Reducing perceptions of discrimination: Pre-analysis Plan

Abstract

This pre-analysis plan documents the intended analysis for a randomized experiment examining how individuals perceive discrimination under three job assignment mechanisms (with varying potential to discriminate) and the effects of the two mechanisms that reduce the scope for discrimination from the status quo on perceived discrimination, effort, performance, future labor supply (reservation wages), and cooperation with and reciprocity towards managers. The study design and randomization ensure that the only differences between the three groups is what they believe about the potential for discrimination in the job assignment process. This plan outlines the study design and hypotheses, outcomes of interest, and empirical specifications.

1 Study design

This project seeks to understand how people form their beliefs that they have been discriminated against, and the effect of those beliefs on worker experiences and performance. To do this, I build a dynamic labor market on Prolific and set up an experiment that varies whether workers are assigned to a harder, higher-paying task or an easier, lower-paying task by three different mechanisms that vary in the degree to which they can discriminate. Two strategies allow me to account for systematic differences in which types of workers each mechanism assigns to the harder task.

1.1 The jobs

In this setting, workers will be hired to proofread and summarize paragraphs from scientific texts, each 50-100 words long. There are two versions of the task: the harder, higher-paying task is to

proofread paragraphs from articles published in leading scientific journals *and* to write a concise, clear, accurate summary of the paragraph, and the easier, lower-paying task is to proofread paragraphs from articles from scientific material geared towards children (e.g. the *Science Journal for Kids)* that target an elementary or middle-school reading level.

Three screening tasks will be used to generate predictive information on workers' proofreading and summarizing skills: a spelling quiz of common scientific words, a grammar quiz, and a general-knowledge science quiz. Scores from these screening tasks will be used to determine workers' later assignments.

1.2 The job assignment mechanisms

There are three job assignment mechanisms of interest. First, the status-quo mechanism: 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 assign the worker to the harder or easier task. Second, a mechanism of growing prevalence in the labor market: a screening algorithm that predicts workers' performance at the harder proofreading task based on their average score on the screening tasks and their education. Finally, the third mechanism is a demographic-blinded manager who only knows a worker's average screening task score and education when making the decision about whether to assign the worker to the harder or easier task.

The manager who knows demographics, the status quo mechanism, is the most likely to discriminate. Using instead a manager who does not know demographics is the cleanest manipulation of perceived discrimination: discrimination based on race or gender is, by definition, impossible, but the other human elements of the status quo mechanism are preserved. The algorithm mechanism is of interest because it does reduce the likelihood of discrimination, but it is not clear if workers will perceive that that is so. Furthermore, there are other aspects of the algorithmic assignment mechanism that could change worker beliefs and behavior, like perceived fairness more generally, their dislike of AI, and so on. To untangle all of this, this experiment will also randomize whether workers have demographic information about who the algorithm assigned to the hard task in the past when the historical sample of workers is almost entirely white men. All workers will know that the algorithm only uses quiz scores and education to predict skill at the harder task. The sub-randomization allows for there to be differences in perceived discrimination that hold those other factors of algorithms fixed. Algorithms are quickly proliferating in the labor market in these types of decision-making scenarios and a large academic literature documents that algo-

rithms can either exacerbate or improve on human biases (Cowgill, 2020; Raghavan et al., 2020; Kleinberg et al., 2018) but how workers perceive algorithmic discrimination is yet unknown. The algorithm is trained on data from a pilot study in which workers, recruited from MTurk, completed the screening tasks as well as the proofreading tasks.

Managers for this labor market will also be hired from Prolific and will evaluate groups of workers. In the main experiment, managers will evaluate groups of approximately 40 workers. The group will be relatively homogenous in terms of education and quiz scores, so that the manager decision is more difficult. Managers will evaluate workers in two stages. First, they will see 5 groups of 8 workers that have been randomly grouped together. They will choose one worker from each of those sets to advance to the next round. In the second round, they will choose 3 of their 5 top workers to do the hard task. They do this for three separate sets of workers (so, 120 workers total).[1] Managers will be paid a small completion fee, but their remuneration will largely be based on the performance of the workers that they choose to assign to the harder task.

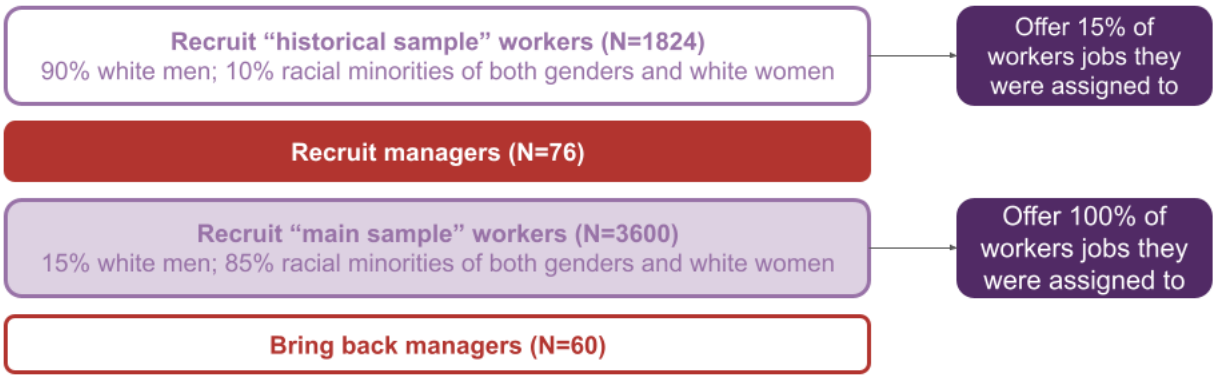
1.3 Generating historical data on manager and algorithm choices

Part of the experimental design, described in the next section, includes showing workers the profiles of other workers who were assigned to the hard task by their manager/the algorithm (depending on their treatment assignment) in a previous round. This "historical sample" round is designed to be more representative of STEM or other industries where white men are particularly over-represented. Generally, the study procedures for the historical sample will be the same as for the main sample; any differences will be explained in the next section.

Figure 1 shows the timeline for recruiting the historical sample workers, the managers, and the main sample workers, and depicts the main differences between the historical sample workers and the main sample workers. The first difference is compositional: workers in the historical sample will be primarily white men whereas workers in the main sample will be primarily not white men. Each sample serves a different purpose: workers in the historical sample are evaluated by one manager and the algorithm (the managers' decisions are implemented) in order to generate data on which workers get assigned to the harder task by the managers and the algorithm in a white-male-dominated field. In the main experiment, workers are told about these decisions that their manager/the algorithm (depending on their treatment assignment) made in a previous round. On the other hand, the main experimental sample over-samples racial minorities and white women

[1] In the "historical sample," described in the next section, managers will evaluate three groups of eight workers and choose one in each group to be assigned to the hard task.

Figure 1: Recruitment timeline



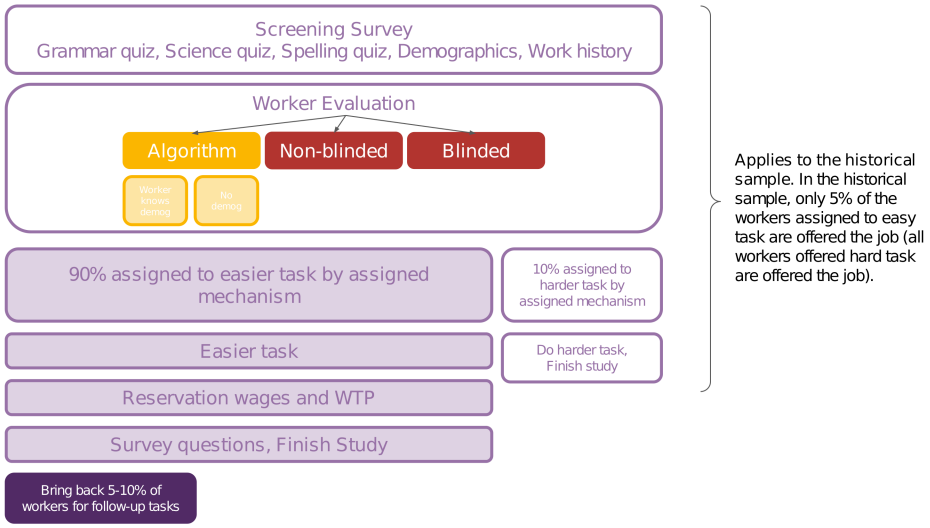
because its purpose is to understand the effects of perceived discrimination. To do that, this sample must be primarily from groups that are historically and currently under-represented and likely to be discriminated against in STEM.

The second difference is logistical. Among the historical sample, only 15 percent of workers will actually be offered the job that they were assigned to by their manager or the algorithm. This is to minimize unnecessary costs because they have served their purpose in the experiment already without doing the main experimental tasks. Workers know that they only have a chance of being offered the later task when the consent to participate in the historical sample. Among the main study sample, all workers will be offered the job that they were assigned to by their manager or the algorithm, because this is where their outcomes on the main experimental tasks are measured.

Note that the same managers evaluate the historical sample and main sample (with some over-sampling initially to account for up to 20 percent attrition among managers between their two rounds). This is so that workers in the main experiment can be told about the exact decisions their manager made in a previous round. However, as shown in Figure 1, the ratio of workers to managers is different in the two samples. When evaluating the historical sample, managers will each evaluate 24 workers (3 groups of 8) and choose one per group to do the harder task. When evaluating the main study sample, managers will evaluate 40 workers (5 groups of 8). They will advance one per group to a second round where they will choose 3 out of the top 5 to assign to the harder task.[2]

[2]They will actually evaluate 3 sets of 40 in this way, but their decisions will be implemented for one randomly chosen set of 40. More on this in Section 1.5.

Figure 2: Experimental design



1.4 Sampling strategy

The study will take place on the online survey platform Prolific where researchers recruit study participants. As described above, the main experiment will over-sample workers from minority groups to allow for comparisons across minority racial groups and to focus on workers who are more likely to experience discrimination: the sample will be approximately 33 percent white and non-Hispanic, 33 percent Black, 16 percent white and Hispanic, and 16 percent Asian. Within each racial category, approximately 50 percent of the participants will identify as women. Participants in the historical sample will be 90 percent white men, and the remaining 10 percent will be non-white men or women or white women (and otherwise unconstrained). This is to create a realistic setting in which most of the workers who have been historically promoted to higher positions are white men. Participants who play the role of managers will be required to be employed outside of Prolific and to be white men. This is to mimic a STEM field where most managers are still white men, and to preserve the relatively low population of racial minorities on Prolific to participate as workers in the study. All of this demographic selection is feasible using Prolific's pre-screening system.

1.5 Experimental design

The screening survey. Workers will be recruited with a "screening survey," common on Prolific, that qualifies the worker to participate in future high-paying surveys. In this screening survey, their first session, workers will complete the three screening tasks, and answer questions about

their demographics and job history. After all workers have completed the screening survey, their scores and demographics will be aggregated into worker profiles. As described above, workers will be grouped into groups with similar education levels and average quiz scores to be evaluated and assigned to the harder or easier task. Quiz scores are shown as 1-5 stars on a worker profile that are approximate quintiles.[3]

Manager randomization. In the historical sample, each worker profile will be evaluated by one manager. Managers will be randomly assigned to evaluate workers with or without demographic information in their profile (profiles always include average test score stars and education). They will get the same type of information about their workers when they return to evaluate the main study sample. They will also be randomly assigned to evaluate workers of a certain average quiz score level (1-2 stars, 2-3 stars, 3-4 stars, 4-5 stars) and will evaluate workers of the same average quiz score level when they return to evaluate the main study sample.

After managers return to evaluate the main study sample workers, they will be randomly paired with another manager who has the opposite assignment as them (receiving demographics when they don't, or vice versa) and was assigned to evaluate workers of the same average quiz score level. These manager pairs will evaluate the same 120 workers (3 sets of 40, say set A, B, and C). All three sets will also be evaluated by the algorithm. Managers know that their decisions will be implemented for one of the three sets and that the performance of the workers they assign to the harder task in that set will determine their bonus payment.

Worker randomization. Then, each set A, B, and C will be randomly matched with one of the three job assignment mechanisms, without replacement. For example, set B could be assigned to jobs following the decisions of the demographic-blind manager, set A could be assigned to jobs following the decisions of the manager with access to demographics, and set C could be assigned to jobs following the decisions of the algorithm. Workers are only told about the mechanism that determined their assignment. The assignments by the other mechanisms only generate data to be used in the analysis of the experiment (see Section 3).

Work task and survey. After all workers have been evaluated (see Section 1.2 for a description of how workers are assigned), 2-3 weeks after workers complete the screening task, any worker who is assigned to the harder task by their manager or the algorithm (depending on their treatment

[3]Because of the rightward skew in test scores, actual quintiles provide unintuitive cutoffs. To make homogenous groups that only vary by one star, we also need fewer participants who receive 1 or 5 stars than the other numbers of stars. So, cutoffs were chosen such that approximately 12 percent of a pilot sample (281 participants) received 1 star (scored less than 40 percent on average), 20 percent received 2 stars (scored between 40-55 percent), 30 percent received 3 stars (scored between 55 and 70 percent), 30 percent received 4 stars (scored between 70 and 85 percent), and 10 percent received 5 stars (scored more than 85 percent). Assuming the distribution is similar in the full experiment, I will use the same cutoffs. Otherwise, I will use the same percentiles to create new cutoffs.

assignment) will be offered the harder task. They will do the harder task and finish the experiment (about 8 percent of workers).

At the same time, any worker who is assigned to the easier task by their manager or the algorithm (depending on their treatment assignment) will be offered the easier task. Among this sample of interest, after agreeing to take the follow-up survey, workers will be told about how they were assigned to the easier, lower-paying task. If they were assigned by a manager, they will see some demographic characteristics of their manager. All workers will see three profiles of the workers that their manager/the algorithm (depending on their treatment assignment) assigned to the harder task in a previous round (i.e. the historical sample). If their manager knew their and other workers' demographics, those demographics will be included in the profiles along with quiz scores and education. If their manager did not know anyone's demographics, the profiles will only include test score stars and education. Workers assigned to be evaluated by the algorithm will be randomized into two sub-groups: those who are shown the race/gender/avatar of those previously assigned by the algorithm to the hard task, and those who are not. All of them will know that the algorithm only *used* data on education and quiz scores, and will see that data. Comparisons of these two groups isolates a difference in perceived discrimination within the algorithm arm, and provides data on how common perceived discrimination might be even in a setting where a worker knows that an algorithm was fair, just from seeing that historically the algorithm has hired or promoted (almost) entirely white men.

One-third of the sample will be randomized into the demographic-blind manager arm, another one-third will be randomized into the arm where managers knew demographics, one-sixth will be assigned by the algorithm and shown the race and gender of those who were previously assigned to the hard task, and another one-sixth will be assigned by the algorithm and not shown the race and gender of those who were previously assigned to the hard task.

After they are told about how they were assigned, workers are asked a free-response question about what they think would have needed to be different about their profile in order to be assigned to the harder task, and then are asked how many stars they think they would have needed to score on the screening quizzes in order to be assigned to the harder task by their manager.

Then, workers will do the easier proofreading task. Workers will know that they have to proofread at least six paragraphs to receive their completion payment and that they are able to proofread up to eighteen paragraphs (each for a bonus). They know that they will be eligible to be evaluated again to do the harder task for a higher wage in a future survey (though they could also be assigned again to the easier task). After either the first, third, or fifth paragraph (randomly assigned), workers will complete an affective well-being scale.

After finishing the easier proofreading task, workers' reservation wages to do this work again will be elicited, first under the assumption that the assignment 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 offered the harder job (a cutoff rule). One set of wages will be selected and approximately five percent of workers will be randomly selected to have their choices implemented, for each assignment type.

Next, workers in a manager arms will be asked at what wage they would want to work together with their manager on a similar but collaborative (instead of hierarchical) task in the future, how much they would be willing to give up in wages to be able to choose their own manager (instead of a default of working with the same manager who assigned them in the main experiment), and how they would share a thank-you bonus with their manager. Each of these choices will be implemented for a randomly selected subset of participants.

Then, workers will answer questions about their self-efficacy to do the easier or harder job, job satisfaction, complaints about the promotion process, incentivized questions about whether they think workers in each race*gender group were over-represented, under-represented, or neither, conditional on their quiz scores, among workers assigned to the harder task (measures of the perceived existence of discrimination), and whether they think they would have been assigned to the harder task if they were assigned by the same mechanism but had a different race or gender (explicit measures of perceived discrimination). After they answer this question, they will be asked to imagine that they are a worker with a different (fictitious) profile with randomly assigned characteristics, and asked how many stars they think they would have needed to score on the screening quizzes in order to be assigned to the harder task. Differences between these answers for fictitious workers of different races and genders provides a measure of implicit perceived discrimination.

1.6 Sample size and statistical power

All calculations assume power of 80 percent and a significance level of 0.95. I will recruit 3,600 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 80 percent of workers complete the follow-up survey and 92.5 percent of workers are assigned to the easier task, the final analysis sample will be around 2,664 workers assigned to the easier task by their randomly assigned mechanism or around 2,304 workers assigned to the easy task by *any* of the three mechanisms (assuming 20 percent of workers are assigned to the hard task by at least one mechanism); see Section 3 for a discussion of why the analysis might

use either of these samples.

1.6.1 First stage

Main effects. Regressions to test whether the demographic-blinded manager reduces perceived discrimination relative to the manager who knows demographics are powered to detect effects larger than 6.5 or 7 percentage points (for sample sizes of 2,664 or 2,304, respectively). Similarly, testing whether one of the algorithm sub-groups reduces perceived discrimination relative to the manager who knows demographics is powered to detect effects larger than 8 percentage points for either sample size.[4] Given the results from a pilot study, the effect sizes are expected to be larger than these MDEs.

Heterogeneity. Treatment effect heterogeneity is powered as follows: tests of whether the effect of the algorithm depends on whether the worker knows the race of the historically assigned workers are powered to detect differences larger than 9.5 or 10 percentage points (for sample sizes of 2,664 or 2,304, respectively). Tests of whether the effect of the demographic-blind human differs from one algorithm sub-group are powered to detect differences of 8 percentage points and tests that the effect of the demographic-blind human differs from both algorithm subgroups (which are pooled and don't differ from each other) are powered to detect differences larger than 6.5 or 7 percentage points (for sample sizes of 2,664 or 2,304, respectively).

Racial and gender heterogeneity is powered as follows: when testing for heterogeneity in the effects of the blinded manager, gender heterogeneity among non-white participants and racial heterogeneity among men are powered to detect differences in the treatment effect of 15 percentage points, gender heterogeneity among white participants is powered to detect differences in the treatment effect of 18 percentage points, and racial heterogeneity among women is powered to detect differences in the treatment effect of 20 percentage points. For each group, testing for heterogeneity in the effects of the algorithm are powered to detect MDEs about 3 percentage points larger than the MDEs for differences in the effects of the blinded manager. These MDEs come from simulations with a sample size of 2,664; with a sample size of 2,304 each MDE is about 1 percentage point larger.

[4]The percent of the population perceiving discrimination in the status quo group is assumed to match the rate of perceived discrimination in each race*gender cell in a pilot study describing the main experiment as a hypothetical scenario. 10 (40) percent of white, non-Hispanic men (women) are expected to perceive discrimination in the status quo group, as compared to 25 (35) percent of white, Hispanic men (women), 45 (70) percent of Black men (women), and 40 (70) percent of Asian men (women).

1.6.2 Reduced form

Regressions to test the effects of the demographic-blind manager on the binary measures of retention (completing only the minimum 6 paragraphs, or completing all 18 paragraphs) are powered to detect effects larger than 5 or 5.5 percentage points (for sample sizes of 2,664 or 2,304, respectively), and tests of the effect of one algorithm sub-group are powered to detect effects larger than 6 or 6.5 percentage points (for sample sizes of 2,664 or 2,304, respectively). In pilot data, 12 percent of workers completed only 6 paragraphs and 68 percent complete all 18 paragraphs, which is assumed in these calculations.

All other outcomes are continuous. Regressions to test the effects of the demographic-blind manager are powered to detect effects larger than 0.14sd or 0.15sd (for sample sizes of 2,664 or 2,304, respectively), and tests of the effect of one algorithm sub-group are powered to detect effects larger than 0.17sd (for either sample size).

1.6.3 Two-stage least squares

The two-stage least squares power calculations assume that the effects of the treatments on perceived discrimination are quite large, effectively taking the rate of perceived discrimination to zero in the demographic-blind manager group and both algorithm sub-groups. This is consistent with piloting (though in very small samples).

Then, two-stage-least-squares regressions are powered to detect effects of reducing perceived discrimination that are larger than 4 or 5 percentage points on the binary outcomes (for sample sizes of 2,664 or 2,304, respectively), and effects larger than 0.12sd on the continuous outcomes (for either sample size).

2 Outcomes

2.1 Primary outcomes

Perceived discrimination. The survey will measure perceived discrimination against oneself implicitly and explicitly, and perceived discrimination against different groups in general.

- **Explicit perceived discrimination against oneself:** Right after workers are told about how they were assigned to the easier task, 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 assigned to the higher-paying, harder 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 race, gender, bias, or discrimination in that free text response is the main measure of explicit perceived discrimination against themselves.

- **Implicit perceived discrimination against oneself:** Following the free-response question about what needed to be different about their profile to be assigned the harder task, workers will answer questions that provide an implicit measure of perceived discrimination. Workers will be asked to how many stars they think they would have needed to score on the screening quizzes in order to be assigned to the harder task by their manager or the algorithm. Differences across workers of different races and genders but with the same number of stars and education level is the (cross-sectional) main measure of implicit perceived discrimination.

- **Perceptions of discrimination in general:** Finally, at the very end of the survey (but before the secondary measures of perceived discrimination described below) workers are asked whether they think each race*gender group is under-represented, over-represented, or neither among workers assigned to the hard task, conditional on their screening quiz scores. These terms are defined and workers answer for eight race*gender groups. To incentivize careful responses, they are told that one group will be randomly selected and they will earn a small bonus if they are correct. Workers' believing that their own group is under-represented is the final main measure of perceived discrimination.

For both the implicit and explicit measures of discrimination, the main analysis will use an indicator for whether a worker perceived either race *or* gender discrimination. Secondary analysis will separately analyze gender versus racial discrimination, especially when considering heterogeneous effects by race and gender.

Future labor supply. Directly following their completion of the proofreading task, workers' reservation wages to be evaluated again are elicited. They are told that some workers will have another chance to be evaluated and assigned to more proofreading tasks in the future. In one future round, the mechanism used to assign people to the two tasks 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 assigned to the harder task. In both cases, workers are asked at what wages they would be interested in being evaluated again under that task assignment 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 offered the job at that wage. The wages vary from \$0.05 per each high-quality

easy paragraph and $0.10 for each high-quality hard paragraph to $0.50 and $1.00 in increments of $0.05 and $0.10, respectively.

Later in the survey, workers are also asked to indicate how much they agree or disagree with the following on a scale of 1 (strongly disagree) to 5 (strongly agree):

- I want to complete more surveys for this Requester
- I want to complete more surveys with tasks that are assigned in this way
- I want to complete more proofreading tasks of this level of difficulty
- I want to complete more difficult proofreading tasks

This will be used to understand the rationale behind any effect on the main, incentivized measure of future labor supply.

Intensive and extensive effort and performance. The real effort task provides three measures of worker effort and performance. These effort tasks reward high-quality work and are incentive-compatible. The three components can be combined into an index (by standardizing and averaging the five continuous measures), which will be the main outcome of interest, or considered independently and adjusted for multiple hypothesis testing (including an indicator for finishing all 18 paragraphs).

1. **Retention.** Retention, or extensive effort, is observable through the number of paragraphs that each worker chooses to proofread and whether they choose to proofread all available paragraphs. They are required to proofread six paragraphs to receive their participation payment but can do up to eighteen, with potential bonuses available for each completed paragraph that they do a good job proofreading (above-median number of mistakes correctly highlighted minus number of non-mistakes incorrectly highlighted, though workers are not told the explicit rule for how "a good job" is determined). After each paragraph from the sixth onwards, they are asked if they would like to continue with another paragraph or continue to the final survey questions. Like an offline job, would be costly for the employer to have high turnover because workers improve at the task as they gain experience.
2. **Performance.** The main measure of performance is the number of mistakes that a worker highlights correctly minus the number of non-mistakes highlighted incorrectly. Each component is also of interest on its own, since workers may make different types of mistakes in different emotional states.

3. **Effort.** Conditional on performance, the proofreading task yields two measures of intensive effort: the number of seconds a worker spent on each paragraph (each paragraph is limited to 60 seconds), and the number of times they clicked on each paragraph (i.e. highlighting or unhighlighting a word).

For both the effort and performance variables, I will separately consider total effort and performance over all 18 paragraphs and effort and performance in the first 6 (required) paragraphs, since treatment effects on retention may affect the measures of total effort and performance.

2.2 Secondary outcomes

Secondary measures of perceived discrimination. Several additional measures of implicit and explicit perceived discrimination that will be collected and used to test robustness, as there is no standard way to measure experienced discrimination. In a pilot, these measurse are highly correlated with each other and with the main measures described above. In particular:

- **Explicit perceived discrimination against oneself:** At the very end of the survey, workers are asked several questions that provide additional explicit measures of perceived discrimination. First, whether they have any concerns about the way they were assigned to proofread easier paragraphs rather than hard paragraphs (yes or no). This follows a placebo question asking them about whether they would have preferred the hard task, which also reminds them of how they were assigned to the easier task. If they respond yes, they are able to type their complaint into a text box. Whether they mention potential bias or discrimination in their complaint is the most natural possible measure of perceived discrimination. However, the (psychological or perceived monetary) costs to reporting discrimination may be too high for this measure to capture true perceived discrimination. Thus, several questions will ask explicitly about differential treatment, though they won't use the word discrimination:
  - Do you that think you would have been assigned to the harder task if your gender was different?
  - Do you think that you would have been assigned to the harder task if your race or ethnicity was different?
- **Implicit perceived discrimination against oneself**: In the last questions in the survey, workers will be asked to imagine that they are a worker with a different (fictitious) profile with randomly assigned characteristics, and asked how many stars they think that worker

would have needed to score on the screening quizzes in order to be assigned to the harder task. Differences between these answers for fictitious workers of different races and genders provides a measure of implicit perceived discrimination that can be constructed at the individual level. Workers will do this four times – once where the fictitious profile has the same race as themself but a different gender, twice where the fictitious profile has the same gender but a different race,[5] and again where the fictitious profile has the same race and gender. The other characteristics (income, married, kids, education) are randomly selected among those similar to the worker's own. Differences between the number of stars they think a worker would need and the number they think they would need, or the number they think the fictitious profile with the same race and gender as themselves would need, is an individual-level measure of implicit perceived discrimination.

- **Perceptions of discrimination in general:** As described above, workers are asked whether they think each race*gender group is under-represented, over-represented, or neither among workers assigned to the hard task, conditional on their screening quiz scores. A secondary measure of perceived general discrimination for non-white, non-men is believing that white men are over-represented.

These questions that are more obviously about discrimination than the main measures above are asked *last* in the survey flow in order to minimize the chance that workers learn the topic of the study and change their behavior accordingly. The survey questions are also asked after the effort tasks have been completed to minimize demand effects on those outcomes. A variety of related filler questions surround these questions to try to disguise the purpose of the study. To understand whether (and at what point) workers figure out the purpose of the study (which could affect results, if workers are more likely to figure it out in one of the treatment groups where they see workers' race and gender), workers are asked what they think the purpose of the study is once before the survey questions (after the effort tasks and reservation wage/willingness to pay elicitations) and again after all survey questions have been answered.

Beliefs about future discrimination. One way that perceived discrimination may affect performance and labor supply is through affecting workers' beliefs about the likelihood that they are discriminated against in the future (which changes the return to effort and labor market participation). After eliciting their reservation wages to be evaluated again by the mechanism they were

[5] In particular, in the first profile, Black, Asian, and Hispanic participants will see a White fictional profile, and White participants will see a Black fictional profile. In the second profile, Black participants will see a Hispanic fictional profile, White participants will see an Asian fictional profile, and Asian and Hispanic participants will see a Black fictional profile.

assigned to in the main experiment or using a cutoff rule and screening quiz scores, workers are asked what they think is the probability that they will be assigned to the harder task in each case. Conditional on their screening quiz scores, this probability identifies beliefs about future discrimination without expressly asking about the likelihood of future discrimination (which would be too revealing to ask).

Cooperation with and reciprocity towards managers. Another way that perceived discrimination may affect performance and labor supply is through willingness to work for or with one's manager. Workers who believe they have a discriminatory manager may not want to be as productive for that manager, or may just want to minimize interactions with them. To understand this mechanism, in the manager arms (67 percent of the sample), workers will be asked to make three incentive-compatible decisions related to working with their manager. They know that each of these choices will be implemented for some random subset of the sample. The three measures can be combined into an index (by standardizing the three measures), or considered independently and adjusted for multiple hypothesis testing.

1. **Reservation wage** to work cooperatively with the same manager as in the main experiment in a cooperative rather than hierarchical task. The task is for workers to summarize complicated scientific paragraphs. Their manager will review the summaries, leave comments, and choose a bonus payment for the worker. Workers will have a chance to revise their work and be paid a base payment per high-quality summary, in addition to the bonus determined by the manager. Workers are asked **at what base payment per paragraph** they would be willing to accept this work, for each multiple of 0.05 between $0.05 and $1.00. I will randomly choose one of these prices, and of all the workers who said they would want the job at that price, twenty workers will be randomly selected to do the work (with the same or a similar manager they had in the initial round of the experiment).

2. **Willingness to pay** to be able to choose your own manager (relative to the default of working with the same manager as in the main experiment). Workers are told to assume that the same task as above will pay $1.00 per high-quality summary and that they are interested in the task at this wage. They are then asked if they would want to keep their same manager or pay to choose their own manager from a group of five, for each price that is a multiple of 0.05 between $0.05 and $1.00 *per summary*. I will randomly choose twenty workers who wanted the job if it paid a base rate of $1.00 per paragraph, and randomly choose one of these prices, and implement worker choices over paying to choose a new manager or working with their old manager.

3. **Reciprocity** towards managers. Workers are told that 20 workers will be randomly selected to receive a $20 thank-you bonus for their participation in the study. If they are chosen, they have the option to allocate some of their thank-you bonus to the manager that assigned them to the easy task. All workers are asked how much they would allocate if they are chosen. Twenty workers will be randomly selected to receive the thank-you bonus and have their choice implemented.

Self-efficacy. A participant's work self-efficacy is a key psychological channel through which perceived discrimination could affect labor supply and performance. Work self-efficacy is one's confidence in their ability to do the tasks required of them in a particular job. To assess participants' work self-efficacy about the tasks at hand, workers are asked how much they agree or disagree with the following statements on a Likert scale from 1 (strongly disagree) to 5 (strongly agree):

- I am capable of doing the harder proofreading job well
- I would have liked a chance to do the harder proofreading task
- I am confident in my ability to work under pressure
- I am capable of doing the easier proofreading job well
- I did a good job on the proofreading task today
- I was able to improve as I proofread more paragraphs

And to understand participants' self-efficacy related to the underlying skills they posses, they are asked to indicate their skill level in the following areas from 1 (not at all skilled) to 5 (completely skilled):

- Written communication
- Oral communication
- Problem solving
- Numeracy
- Motivation
- Learning new material

*The most commonly used psychometric scale to assess work self-efficacy is the Work Self-Efficacy Inventory (WS-EI; Raelin (2008)), however, the WS-EI was not adaptable to the Prolific context.

Affective well-being. Affective well-being is another psychological channel through which perceived discrimination might affect productivity and effort. Early in the proofreading task (randomized to be after the 1st, 3rd, or 5th paragraph), workers are asked to indicate to what extent they feel each of the following emotions *right now,* on a scale from 1 (not at all) to 6 (very much): happy, at ease, anxious, annoyed, motivated, calm, tired, bored, gloomy, active. This is the standalone short-form 10-item Daniels five-factor measure of affective well-being (D-FAW; Russell and Daniels, 2018). Mixed in with the standard items (in random order), they are also asked how discouraged and upset they are in order to validate responses to how motivated and at ease they are, since these are some of the key hypothesized emotional states that may be affected by perceived discrimination (the direction of the effect on motivation/discouragement may go either way, but perceived discrimination is expected to make workers less at ease and more upset).

Job satisfaction. Survey measures of job satisfaction have been related to measures of perceived discrimination in past work (Mukerjee, 2014), so they are included here as well as supplemental outcomes. The measure of job satisfaction uses the phrasing of standard job satisfaction surveys, and asks workers how satisfied they are, on a scale of 1 (very dissatisfied) to 4 (very satisfied), with the following:

- How satisfied are you on the whole with the work that you were offered in this survey?
- How satisfied are you on the whole with the work that is available on Prolific?
- How satisfied are you on the whole with the work that you do outside of Prolific (if applicable)?

3 Empirical specifications

3.1 Solving selection issues

As described above, workers are randomly assigned to be evaluated by a manager without access to demographics, a manager with access to demographics, and an algorithm. The decision made by that job assignment mechanism is then implemented for the worker, and those who are assigned to the easy task (~92 percent of the original sample) make up the experimental sample of interest.

There are two possible ways that the job assignment mechanisms could change average worker outcomes among the workers randomly assigned to be evaluated by them:

1. The mechanisms have different potential to discriminate against workers, and thus may effect perceptions of discrimination (the effect of interest)

2. The three mechanisms could assign different types of workers to the hard versus easy tasks (a "selection effect")

In order to isolate the effect of interest, the analysis needs to account for the possible selection effect. Luckily, this is easier than dealing with a similar type of selection effect in observational data. I will use two methods to do so:

3.1.1 Controlling for observed differences

All of the regressions outlined in the next section include controls for race, gender, education, and screening quiz scores. I can additionally add controls for the hairstyle, hair color, and skin color the participant selected for their avatar. Unlike in the real world, we *know* that this is all (or more) information that the manager or algorithm knew about each worker when they decided on who to assign to the harder task. Including these controls should account for any systematic differences in what types of workers the different mechanisms assigned, and this can be tested with balance tests on all of the other demographic and screening quiz score data that we observe on the participants.

The benefit of this approach is that the full sample of workers who do the easy task can be included in the sample for analysis (expected to be about 2664 workers, if 80 percent of the sample who does the screening survey returns for the proofreading tasks). The downside is that there may be some residual differences in the types of workers that each mechanism assigns to the hard task, though this is unlikely.

3.1.2 Restricting to those assigned to the easy task by all three mechanisms

A second approach to dealing with these selection concerns has been baked into the experiment design. All workers are randomly assigned to be evaluated by one mechanism, and that mechanism's decisions are implemented for that worker. However, the evaluation system has been designed such that every worker is actually evaluated by *all* three mechanisms (described in Section 1.5) and then randomly assigned to have one of those mechanism's decisions implemented.

Thus, to remove selection concerns, another approach is to restrict the sample to those who *would have been assigned to the easy task by all three mechanisms.* Again, we can test for balance among this sample.

The benefit of this approach is that these workers were all evaluated the same way by all three mechanisms, so any difference in their outcomes must be due to how they reacted to learning how they were assigned by the mechanism they were randomly assigned to. In other words, it eliminates the selection effect concern completely. The downside is that this is a smaller sample, since the decision-making of the three mechanisms is slightly but not strongly correlated. Assuming that 80 percent of the sample is assigned to the easy task by all three mechanisms (i.e. that the mechanisms' decisions are slightly but not strongly correlated), this is expected to be about 2304 workers.

The outcomes of the balance tests and the degree to which precision is a concern for the results will determine which of these two approaches is the primary one to be used in the paper.

3.2 ITT: Effects of job assignment mechanisms

The main results will focus on the intent-to-treat (ITT) effect of using job assignment mechanisms with less scope for discrimination on the outcomes of interest.

The following ITT specification will be used to estimate the average effect of the job assignment mechanisms:

$$y_i = \alpha + \beta_1 T_i^{Algo_1} + \beta_2 T_i^{Algo_2} + \beta_3 T_i^{NoDemog} + \Phi X_i + \nu_i + \varepsilon_i \quad (1)$$

Where y_i is the outcome of interest for respondent i. The omitted group is those who were assigned to the easier task by a manager with access to demographics. T^{Algo_1} is an indicator for being assigned by the algorithm and not being shown demographics when seeing historical worker assignments, T^{Algo_2} is an indicator for being assigned by the algorithm and being shown demographics when seeing historical worker assignments, and $T^{NoDemog}$ is an indicator for being assigned by a manager who did not have access to demographics. X is a vector of participant controls, including 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. Recall that workers are evaluated in groups of 40 by the algorithm and managers; ν_i are fixed effects for those groups. β_1, $\beta_2$and β_3 are the coefficients of interest. Standard errors will be clustered by the groups of 40 (the level at which treatment is assigned).

Heterogeneity. I anticipate important heterogeneity by race and gender and by whether a participant reported in the screening survey that they've ever experienced discrimination at work in the past. I will also examine heterogeneity by workers' degree of over-confidence in their ability, as measured as the difference between their (incentivized) guesses of how many they got right and their actual scores on the screening tasks.[6] Each worker's three values of over-confidence on the three screening tasks will be summarized in an index and used for heterogeneity analysis. Finally, I will examine heterogeneity by workers' beliefs about how often workers in their own race*gender group report having experienced discrimination in my sample (an incentivized belief elicitation in the screening survey). Finally, I anticipate heterogeneity by one's beliefs about the prevalence about discrimination against one's group at baseline. A question in the screening survey asks participants to report what they think is the fraction of the sample that reports experiencing discrimination at work in the past for each race*gender demographic group; their beliefs about their own group will be used to understand this dimension of heterogeneity. [7]

3.3 2SLS: Effects of reducing perceived discrimination

A second set of results will restrict to the sample that was assigned to the easier task by a manager, who either had access to demographics or did not. Being evaluated by a manager without access to demographics can be used as an instrument for perceived discrimination in two-stage least squares (2SLS) models that estimate the effect of reducing perceived discrimination. Similarly, restricting to those who where assigned to the easier task by the algorithm, being aware of the demographics of workers previously assigned to the hard task by the algorithm can be used as an instrument for perceived discrimination in analogous 2SLS models.

Assuming that there is a sufficient first stage effect of each of these assignments on perceived discrimination, the effect of reducing perceived discrimination can be recovered by instrumenting for perceived discrimination with treatment assignment, separately for workers who were evaluated by a manager and those evaluated by the algorithm. Among workers evaluated by a manager, being

[6] An un-incentivized question will ask workers to report what they believe to be the probability density function over the support of all possible scores, which will yield a second measure of over- and under-confidence as being whether their scores were outside of their 90-percent confidence interval and a third measure of their certainty, both of which will be used for robustness.

[7] In a broader context, there may also be important heterogeneity by age, since age discrimination is common in the workplace and may be especially salient in an online setting where older adults are particularly under-represented, like Prolific. However, because the sampling stratification by race and gender already yields quite small cells of available participants, I am not able to also stratify by age in a way that creates a large enough sample of older adults. As as result, I have not included age as a characteristic in the worker profiles in order to eliminate this vector of possible (perceived) discrimination.

evaluated by a manager without access to demographic information is expected to reduce perceived discrimination. Among workers evaluated by the algorithm, not knowing the demographics of the workers previously assigned to the hard task by the algorithm is expected to reduce perceived discrimination.

The first stage estimating equation would be estimated separately for these two groups, and can be written as:

$$NPD_i = \alpha + \delta_1 T_i^{LessDiscr^1} + \Phi X_i + \nu_i + \mu_i \tag{2}$$

where NPD_i is a measure of whether an individual **did not** believe that they were discriminated against when assigned to the easier task and all other variables are defined as above. $T_i^{LessDiscr}$ is an indicator for being in the treatment group that is hypothesized to reduce perceived discrimination (i.e. the manager without access to demographics in the manager sample, and not knowing the demographics of the historical workers in the algorithm sample).

The corresponding second stage estimating equation is:

$$y_i = \alpha + \gamma \widehat{NPD_i} + \Phi X_i + \nu_i + \eta_i \tag{3}$$

where $\widehat{NPD_i}$ is the fitted value from the first stage regression.

Heterogeneous treatment effects will be assessed over the same variables as the ITT specification.

4 Hypotheses

I will test two main hypotheses:

1. Whether assignment by a demographic-blind manager or algorithm lead to lower rates of perceived discrimination than the status quo (manager with access to demographics).
    (a) Whether the perceived discrimination associated with assignment by an algorithm depends on whether the algorithm is known to be mainly assigning white men to the harder task, even when all workers know that the algorithm's only inputs are quiz scores and education
2. Whether assignment by a demographic-blind manager or algorithm (and a hypothesized reduction in perceived discrimination) improve future labor supply, effort and performance relative to the status quo

For both questions, I will also test:

1. Whether effects are driven by racial minorities and women, people who report experiencing discrimination at work in the past, people who are over-confident about their abilities, and/or people who think that people in their own demographic group are more likely to report being discriminated against in general.

2. How the effects of the demographic-blind manager and algorithmic assignment compare.

To understand mechanisms, I will examine effects on beliefs about future discrimination, cooperation with and trust in managers, self-efficacy, affective well-being, and job satisfaction.

References

Cowgill, Bo, "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening," *Working Paper*, 2020.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan, "Algorithmic Fairness," *AEA Papers and Proceedings*, May 2018, *108*, 22–27.

Mukerjee, Swati, "Job satisfaction in the United States: Are Blacks still more satisfied?," *The Review of Black Political Economy*, January 2014, *41* (1), 61–81.

Raelin, Joseph, "Validating a new work self-efficacy inventory," in "Unpublished manuscript," Boston, MA: Northeastern University, 2008.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy, "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 2020, pp. 469–481. arXiv: 1906.09208.

Russell, Emma and Kevin Daniels, "Measuring affective well-being at work using short-form scales: Implications for affective structures and participant instructions," *Human Relations*, November 2018, *71* (11), 1478–1507. Publisher: SAGE Publications Ltd.