Reducing perceptions of discrimination (follow-up to AEARCTR-0009592): Pre-analysis Plan

July 2023

Abstract

This pre-analysis plan documents the intended analysis for an experiment that follows up on AEARCTR-0009592. This follow-up randomized experiment examines how individuals perceive discrimination, further (relative to the original experiment) varying the methods used to hire workers and what workers know about them to understand certain mechanisms behind the original treatments that reduce perceptions of discrimination. The main outcome is the rate of perceived discrimination in each of 6 treatment arms (four of which replicate the original experiment). The follow-up will also replicate and extend the results of the original experiment on the effects of perceived discrimination on future labor supply, and, unlike the original experiment, will measure comprehension of the various treatments. This plan outlines the study design and hypotheses, outcomes of interest, and empirical specifications.

1 Study design

Given the parallels of this experiment to the original experiment,[1] this pre-analysis plan will outline the deviations of this follow-up from the original experiment and build on the initial pre-analysis plan, referencing readers to that original plan as needed.

This follow-up seeks to understand the mechanisms behind certain results from the main experiment. Specifically, whether perceptions of algorithmic discrimination in the main experiment are due to (1) inattention or misunderstanding (since the algorithm did not use demographics in its decision-making) or (2) sophistication about the potential for algorithms that do not explicitly use demographics to still discriminate. I will also replicate the effects of perceived manager

[1] Details at https://www.socialscienceregistry.org/trials/9592

discrimination on future labor supply and provide more precise estimates of the effects of perceived algorithmic discrimination on future labor supply (if there is a first stage; that is, if there are still positive perceptions of algorithmic discrimination in one or more arms of the follow-up experiment). I will also directly measure worker comprehension of the various mechanisms.

To do this, I extend the original dynamic labor market on Prolific and set up an experiment that varies what mechanism (and what workers know about it) is used to hire *new* workers to do the harder, higher-paying proofreading task from the original experiment. In addition to the three mechanisms in the original experiment (managers with or without demographics and an algorithm that doesn't use demographics), the follow-up adds evaluation by an algorithm that does use demographics in its hiring decisions. As in the main experiment, the follow-up varies whether workers know the demographics of those hired in the past among workers evaluated by managers or algorithms that don't use demographics to make decisions,. This generates six treatment arms (including the four from the original experiment). Changing the context from promotion (as in the original experiment) to hiring allows an additional extension, which is to replicate the main results in a hiring context rather than promotion (for perceived discrimination outcomes and future labor supply).

1.1 The hiring mechanisms

There are three hiring mechanisms of interest **which are identical to** the job assignment mechanisms in the original experiment: a manager (also recruited on Prolific) who knows a worker's demographic characteristics, education, and average score on the three training tasks when making the decision about whether to hire the worker, a screening algorithm that predicts workers' performance at the harder proofreading task based on their average score on the screening tasks and their education, and a demographic-blinded manager who only knows a worker's average screening task score and education when making the hiring decision. **A fourth hiring mechanism new to the follow-up experiment** is an algorithm that allows the predictive relationship of screening scores and education to vary by race and gender. In reality, this can make the algorithm less biased. Workers may perceive "an algorithm that uses race and gender" (along with quiz scores and education) to be more biased, however, without more information about how the algorithm works.

The manager task (managers are also recruited from Prolific) is exactly the same as in the main experiment, though to minimize costs they will each evaluate **nine** sets of groups of 40 workers, hiring one in each group to do the hard proofreading task (and whose performance determines their bonus payment). The managers (who are the same managers who participated in the original

experiment) keep their prior randomly-assigned status as a "demographic-blind" or "non-blind" manager and evaluate workers with quiz scores in the same range as the workers they evaluated in both the historical and main sample in the original experiment. They are paid a bonus based on the performance of the worker in one randomly-selected group of 40 who they chose to hire.

As in the original experiment, all workers will actually be evaluated by all four mechanisms in order to generate counterfactual data on whether they would have been hired if they had been randomly assigned to a different treatment arm. Only the decision of their actually randomly-assigned arm is implemented and this is what they are told about in the main treatment.

1.2 Generating historical data on manager and algorithm choices

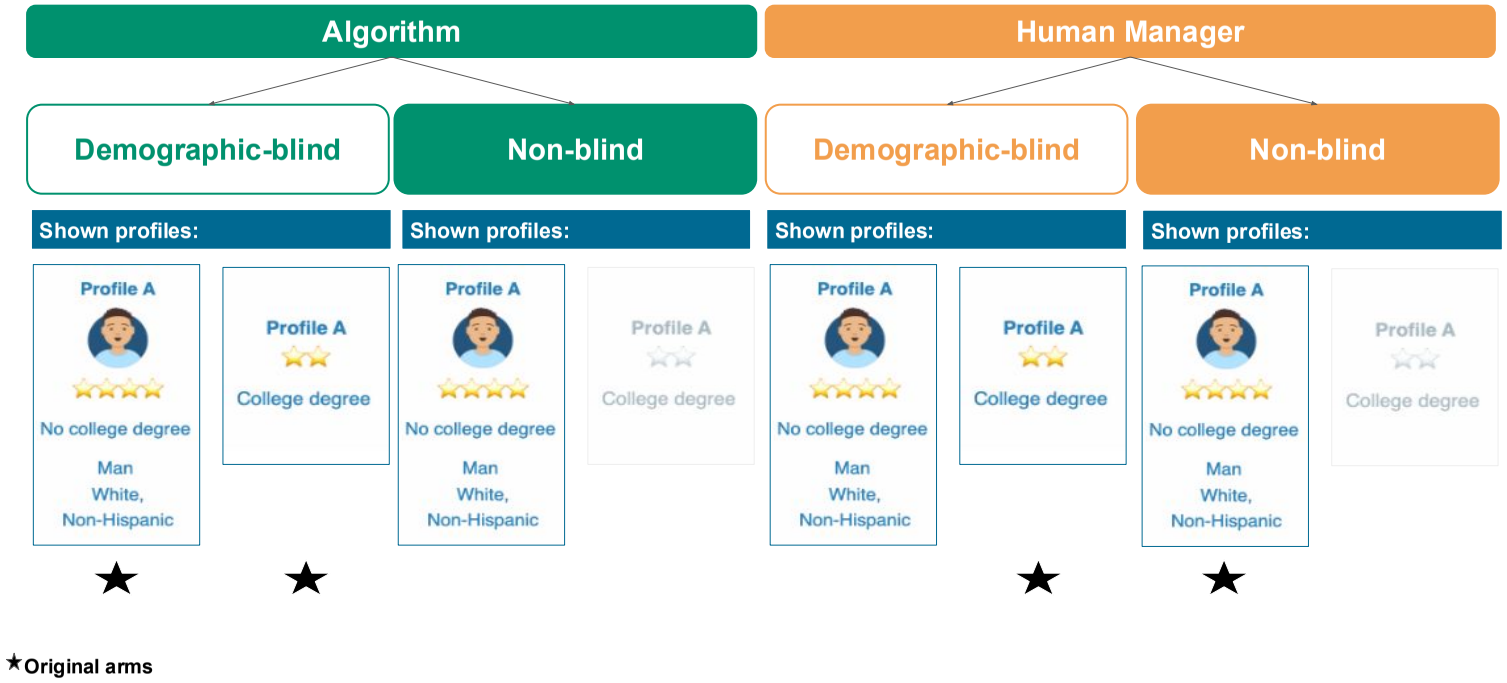
The managers that participate in the follow-up experiment (evaluating workers and deciding which ones to hire for the hard proofreading task) will be the **same** Prolific recruits as participated in the main experiment. This eliminates the need to recruit new "historical workers" to generate data on managers' past decision-making. The new algorithm can also be run on this old sample to generate historical decisions for that new mechanism.

1.3 Sampling strategy

As in the main experiment, the follow-up will take place on the online survey platform Prolific where researchers recruit study participants. The main experiment over-sampled particular demographics to allow for comparisons across minority racial groups and to focus on workers who are more likely to experience discrimination: the sample was approximately 50 percent white and non-Hispanic women, 7 percent each Black, Asian, and Hispanic women, 7 percent each Black, Asian, and Hispanic men, and 8 percent white and non-Hispanic men. This differs from the original demographic makeup of the original experiment's pre-analysis plan due to constraints on the number of racial minority men and women I was able to recruit on Prolific.

In the follow-up, I will recruit only women and racial minority men to the experiment, since the original experiment documents that white men do not perceive discrimination in this context. I do not plan to be powered to test for racial or gender differences (unlike the main experiment) since the only racial minority workers who will be available to participate are those who joined Prolific since the original experiment was run (since all available racial minority workers willing to take this survey were previously recruited). Thus, the sample will be primarily white women.

Figure 1: Experimental design



1.4 Experimental design

The follow-up will follow the same timeline as the original experiment: workers will take a baseline survey, be evaluated by different mechanisms, randomly assigned to have the decision of one mechanism implemented, and brought back for the experimental survey if they aren't hired to do the proofreading task by their assigned mechanism. **The differences are as follows:**

1. Workers will take a **shortened** baseline survey to minimize costs. The survey still includes all key information needed to evaluate workers (the same quizzes, avatars, questions about demographics, etc.).

2. Workers will be evaluated by **four** different mechanisms rather than three.

3. Workers will be brought back for the experimental survey if they aren't hired to do the proofreading task (those that are hired are offered that task), **but this survey does not include a proofreading task.**

See Figure 1 for the full experimental design. Workers are randomly assigned to one of six treatment groups, which vary in which evaluation mechanism was used and what they know about it. The design nests the original four treatment arms, marked with stars. The "treatment" is similar to the original experiment – when workers return for the experimental survey, they are told about how they were evaluated and that they weren't hired. As before, the experiment will also randomize whether workers have demographic information about who their mechanism hired in the past when their mechanism *did not* have access to demographic information. However, in those arms, **it will be more salient in the follow-up** that the mechanism did not have access to demographic information. Thus, this is a 4x2 design with two irrelevant arms – where the hiring mechanism did know demographics but workers don't know the demographics of the historically promoted workers (grayed out in Figure 1).

The outcomes are collected after workers receive all of this information.

Worker randomization. Workers are randomly assigned to groups of forty (call this an "assignment group") with similar test scores that are jointly assigned to one of the six arms in Figure 1. Pairs of managers (one demographic-blind, one not) evaluate 9 sets of these groups of 40 (call this an "evaluation group"), and their decisions are implemented for randomly selected assignment groups. The algorithms are also used to evaluate each group of 40, hiring the worker with the highest predicted performance in each group to do the proofreading task. There will be 11 evaluation groups (3960 workers will complete the screening survey). Then, among each evaluation group,

one assignment group is randomly assigned to each of three of the arms in Figure 1 and two assignment groups are randomly assigned to each of the other three arms. The arms that get one versus two in each evaluation group are randomly assigned so that there are equal numbers of workers per arm and there is no systematic relationship between the arms that got multiple assignment groups within one evaluation group. Managers are randomly paired together and are randomly assigned to evaluation groups (conditional on evaluating workers with similar quiz scores as those they evaluated in the past, and with preference to managers who participated in the main experiment as they are more likely to quickly return to evaluate workers and those who historically assigned three white men in the past, as that is shown in the main experiment to increase rates of perceived discrimination which will improve power).

Work task and survey. After all workers have been evaluated, any worker who is hired to do the proofreading task by their manager or the algorithm (depending on their treatment assignment) will be offered that task. They will do the task and finish the experiment (2.5 percent of workers).

At the same time, any worker who is not hired by their manager or the algorithm (depending on their treatment assignment) will be offered the experimental survey. Since the study is framed as being about hiring mechanisms, asking workers to come back to answer questions about the mechanisms should seem relatively natural, especially because some of the questions are about their interest in future work. Among this sample of interest, after agreeing to take the follow-up survey, workers will be told about how they were evaluated (and not hired). **They will see information identical to the main experiment with three main exceptions:**

- In the (new) "algorithm using demographics" arms, workers will know that the algorithm used demographics in its prediction of who would do the best at the difficult proofreading task
- When workers are shown the demographics of the historical workers hired in the past, but the mechanism that evaluated them did not actually use demographics, they will first see their own profile with the information the mechanism *did* use (quiz score stars and education) before seeing the profiles of the historically hired workers that include those workers' avatars and demographics. In the arms that did use demographics, they will see their profile at the same point with their demographic information included. *This is the only change to the four arms that are replicated from the main experiment, aside from transitioning from a promotion to hiring context. Comparing these arms in the follow-up to those in the main experiment will help illuminate whether workers were confused/forgetful or sophisticated in how they thought about algorithmic bias, though this is not the key to understanding this result.*

As in the original experiment, after they are told about how they were evaluated, workers are asked a free-response question about what they think would have needed to be different about their profile in order to be hired, and then are asked how many stars they think they would have needed to score on the screening quizzes in order to be hired.

Unlike the original experiment, workers do not do a proofreading task.

After they learn about their evaluation, I ask workers about their interest in future work, again **identical to the original experiment *except* that workers' reservation wages are for the chance to be evaluated again and potentially hired, rather than assigned to the easier or harder task**. Specifically, workers' reservation wages to do this work again will be elicited, first under the assumption that the hiring mechanism is the same as in the experimental treatment (as depends on their treatment assignment) and again under the assumption that the workers with the top screening quiz scores will be hired (a cutoff rule). One set of wages will be randomly selected and a random subset of workers will be selected to have their choices implemented at that wage, separately for the original and the cutoff hiring mechanisms. They also report their beliefs about the likelihood they will be hired in each case.

Then, workers will answer questions about their self-efficacy to do the difficult proofreading task, affective well-being, and three measures of explicit perceived discrimination: whether they have any complaints about the hiring process, and whether they think they would have been hired if they had a different race or gender. Finally, they will answer incentivized comprehension questions about the hiring mechanism used to evaluate them.

1.5 Sample size and statistical power

All calculations assume power of 80 percent and a significance level of 0.95. I will recruit 3,960 participants to complete the screening survey. I expect take-up for the follow-up survey to be high, since the screening survey's initial description will indicate that there is a well-paid follow up survey. Assuming that 82 percent of workers complete the follow-up survey (as in the original experiment) and 98.5 percent of workers are not hired (by design), the final analysis sample will be around 3166 workers who weren't hired by their randomly assigned mechanism or around 2,850 workers who weren't hired by *any* of the four mechanisms (assuming 10 percent of workers are hired by at least one mechanism as in the original experiment); see Section 3 for a discussion of why the analysis might use either of these samples.

This is approximately 530 (475) participants in each of the six treatment groups in Figure 1. The primary research question in this follow up is whether and how algorithms are effective at reducing

perceptions of discrimination. Thus, I am most interested in comparing each of the other treatment groups to the non-blind manager, and I calculate conservative MDEs by focusing on the comparisons of two groups at a time without control variables – pooling arms and adding controls would improve the precision of the estimates.

With this sample size, I will be powered ($\alpha = 0.05$, power = 80 percent) to detect a 8.5pp (9pp) difference in either direction between each treatment group and the non-blind manager group (based on analytical power calculations with a total sample size of 1060 (950), and assuming that 40 percent of participants perceive discrimination in the non-blind manager group, as in the original experiment (among women and racial minority men, who will make up the whole sample for the follow-up). Given the large differences in perceived discrimination (20pp or more) between arms in the main experiment, this is a reasonable minimum detectable effect (MDE).

Given the sample size needed to obtain the power described above, I can also calculate the MDEs for the differences between other treatment groups, depending on the rate of perceived discrimination in the less-discriminatory group, all of which would be better-powered based on the results from the main experiment. For example, I am interested in testing whether the algorithm that uses demographics is perceived to discriminate more than the algorithm that doesn't, as well as the difference between the arms with the blind manager and the algorithm without demographics in which workers know that mostly white men were hired in the past. In these cases, the relevant control mean is 20%, not 40%, so I would be powered to detect differences larger than 7.3pp (7.7pp) with 530 workers per group or 475 workers per group, respectively. Below is a table of the rest of the implied MDEs (calculated analytically using the same assumptions) for the differences between two groups when the comparison group perceives discrimination at various rates:

Comparison group: percent perceiving discrimination	MDE if effect is positive, N per group = 530	MDE if effect is positive, N per group = 475
1 percent	2.5pp	2.7pp
5 percent	4.5pp	4.7pp
10 percent	5.8pp	6.1pp
15 percent	6.7pp	7.1pp
20 percent	7.3pp	7.7pp
25 percent	7.8pp	8.2pp
30 percent	8.2pp	8.6pp
40 percent	8.5pp	9.0pp
50 percent	8.6pp	9.0pp

The second outcome is reservation wages for future work, which, between the manager arms is a replication of the original experiment and will only be possible in the algorithm arms if there are still positive rates of perceived of discrimination in some of the algorithm arms. Again focusing on comparing just two arms, the MDE for the effect on a continuous variable is about 0.17sd for either sample size. Instead, pooling the three arms where there will almost certainly be no perceived discrimination (based on the results of the original experiment) and pooling the three arms where there will most likely be positive rates of perceived discrimination between 20-40 percent (based on the results of the original experiment), the MDE is about 0.1sd for either sample size of N=2100 (6 arms of 530 each) or N=1890 (6 arms of 475 each).

2 Outcomes

2.1 Primary outcomes

Perceived discrimination. Right after workers are told about how they were evaluated, they see the following*: "We'd like to know what you think would have needed to be different about your profile for you to be hired to do the proofreading task. For example, would it have helped if you scored higher on the quizzes, or had more education? What do you think?*" and they respond in an open text box. Whether they mention demographics in that free text response is the main measure of explicit perceived discrimination against themselves. This is the same as in the original experiment

Future labor supply. Directly following learning that they were not hired, workers' reservation wages to be evaluated again are elicited. They are told that some workers will have another chance to be evaluated and potentially hired in the future (and that the outcome may be different i.e. they might be hired next time, for example if they are compared to different workers). In one future round, the mechanism used to evaluate workers will be the same as in the experiment (i.e. depends on their treatment assignment). In another, the workers with the highest screening quiz scores will be hired. In both cases, workers are asked at what wages they would be interested in being evaluated again under that hiring mechanism, and that one of the sets of wages will be randomly selected and I use their answer under that wage to determine whether they are evaluated or not (for a randomly selected subset of workers). The wages vary from \$0.10 for each high-quality paragraph (if hired) to \$1.00 in increments of \$0.10, respectively. This is analogous to the original experiment but in the hiring rather than task assignment context.

2.2 Secondary outcomes

The only new secondary outcome is comprehension of the evaluation mechanisms. This was not asked in the original experiment to mask the study purpose, but is asked at the end of this survey in order to understand comprehension, since the results in the original experiment that 20 percent of people perceive discrimination in the algorithm arm cannot separate confusion/inattention from sophisticated thinking about the ability of algorithms to be biased even when they do not use demographics as inputs. Thus, in addition to building in randomization to separate these channels and minimizing the chance for confusion relative to the original experiment, the follow-up also explicitly measures comprehension of the different mechanisms. These questions are incentivized (there are two questions, and 100 randomly-selected workers can earn a \$0.25 bonus for each question they get right). The questions ask:

1. What information did [your manager/the algorithm] have when deciding who to hire?
2. [In the manager arms] What was the basis of your manager's bonus?
3. [In the algorithm arms] What do you know about how the algorithm was designed?

Each question is a "select all that apply" multiple choice question with decoy options.

The remainder of the secondary outcomes are a subset of the secondary outcomes in the original experiment:

- *Secondary measures of perceived discrimination:* https://www.socialscienceregistry.org/trials/9592
- *Beliefs about the likelihood of being hired in the future*
- *Self-efficacy*
- *Affective well-being*

3 Empirical specifications

3.1 Solving selection issues

As described above and in the original experiment's pre-analysis plan, workers are randomly assigned to be evaluated by different mechanisms. The hiring decision made by one randomly-assigned mechanism is then implemented for the worker, and those who are not hired by their randomly-assigned mechanism (~98 percent of the original sample) make up the experimental sample of interest. To eliminate bias caused by differential selection (the mechanisms might hire different types of workers, introducing differences between groups other than their perceptions of discrimination) I will restrict the sample to the workers who are not hired *by all four mechanisms.* I can also include extensive controls for all worker characteristics that the algorithm/managers know in the full sample.

I will present results for both methods. In the original experiment, the mechanisms do not differentially assign certain types of workers to the hard versus easy task, so I anticipate that this will continue to not be an issue in the follow-up study. Results in the main experiment are similar using either method.

3.2 ITT: Effects of hiring mechanisms

The main results will focus on the intent-to-treat (ITT) effect of treatment assignment on perceptions of discrimination and mirror the original experiment. I will control for indicators for age categories, race, gender, education level received, annual household income categories, and employment status (all measured during the screening survey), as well as measures of screening task performance. As in the original experiment, I will show that the results are robust to lasso-selected controls. Recall that randomization occurs conditional on workers quiz score quintile groups; I

also control for fixed effects for each group (quintile 1-2, 2-3, 3-4, or 4-5).[2] Standard errors will be clustered by the "assignment groups" of 40 (the level at which treatment is assigned), plus further pooling the groups that were assigned the same treatment status within each "evaluation group," since these groups of 80 will have seen the same information about historical workers and/or their manager.

3.3 2SLS: Effects of perceived discrimination

Again, as in the original experiment, I will instrument for perceived discrimination with treatment assignment to look at effects on labor supply.

4 Hypotheses

I will test three main hypotheses:

1. What drives the result in the original experiment that 20 percent of workers perceive discrimination when they see that the algorithm historically promoted mostly white men, even if they know the algorithm did not know race and gender? Specifically, is it confusion/inattention, or a sophisticated understanding of algorithmic bias?
   (a) Does this persist when the information the algorithm used is more salient?
   (b) Do workers draw similar or different conclusions when they know a manager who didn't know demographics mostly hired white men in the past?
   (c) Do workers understand and remember the details of the evaluation mechanisms?
2. How do workers perceive an algorithm that used demographic information?
3. Do the effects of perceived manager discrimination in the original study on future labor supply replicate? What is the effect of perceived algorithmic discrimination on future labor supply?

[2]This is a deviation from the original pre-analysis plan but will match the final analysis of the original experiment, though results are nearly identical if including the pre-registered fixed effects for "evaluation groups" (randomly-assigned groups of 120 in the main experiment, conditional on these quiz score quintile groups) instead. This change allows me to include the same controls in the manager and algorithm regressions in the analysis of the original experiment.

5 Update to Pre-Analysis Plan, 9-11-2023

While participant screening was underway (but before any worker had been randomized to a treatment group or started the experimental survey) this pre-analysis plan was updated to reflect a change in the study design. Namely, the original design included treatment arms that provided explicit information about the algorithm design being unbiased, which have been removed from the experiment. This simplified the experiment from 9 arms to 6, which will improve power in the arms that were included specifically to better understand the results of the main experiment, rather than introducing a new research question about how workers respond to information about algorithm design. The arms that provided information about the algorithm design being unbiased would also have been particularly difficult to extrapolate from an online lab-style experiment to the field, since the person or organization providing the information about unbiasedness and their trustworthiness would likely have a large role in its effects. This change created some logistical differences in worker randomization and the study is now better-powered, but had no bearing on answering the main research questions outlined in the original pre-analysis plan.