p<0.05, \*\*\* p<0.01.

Table A.13: Perceived relative recruiter gender bias, by task performance (robustness check)

*A. Dependent variable: Preference for human recruiter*

	Faster than competitor		Slower than competitor	
	(1)	(2)	(3)	(4)
Treated	-0.129 (-0.88)	-0.130 (-0.91)	0.266* (1.87)	0.303** (2.05)
N	202	202	205	205

*B. Dependent variable: Willingness to pay*

	Faster than competitor		Slower than competitor	
	(5)	(6)	(7)	(8)
Treated	12.57 (1.16)	10.18 (1.00)	-8.198 (-0.78)	-5.582 (-0.55)
N	202	202	205	205
Additional controls	No	Yes	No	Yes

*Notes:* Estimates for the gender bias experiment (subsample of workers who gave the correct answer to both check questions). Treated workers are those paired with a different-gender competitor. All models are estimated with OLS and control for the date of data collection. Additional controls are  $\psi_i$  from equation 2. *t* statistics in parentheses; \* p<0.1 \*\* p<0.05, \*\*\* p<0.01.

Table A.14: Performance signals and gain from recruiter choice (robustness check)

*Dependent variable: pay-off from recruiter choice*

	Full sample		Algorithm chosen		Human chosen	
	(1)	(2)	(3)	(4)	(5)	(6)
Observed task performance	-1.768*** (-5.35)	-1.985*** (-6.02)	6.530*** (26.57)	6.521*** (42.61)	-6.392*** (-30.72)	-6.247*** (-36.72)
Average task performance	0.954*** (3.05)	0.953*** (2.99)	-0.378* (-1.86)	-0.142 (-1.07)	0.474** (2.19)	0.0434 (0.24)
N	747	747	402	402	345	345
Additional controls	No	Yes	No	Yes	No	Yes

*Notes:* Estimates for the information experiment (subsample of workers who gave the correct answer to both check questions). Pay-off from choice are measured in US Dollars. Observed task performance is the negative of the number of seconds the worker took to solve the observed task, standardized to have a zero mean and unit standard deviation. Average task performance is the negative of the number of seconds the worker took to solve all tasks on average, standardized to have a zero mean and unit standard deviation. All models are estimated with OLS and control for the date of data collection. Additional controls are  $\phi_i$  from equation 3.  $t$  statistics in parentheses; \*  $p < 0.1$  \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table A.15: Performance signals and recruiter choice (robustness check)

*Dependent variable: Preference for human recruiter*

	(1)	(2)
Bad signal	0.0730*** (3.32)	0.0627*** (2.88)
$N$	747	747
$R^2$	0.028	0.077
adj. $R^2$	0.014	0.050
Additional controls	No	Yes

*Notes:* Estimates for the information experiment (subsample of workers who gave the correct answer to both check questions). Bad signal is the difference between the number of seconds the worker took to solve round 9 and the number of seconds the worker took to solve all observed tasks on average; standardized to have a zero mean and unit standard deviation. All models are estimated with OLS and control for the date of data collection. Additional controls are gender, age, race, education, personality characteristics and risk aversion.  $t$  statistics in parentheses; \*  $p < 0.1$  \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .