

A Robust Test of Prejudice for Discrimination Experiments*

Daniel Martin[†] and Philip Marx[‡]

December 11, 2021

Forthcoming in *Management Science*

Abstract

A large experimental literature has been devoted to studying discrimination. An important question for policymakers and firms is what drives the discrimination uncovered by those experiments. However, motivations are hard to determine when decision-makers pay selective attention to information because their learning is private. We overcome this challenge by deriving conditions on average outcomes that reveal decision-makers are prejudiced no matter what they learned about individuals in each demographic group before making their decisions. This provides a test of prejudice that is general, simple, and robust, and that can potentially be used to identify prejudice in a wide range of important settings, such as hiring, consumer lending, and housing access. We demonstrate our test of prejudice using the labor market experiments of Reuben, Sapienza, and Zingales (2014) and Bertrand and Mullainathan (2004).

Key words: Discrimination, prejudice, experiments

JEL codes: D60, D83, D91

*We thank Roland Bénabou, Quoc-Anh Do, Hanming Fang, Laura Gee, Alex Imas, Peter Klibanoff, Nicola Persico, Ernesto Reuben, Sigrid Suetens, our editors, and two anonymous referees for valuable feedback.

[†]Kellogg School of Management, Northwestern University.

[‡]Department of Economics, Louisiana State University.

1 Introduction

Discrimination is a pressing issue for society and the management and regulation of firms, and a large experimental literature has been devoted to its study. Experiments – both in the lab and the field – have provided evidence in a wide range of settings that decisions can change when they are made about members of different demographic groups.¹ For example, Bertrand and Mullainathan (2004) find that hiring managers are less likely to call back candidates with otherwise identical resumes that have traditionally African-American names, and Reuben, Sapienza, and Zingales (2014) find that employers in a lab experiment are less likely to hire a female candidate to complete a task in which females perform equally as well as males.

For policymakers and firms, it is important to know what drives such discrimination. The economic literature distinguishes primarily between preference- and belief-based channels of discrimination. *Prejudice* (taste or preference-based discrimination) occurs when decisions differ across groups because the decision-maker obtains different utility from outcomes depending on group identity (Becker 1957). *Statistical discrimination* (belief-based discrimination) occurs when decisions differ across groups because a decision-maker holds different but correct beliefs about each group (Arrow 1971; Phelps 1972).

Unfortunately, a decision-maker’s motivations can be hard to determine when their learning is *private* (not observable to outsiders) because an analyst cannot directly assess all of the factors that enter into the decision-maker’s choices. For instance, what aspects of a candidate’s appearance factor into a hiring manager’s decision about whether to hire that candidate? Or when quickly scanning a candidate’s resume, what information does a hiring manager extract before deciding whether or not to call back that candidate? This identification challenge is especially pronounced in settings where discrimination is impacted by selective attention that depends on group identity. For example, Bertrand and Mullainathan (2004) note that: “Employers receive so many resumes that they may use quick heuristics in reading these resumes. One such heuristic could be to simply read no further when they see an African-American name.”² Without knowing the information decision-makers have attended to, it becomes even harder to determine what factors entered into the decision-maker’s choices.

We overcome this identification challenge by deriving conditions on outcomes that reveal prejudice regardless of what decision-makers have learned about individuals in each demo-

¹Prominent settings include hiring, consumer lending, and housing access. For recent reviews, see Riach and Rich (2002), Anderson, Fryer, and Holt (2006), Lane (2016), Bertrand and Duflo (2017), Baert (2018), Neumark (2018), and Wozniak and MacNeill (2020).

²Bartoš, Bauer, Chytilová, and Matějka (2016) develop a related model of discrimination based on selective attention and use it to explain empirical disparities in labor and housing markets.

graphic group. These conditions provide a test for prejudice that is *robust* to any form of private learning. The key to this test is to compare outcomes across decisions and groups, which is possible with the right experimental design. For example, we show that decision-makers are prejudiced regardless of what they could have learned if *unhired* women are more productive than *hired* men.³

If our test does not indicate prejudice, then the decision-maker’s behavior can be explained *as if* their choices are free of prejudice for some private learning. This does not mean that the decision-maker’s choices are *actually* free of prejudice given what they learned. The fact that our test allows for any form of private learning means that we give as many opportunities as possible for the decision-maker’s behavior to be explained as if it is free of prejudice. Thus, our test has lower power but provides strong evidence of prejudice. Despite its power, we find that our test provides new evidence of prejudice in two well-known discrimination experiments.

We first demonstrate our test using the lab experiment of Reuben, Sapienza, and Zingales (2014), in which employers were incentivized to hire the more productive of two candidates based on the information provided in each treatment: appearance, past performance, and/or candidate predictions for future performance. Our robust test provides suggestive evidence of prejudice against women in their “Decision Then Cheap Talk” treatment (in which initial hiring decisions were made based only on the appearance of candidates) because *unhired* women were more productive than *hired* men.

This condition reveals prejudice because it implies that the employer’s threshold belief for hiring women must be above their threshold belief for hiring men. In this treatment, the probability of an *unhired* woman being more productive was 52.2% and the probability of a *hired* woman being more productive was 64.4%, so the employer’s threshold belief of productivity for hiring women is bounded between these rates. Likewise, the probability of an *unhired* man being more productive was 35.6% and the probability of a *hired* man being more productive was 47.8%, so the threshold belief for hiring men is bounded between these rates. Because the probability of an *unhired* woman being more productive (52.2%) exceeds the probability of a *hired* man being more productive (47.8%), the employer’s threshold belief for hiring women must be above their threshold belief for hiring men, indicating the employer is prejudiced against women. This conclusion holds regardless of what employers learned about male and female candidates based on their appearance.⁴

³Our test can be extended beyond binary decisions and states by applying the No Improving Action Switches (NIAS) conditions of Caplin and Martin (2015) to group-specific choice data. Rambachan (2021) extends our test further to cover important settings in which there is missing data (e.g., screening decisions) and highlights the key role of exclusion restrictions in identification.

⁴This evidence is more suggestive than conclusive because the ordering of productivity rates between hired men and unhired women is not statistically significant at traditional levels of significance, is not robust to pooling decisions across similar treatments, and can also be explained by incorrect prior beliefs. In Appendix

We also extend our test to allow for prejudice in the decision-maker’s *selection motive*. This occurs when the decision-maker positively selects for a trait in one group and negatively selects for the same trait in another group.⁵ For example, an employer calls back more productive White applicants yet – perhaps to abide by anti-discrimination laws in letter but not in spirit – calls back less productive African-American applicants.⁶ To increase its applicability, we also show that our test for prejudice in selection motive remains true as long as observed outcomes correlate sufficiently with true outcomes.⁷

We demonstrate this form of prejudice using the field experiment of Bertrand and Mullainathan (2004). In their experiment, names that strongly signal gender and race were randomly added to fictitious resumes of subjectively high and low quality. When these resumes were sent to prospective employers, Bertrand and Mullainathan (2004) observed a strong disparity in callbacks depending on the race of the name applied to a resume. Revisiting their data, we find evidence of prejudice in selection motive at the intersection of gender and race. In contrast to all other intersectional groups, the probability of a callback for an African-American male *decreases* with resume quality, from 7.4% for low-quality resumes to 4.3% for high-quality resumes. To the best of our knowledge, the negative return to resume quality at the intersection of race and gender in the experiment of Bertrand and Mullainathan (2004) has not been documented previously. Such a discrepancy in the *sign* (as opposed to the magnitude) of the effect of quality provides evidence that employers are prejudiced in their selection motive regardless of what information they gleaned from the resumes they received. Moreover, a difference in callbacks by resume quality implies that hiring managers are paying attention to resume details.

Our paper provides three main contributions. First, we introduce a test of prejudice that is both general and simple, which we demonstrate using well-known labor market experiments by Bertrand and Mullainathan (2004) and Reuben, Sapienza, and Zingales (2014). With the appropriate experimental design, researchers, policymakers, and firms can use our test to look for prejudice in other important settings, such as consumer lending and housing access decisions.

Second, by leveraging data on outcomes across decisions and groups we are able to offer an outcome test of prejudice that does not require observing marginal decisions. In the first outcome test, Becker (1957) showed that a decision-maker is prejudiced if there are differences in outcomes across groups at the margin. For instance, his test identifies prejudice against

A, we elaborate on the role of prior beliefs in our test.

⁵To the best of our knowledge, this form of prejudice has not been proposed in the literature.

⁶We follow the National Association of Black Journalists (NABJ) recommendation from June 2020 to capitalize all racial categories.

⁷This allows us to use our test to infer prejudice in correspondence studies that exogenously vary observable non-demographic characteristics that correlate with quality. For a review of correspondence studies, see Quillian, Pager, Hexel, and Midtbøen (2017), Baert (2018), and Gaddis (2018).

applicants if *at the margin* the hired applicants of one group are more productive than hired applicants of another group. However, a limitation of the Becker test is that it is often difficult to identify marginal decisions, and it has been shown that the test can produce misleading conclusions about prejudice if it is applied to average (infra-marginal) outcomes (see Ross and Yinger 1999 and Ayres 2002). In a groundbreaking paper, Knowles, Persico, and Todd (2001) show that a comparison of average observed outcomes across groups is a valid test of prejudice in their game-theoretic model because average and marginal outcomes coincide in equilibrium. In other words, there is no selection on outcomes in equilibrium. However, experimental data on treated and untreated outcomes allows us to simultaneously test (and reject) this lack of selection, as well as offer a new test that does not suffer from the inframarginality problem of the Becker outcome test.

Third, by leveraging data on outcomes across decisions and groups we are also able to offer a test of prejudice that is more robust to private learning than existing outcome tests. This robustness is especially valuable in settings where there is unobservable and selective attention. However, because robustness can decrease the power of a test, we view our test as complementary to these existing tests. For instance, our test can easily be run alongside the test of Anwar and Fang (2006), who develop an alternative outcome test that looks for differences in the rank-order of average outcomes across decision-makers of different demographic groups. In addition, Arnold, Dobbie, and C. S. Yang (2018) and Marx (2020) develop more powerful tests that jointly use information on decisions and outcomes. A common theme of these existing tests is that they assume away variation in information across decision-makers in order to attribute exogenous variation in observed behavior to differences in preferences. However, such assumptions have been questioned recently in settings such as judicial decision-making (Frandsen, Lefgren, and Leslie 2019; Gelbach 2021). This underscores the value of a robust approach.

The rest of the paper is structured as follows. Section 2 provides our model of decision-making across groups, and Section 3 formally introduces our outcome test and provides a demonstration using the experiment of Reuben, Sapienza, and Zingales (2014). Section 4 provides our extension to prejudice in selection motive and demonstrates this extension using the experiment of Bertrand and Mullainathan (2004). Section 5 discusses the implications of incorrect beliefs for our test and this extension.

2 Model of Decision-Making

We first present the simple model of decision-making across groups that motivates our test. There is a continuum of individuals, each of whom belong to an observable group $g \in \{m, w\}$. For each individual there is an imperfectly observed state $s \in \{0, 1\}$, which can be interpreted

as their type. There is also a decision-maker (DM) who makes a decision $d \in \{0, 1\}$ about each individual. For example, this can be an employer who decides whether to hire ($d = 1$) or not hire ($d = 0$) candidates of different race/ethnicity (*minority* or *white*) or of different gender (*men* or *women*),⁸ when each candidate can be of high ($s = 1$) or low ($s = 0$) future productivity. Let $P_g(d, s)$ denote the joint probability of decision d and state s for group g . With a slight abuse of notation, we also refer to the marginal distributions of decisions and states by $P_g(d)$ and $P_g(s)$, respectively.

We assume the DM makes each decision as follows. First, for each individual in a group, the DM receives a signal of the state and forms a posterior belief γ about the probability of state $s = 1$ by updating a prior belief μ_g . For now we assume that the DM's prior is correct, so that $\mu_g = P_g(s = 1)$.⁹ We summarize the signal process for each group with an information structure, defined as a discrete conditional distribution of posteriors conditional on the state, $\pi_g(\gamma|s)$, with the unconditional distribution of posteriors denoted by $\pi_g(\gamma) = \mu_g\pi_g(\gamma|s = 1) + (1 - \mu_g)\pi_g(\gamma|s = 0)$. The prior and information structure may each vary by group. However, the DM's beliefs are internally consistent with Bayes' Rule:

$$\gamma = \frac{\mu_g\pi_g(\gamma|s = 1)}{\pi_g(\gamma)} \quad (1)$$

for all groups g , states s , and posteriors γ reached with positive probability given the information structure.

Given posterior beliefs γ , the DM implements for each group g the decision d with probability $\sigma_g(d|\gamma)$. The joint probability of deciding d in state s is thus:

$$P_g(d, s) = P_g(s) \sum_{\gamma} \pi_g(\gamma|s) \sigma_g(d|\gamma) \quad (2)$$

The decision rule σ_g maximizes expected utility based on a possibly group-dependent Bernoulli utility function $u_g(d, s)$, with $u_g(0, s) \neq u_g(1, s)$ for some state s . When the DM wants to match high states with high actions,¹⁰ it is without loss of generality to parameterize the utility function as:

$$u_g(d, s) = d[s - t_g] \quad (3)$$

where $t_g \in [0, 1]$. The parameter t_g is a cost that determines the threshold posterior belief above which it is strictly optimal for the DM to take the decision $d = 1$.

The DM is defined to exhibit *prejudice* against group w if:

$$t_w > t_m. \quad (4)$$

⁸Our framework can easily be expanded to consider more than binary identities if that distinction is recorded in the data.

⁹We consider the case of incorrect prior beliefs in Section 5.

¹⁰We consider the alternate case and study prejudicial disparities in this selection motive in Section 4.

A prejudiced DM may have different preferences over decisions across groups, even when beliefs about the state are the same.

An analyst observes the group-conditional joint distributions $P_g(d, s)$ for each group g . For simplicity we restrict attention to observed distributions where $P_g(d, s) \in (0, 1)$ for all d, s . The analyst wants to determine whether the DM is prejudiced, and against whom. Next, we propose such a test.

3 Our Test

In what follows we refer to $P(s = 1|d)$ as the *outcome probability* conditional on decision d . Our test for prejudice bounds the threshold t_g by the conditional outcome probabilities and finds evidence of prejudice when the bounds across groups do not overlap.

Theorem 1. *For each group g , suppose that $P_g(s), P_g(d) \in (0, 1)$ and that the DM behaves according to our model with correct prior beliefs $\mu_g = P_g(s = 1)$. Then for each group g , the threshold t_g is sharply bounded by conditional outcome probabilities:*

$$P_g(s = 1|d = 0) \leq t_g \leq P_g(s = 1|d = 1). \quad (5)$$

Hence there is evidence of prejudice against group w if:

$$P_w(s = 1|d = 0) > P_m(s = 1|d = 1). \quad (6)$$

In the context of hiring decisions, our test reveals prejudice against women if unhired female applicants are more productive (henceforth on average) than hired male applicants. In that case, an unbiased employer could have done better by replacing hired male applicants with unhired women.

The bounds on thresholds contain the overall outcome probability $P_g(s = 1)$, and therefore $t_g = P_g(s = 1)$ is always consistent with the model. This implies that if prior probabilities are equal across groups ($P_w(s = 1) = P_m(s = 1)$), then no prejudice ($t_w = t_m$) is always consistent with the model. Specifically, all variation in choices between groups can be attributed to variations in a simple form of learning: for each group, receiving one signal for hiring (above the common threshold $t_w = t_m$) and one signal for not hiring (below the common threshold), but having these signals arrive with different probabilities for each group. Thus, a necessary condition for our test to uncover robust evidence of prejudice is that the outcome probabilities differ by group. Moreover, to uncover robust evidence of prejudice against a group, the outcome probabilities must be higher for that group. For example, a necessary condition for uncovering robust evidence of prejudice against women in the hiring context is that women must be more productive, $P_w(s = 1) > P_m(s = 1)$. Otherwise hired

male employees will be more productive than average female applicants, who are in turn more productive than unhired female applicants.

We interpret the necessary ranking between unconditional outcomes in two more constructive ways. First, the condition suggests the kinds of tasks or applications that firms, regulators, and researchers may wish to focus on if their aim is to test for prejudice in a theoretically robust way: specifically, tasks in which the discriminated group performs better (Theorem 1) or is perceived to perform better (Proposition 3). Second, the logic of Theorem 1 can be applied conditional on any realizations of covariates that are also in the decision-maker’s information set. For example, suppose the decision-maker and researcher observe a correlate of applicant quality that does not otherwise enter preferences, such as GPA; then our test would uncover prejudice if unhired African-American applicants with high GPAs were more productive than hired White applicants with low GPAs.

3.1 Empirical Application

In an already influential experiment, Reuben, Sapienza, and Zingales (2014) investigated how stereotypes about gender and mathematical ability affect the career opportunities of women relative to men, and how this varies with the provision of information to prospective employers. Experiment participants were assigned to one of four treatments that varied employers’ information about candidates’ performance on a math task.¹¹

In the “Cheap Talk” treatment, employers were provided candidates’ self-reported expected performance, and in the “Past Performance” treatment, employers were provided verifiable information about candidates’ performance on a previous task. We concentrate our attention on the other two treatments. In the “Decision Then Cheap Talk” treatment, employers made an initial employment decision with no additional information beyond appearance and then made a second employment decision after being provided information about self-reported expected performance. In the “Decision Then Past Performance” treatment, employers also made an initial employment decision with no additional information beyond appearance and then made a second employment decision after being provided information on performance on a previous task. In the subsequent task, employers were incentivized to hire the better-performing candidate, who we therefore label as being productive. Since we only analyze decisions over mixed-gender pairs, we define the female and male productivity rates as the percentage of times candidates of each gender were productive, and we say one group was more productive if it had the higher productivity rate.

¹¹Within each treatment, pairs of participants were selected as candidates for employment, and remaining participants were “employers” tasked with hiring one of the two candidates for a subsequent task. In total, the data analyzed from the experiment consists of 932 employer decisions over 76 mixed-gender candidate pairs.

Their study found large differences in hiring rates between male and female candidates when employers had no information beyond appearance (the initial employment decisions in the “Decision Then Cheap Talk” and “Decision Then Past Performance” treatments), despite the fact that men and women were on average similarly productive in the task. Additional information about candidates’ self-reported expected performance in the “Decision Then Cheap Talk” treatment did not reduce these differences in the second hiring decision because employers did not fully internalize that male candidates relatively overstated their expected performance. Additional information about candidates’ performance on a previous task in the “Decision Then Past Performance” treatments did reduce differences between men and women in the second hiring decision but did not eliminate them.

A natural question is whether the observed hiring differences between male and female candidates provide evidence of prejudice, or whether this behavior can instead be rationalized by (correct) statistical discrimination. To answer this question, we apply our robust outcome test established in Theorem 1. We begin by focusing attention on the experiment’s “Decision Then Cheap Talk” treatment because our test provides evidence of prejudice in this treatment and because decisions in this treatment provide a useful demonstration of key features of our test. Namely, women’s superior productivity in this treatment makes it possible to find evidence of prejudice against women, while the variation in informativeness illustrates the power of our test.

Figure 1 provides a visual summary of productivity rates both before and after receiving cheap talk information. Based on these rates, our test finds evidence of prejudice when employers received no information beyond appearance (left panel) because unhired women are more productive (52.2%) than hired men (47.8%).¹² This evidence is only suggestive because we fail to reject the null hypothesis that hired men are at least as productive as unhired women at conventional levels ($p = 0.28$).¹³ However, this test of significance is at the lower bound of prejudice, which is given by the difference between 52.2% and 47.8%. The upper bound of prejudice is given by the difference between the productivity of hired women (64.4%) and the productivity of unhired men (35.6%).

After receiving cheap talk information (right panel), the distribution of productivity is held fixed but employers are better at discerning productivity, and so unhired women are less productive than hired men (42.6% vs. 57.4%). As a result, while hired women are significantly more productive than hired men (84.9% vs. 57.4%),¹⁴ observed outcomes can

¹²A table that details the outcome probabilities for all treatments and decisions in the experiment is available in the Supplementary Appendix.

¹³The p -values are computed using a one-sample test of proportions that unhired women are more productive at least half of the time. We use this one-sample formulation because the outcomes of unhired women and hired men are perfectly correlated by nature of the experimental design. Alternatively, using the clustered standard errors presented and discussed in Figure 1 results in a one-tail p -value of 0.23.

¹⁴Comparing these hired outcome probabilities, the hit rate test of Knowles, Persico, and Todd (2001)

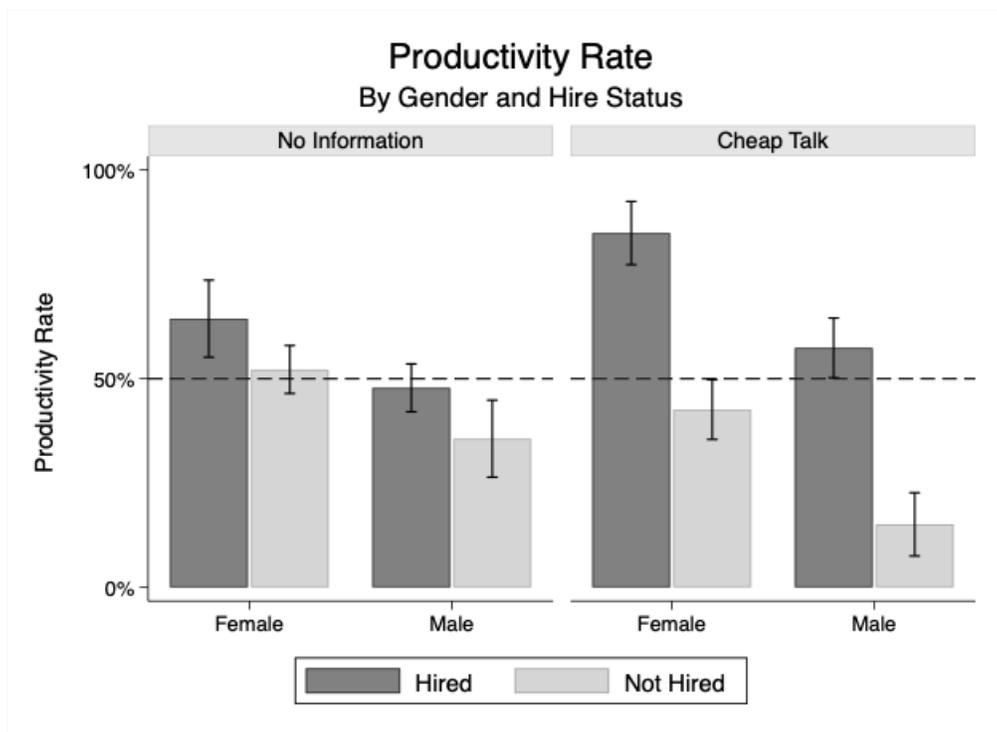


Figure 1: This figure illustrates our test for prejudice in the “Decision Then Cheap Talk” treatment of Reuben, Sapienza, and Zingales (2014). As in their analysis, standard errors are computed from a probit regression with random effects and clustering at the employer level. Since employers were incentivized to choose the higher-performing candidate in each pair, a group’s productivity rate is defined as the percentage of mixed-gender pairs in which candidates of that group were higher-performing. For initial decisions (before employers received “cheap talk” information) our test uncovers evidence of prejudice because unhired women are more productive than hired men (see left panel). For second decisions (after employers received “cheap talk” information), our test cannot rule out that differences in conditional outcome probabilities between men and women are the result of statistical discrimination instead of prejudice (see right panel).

be rationalized without prejudice – as statistical discrimination based on some unobserved learning about the state. Specifically, employer decisions are consistent with having a gender-neutral threshold and receiving a very positive signal of productivity for a small group of women and a weaker but still positive signal for a larger group of men. This does not mean that employers are no longer prejudiced after receiving cheap talk information; it does mean that their choices can be represented *as if* they are no longer prejudiced.

We conclude by relating our test and results to the experimental design. Our test is valid in spite of additional structure in the experiment, namely that for each pair of candidates exactly one candidate is more productive. In fact, this structure simplifies our test: unhired women are more productive than hired men if and only if unhired women are more productive at least half of the time. Our test also shows why, if the researcher’s goal is to assess the existence of prejudice against women, it may be more informative to use a task that is performed better by women than men. In contrast, a justification for the arithmetic task used by Reuben, Sapienza, and Zingales (2014) was that it was performed equally well by men and women. Still, our robust test provides suggestive evidence of prejudice in their experiment because women perform slightly better than men. That our test provides any evidence of prejudice is noteworthy because the unconditional productivity rates are similar across genders.

However, our empirical results are more safely interpreted as a proof of concept of the test in experimental data, rather than as conclusive evidence of taste-based discrimination. We emphasize two caveats in addition to the lack of statistical significance. First, the "No Information" decision in the “No Information Then Cheap Talk” treatment and the “Decision Then Past Performance” treatment are the same *ex ante* (up to forward-looking concerns). Yet our pointwise evidence of prejudice is not robust to pooling the treatment samples because women in the “Decision Then Past Performance” treatment are on average less productive than men.¹⁵ Second, our conclusions assume that prior beliefs are correct in the sense of agreeing with treatment group averages. We discuss the possibility and effects of incorrect prior beliefs in the experiment in Appendix A.1.

4 Selection Motive

So far we have assumed that the DM wants to match decisions to the state: $d = s$. In other words, the DM selects *for* the state. Our focus in this section is on empirically identifying

would find significant evidence of prejudice against women, but their test is invalid in this context because of selection: hired employees are more productive than unhired employees.

¹⁵In the pooled "No Information" sample, unhired women and hired men are more productive 47.76% and 52.24% of the time, respectively. Further details are provided in Table 1.

the selection motive and group-dependent disparities therein. We say that a DM exhibits *prejudice in selection motive* against group m if the decision-maker appears to select *for* the state for group w and *against* the state for group m .¹⁶ Formally, this means:

$$u_w(d, s) = d[s - t_w] \quad \text{and} \quad u_m(d, s) = -d[s - t_m]. \quad (7)$$

For example, a prejudiced employer who has to comply with anti-discrimination laws may do so in letter but not in spirit by calling back (or hiring) White applicants who are more likely to be productive but African-American applicants who are less likely to be productive, in the anticipation that less qualified applicants will not proceed to the next stage. In that case, the employer selects for productivity among White applicants but against productivity among African-American applicants. As with our previous notion of prejudice, prejudice in selection motive is a preference-based source of differences in decisions. Next we show how the selection motive (and thus, prejudice) is identified by a simple comparison of conditional outcome or decision probabilities.

Proposition 1. *For each group g , suppose that $P_g(d, s) \in (0, 1)$ and that the DM behaves according to our model with a prior belief μ_g . Then for each group g , selection for the state is identified by a strict ordering of conditional outcome probabilities:*

$$P_g(s = 1|d = 0) < P_g(s = 1|d = 1) \quad (8)$$

or decision probabilities:

$$P_g(d = 0|s = 1) < P_g(d = 1|s = 1). \quad (9)$$

Analogously, selection against the state is identified by the reverse ordering. Therefore the test finds evidence of prejudice in the selection motive against group m if the conditional outcome probabilities are inversely ranked across groups:

$$\begin{aligned} P_m(s = 1|d = 1) < P_m(s = 1|d = 0) \\ \text{and} \\ P_w(s = 1|d = 1) > P_w(s = 1|d = 0). \end{aligned} \quad (10)$$

An analogous and equivalent condition holds in terms of decision probabilities.

Identification of the selection motive is even more robust in the sense that it does not require perfect observability of the state s . Instead, let $\hat{s} \in \{0, 1\}$ denote an imperfect proxy for the DM's state that is observed by the researcher, and let $\hat{\pi}_g(\gamma|\hat{s})$ denote an information structure of posteriors conditional on the observed proxy. For example, in a

¹⁶Motivated by our subsequent application, we use m and w in this section to denote *minority* and *white* applicants, respectively.

correspondence CV study, the researcher may devise “good” and “bad” resumes \hat{s} which correlate with the true productivity or qualifications s that employers want to select for. To identify the selection motive, it is enough to assume that higher observed proxy realizations induce stochastically higher posterior beliefs over the state.

Proposition 2. *Suppose the proxy $\hat{s} = 1$ leads to stochastically higher posterior beliefs over the state (distributions of posteriors across proxies are first-order stochastically ordered):*

$$\hat{\pi}_g(\cdot|\hat{s} = 1) \succeq_{FOSD} \hat{\pi}_g(\cdot|\hat{s} = 0). \quad (11)$$

Then the selection motive is identified as in Proposition 1 upon replacing the true but unobserved state s with the observed but imperfect proxy \hat{s} .

Next we apply our generalized result for identifying prejudice in selection motive to the correspondence CV study of Bertrand and Mullainathan (2004).

4.1 Empirical Application

In an influential study, Bertrand and Mullainathan (2004) randomly assigned names that strongly signal race and gender to fictitious resumes and found significant evidence of differences in decisions in the labor market: candidates with African-American names were called back significantly less often by employers relative to candidates with White names. In addition, the study found that the returns to resume quality were lower for candidates with African-American names.¹⁷

As is well-known, data on decisions alone cannot identify whether a difference in decisions across groups is the result of preference-based prejudice, information-based statistical discrimination, or both. However, Proposition 2 provides a test of preference-based prejudice in selection motive if we assume that resume quality is an imperfect proxy for the true state important to employers (e.g., productivity). If there is no prejudice in selection motive, then the ordering of callback rates across resume quality (or resume quality across callback rates) should be independent of race. Allowing for the possibility of intersectional prejudice, the same ordering should be independent of race interacted with gender. For consistency with the original study, we apply our test in terms of callback rates.

¹⁷More specifically, the study randomly assigned 4,870 resumes to names that were selected for being strongly suggestive of race and gender. To measure differences in the returns to qualifications across race, the resumes were subjectively classified and further manipulated to be of either “high” or “low” quality. High quality resumes had on average more experience, fewer employment gaps, an email address, foreign language skills, and additional certifications or honors. Each employment ad received four experimentally-generated resumes: a high and low quality resume with a typically African-American or White name. Employment ads were answered in Boston and Chicago and were further classified into “administrative” and “sales” roles. Traditionally female names were sent to ads for administrative jobs, whereas both male and female names were sent to ads for sales jobs.

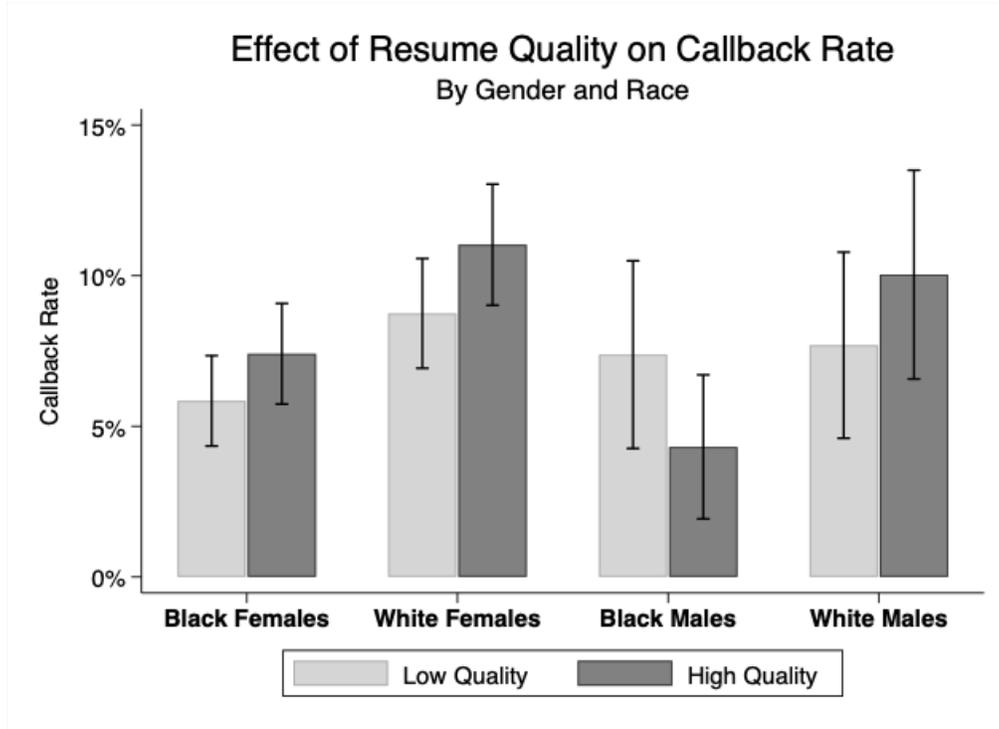


Figure 2: This figure plots the mean callback rates by subjective resume quality across race-gender pairs in the study of Bertrand and Mullainathan (2004). Similarly to their Table 5, standard errors are corrected for clustering at the employment-ad level in a probit regression of the callback dummy on a full interaction of race, gender, and resume quality. The main finding is that callback rates increase in resume quality for all groups, except for African-American men. In our framework, a racial difference in the sign of the effect of resume quality on callback rates constitutes evidence of prejudice in employers' selection motive.

Figure 2 plots the callback rates across resume quality for each intersectional group. Our main finding is that resume quality *decreases* the callback rate (only) for African-American men. The mean callback rate for low-quality resumes with the names of African-American men is 7.4%, yet the mean callback rate for high-quality resumes with the names of African-American men is only 4.3%. The null hypothesis that the callback rate for African-American men is weakly increasing in quality is rejected at the 90% level of confidence ($p = 0.063$).¹⁸ To the best of our knowledge this finding is new. The original study of Bertrand and Mullainathan (2004) finds significant evidence of lower but *positive* returns to quality across race. Such differences in magnitude may be a product of statistical discrimination. In contrast, our results disaggregated by race and gender indicate *negative* returns to quality among African-American males. Such differences in sign are not easily explained by statistical discrimination. In our simple framework, a difference in sign constitutes evidence of preference-based prejudice in the selection motive.¹⁹

Furthermore, the finding of negative returns to quality for African-American men has an implication for selective attention. In order for there to be a differential in callback rates across quality at all, it must be that employers are paying at least some attention to resume details beyond the names provided. However, the employers then appear to use the acquired information to select against quality among African-American men.

We conclude by discussing interpretations of our result. In a model with binary actions and states, it is without loss of generality to interpret the observed decision behavior as being generated by a preference to select against quality among African-American men. For example, an employer calls back more productive White applicants yet – perhaps to abide by anti-discrimination laws in letter but not in spirit – calls back less productive African-American applicants. Alternative explanations of our finding are also possible, particularly if we enrich the state space underlying decision-maker preferences. For example, employers could favor African-American men with lower-quality resumes, in the sense that the employers seek to help candidates they perceive to have been under-privileged.²⁰ However, in that case we might expect to see higher callback rates among low-quality African-American men than their low-quality peers, which we do not. Alternatively, employers may be less likely to call back high-quality African-American men if callbacks are costly for employers, and employers view African-American men with high-quality resumes as unlikely to accept their offer because of strong competing offers.²¹ Also, the nature of the jobs that call back low- and high-quality applicants may differ in a way that correlates with race. For example,

¹⁸The p -values are computed from two-sample one-sided tests of proportion. Alternatively, the one-sided test using the clustered standard errors in Figure 2 is smaller and significant at the 95% level of confidence ($p = 0.047$).

¹⁹In the Supplementary Appendix we provide a robustness analysis for this result.

²⁰We are grateful to Roland Bénabou for this suggestion.

²¹We are grateful to Laura Gee and Quoc-Anh Do for this suggestion.

low-quality African-American men may be receiving callbacks primarily from “lower-quality” sales positions.²² However, we observe prejudice in the selection motive across all three job types to which the resumes of African-American males were sent (managerial positions, sales representative positions, and retail sales positions).

5 Incorrect Beliefs

Many standard tests for prejudice, including the classic test of Becker (1957), assume that the DM’s prior and posterior beliefs are, on average, correct. This may not be the case, especially in settings where there is already a concern about potential discrimination.²³ However, our test for prejudice is robust to many forms of belief updating biases. For example, consider one of the most well-documented forms of updating bias, where the DM is conservative in their updating. In that case, if unhired women are more productive than hired men, the highest (incorrect) posterior belief at which women are not hired will still be above the lowest (incorrect) posterior belief at which hired men are hired.²⁴ This will also be the case for other forms of conservatism, such as confirmation bias (not fully updating after signals that go against the prior) and asymmetric belief updating based on prejudicial preferences (not fully updating after signals that go against preferences).

In addition, our test of prejudice in the selection motive (Proposition 1) does not require the DM to hold correct prior beliefs. On the other hand, our baseline test of prejudice (Theorem 1) does require that prior beliefs coincide with the observed distribution of the state for both groups. This is a general limitation of outcome-based tests for prejudice. As shown recently by Bohren, Haggag, Imas, and Pope (2019), it is not possible to identify prejudice without assumptions or information on prior beliefs, even with strong parametric assumptions and perfect observability of the decision-maker’s learning process.²⁵ Nevertheless, a decision-maker’s prior can be incorrect for a number of reasons. For instance, their beliefs can be distorted by stereotyping (e.g., Bordalo, Coffman, Gennaioli, and Shleifer 2016) or because the experimental distribution of the outcomes deviates from the population in ways decision-makers are unaware of. The latter can happen, for instance, if the DM does not appreciate selection into the experiment. It can also happen in field studies when random

²²We are grateful to Sigrid Suetens for this suggestion.

²³See Bohren, Haggag, Imas, and Pope (2019) for a review of the literature on incorrect statistical discrimination, and Bursztyn and D. Y. Yang (2021) for a recent meta-analysis of field experiments documenting a broad pattern of group-based misperceptions in beliefs.

²⁴See Benjamin (2019) for a review of belief biases, and Albrecht, Von Essen, Parys, and Szech (2013) for experimental evidence of discrimination being driven by conservatism in belief updating.

²⁵The indistinguishability between prior beliefs and taste thresholds has been discussed previously in the context of health care (Chandra and Staiger 2010; Abaluck, Agha, Kabrhel, Raja, and Venkatesh 2016) and is also formalized in Arnold, Dobbie, and C. S. Yang (2018).

assignment of quality by demographic group is independent of the distribution in the field (as is often the case in correspondence studies).

However, there is a simple solution for our test in experimental settings: eliciting prior beliefs. In the Supplementary Appendix, we provide and empirically apply a result (Proposition 3) that expresses joint bounds on a decision-maker's preferences and prior beliefs. These bounds can either be combined with elicited prior beliefs to fully recover our test for prejudice or used to determine the set of prior beliefs that would (not) imply that a decision-maker is prejudiced. We pursue the latter exercise in Appendix A.1.

References

- Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh (2016). “The determinants of productivity in medical testing: Intensity and allocation of care”. In: *American Economic Review* 106.12, pp. 3730–64.
- Albrecht, Konstanze, Emma Von Essen, Juliane Parys, and Nora Szech (2013). “Updating, self-confidence, and discrimination”. In: *European Economic Review* 60, pp. 144–169.
- Anderson, Lisa, Roland Fryer, and Charles Holt (2006). “Discrimination: experimental evidence from psychology and economics”. In: *Handbook on the Economics of Discrimination*, pp. 97–118.
- Anwar, Shamena and Hanming Fang (2006). “An alternative test of racial prejudice in motor vehicle searches: Theory and evidence”. In: *American Economic Review* 96.1, pp. 127–151.
- Arnold, David, Will Dobbie, and Crystal S Yang (2018). “Racial bias in bail decisions”. In: *The Quarterly Journal of Economics* 133.4, pp. 1885–1932.
- Arrow, Kenneth (1971). “The theory of discrimination”. In:
- Ayres, Ian (2002). “Outcome tests of racial disparities in police practices”. In: *Justice research and Policy* 4.1-2, pp. 131–142.
- Baert, Stijn (2018). “Hiring discrimination: An overview of (almost) all correspondence experiments since 2005”. In: *Audit studies: Behind the scenes with theory, method, and nuance*, pp. 63–77.
- Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka (2016). “Attention discrimination: Theory and field experiments with monitoring information acquisition”. In: *American Economic Review* 106.6, pp. 1437–75.
- Becker, Gary S (1957). *The economics of discrimination*. University of Chicago press.
- Benjamin, Daniel J (2019). “Errors in probabilistic reasoning and judgment biases”. In: *Handbook of Behavioral Economics: Applications and Foundations 1 2*, pp. 69–186.
- Bertrand, Marianne and Esther Duflo (2017). “Field experiments on discrimination”. In: *Handbook of economic field experiments 1*, pp. 309–393.
- Bertrand, Marianne and Sendhil Mullainathan (2004). “Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination”. In: *American economic review* 94.4, pp. 991–1013.
- Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope (2019). *Inaccurate statistical discrimination*. Tech. rep. National Bureau of Economic Research.
- Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer (2016). “Stereotypes”. In: *The Quarterly Journal of Economics* 131.4, pp. 1753–1794.
- Bursztyjn, Leonardo and David Y Yang (2021). *Misperceptions about Others*. Tech. rep. National Bureau of Economic Research.

- Caplin, Andrew and Daniel Martin (2015). “A testable theory of imperfect perception”. In: *The Economic Journal* 125.582, pp. 184–202.
- Chandra, Amitabh and Douglas O Staiger (2010). *Identifying provider prejudice in health-care*. Tech. rep. National Bureau of Economic Research.
- Frandsen, Brigham R, Lars J Lefgren, and Emily C Leslie (2019). *Judging judge fixed effects*. Tech. rep. National Bureau of Economic Research.
- Gaddis, S Michael (2018). *Audit studies: Behind the scenes with theory, method, and nuance*. Vol. 14. Springer.
- Gelbach, Jonah B (2021). “Testing Economic Models of Discrimination in Criminal Justice”. In: *Available at SSRN 3784953*.
- Knowles, John, Nicola Persico, and Petra Todd (2001). “Racial bias in motor vehicle searches: Theory and evidence”. In: *Journal of Political Economy* 109.1, pp. 203–229.
- Lane, Tom (2016). “Discrimination in the laboratory: A meta-analysis of economics experiments”. In: *European Economic Review* 90, pp. 375–402.
- Marx, Philip (2020). “An Absolute Test of Racial Prejudice”. In:
- Neumark, David (2018). “Experimental research on labor market discrimination”. In: *Journal of Economic Literature* 56.3, pp. 799–866.
- Phelps, Edmund S (1972). “The statistical theory of racism and sexism”. In: *The american economic review* 62.4, pp. 659–661.
- Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen (2017). “Meta-analysis of field experiments shows no change in racial discrimination in hiring over time”. In: *Proceedings of the National Academy of Sciences* 114.41, pp. 10870–10875.
- Rambachan, Ashesh (2021). *Identifying prediction mistakes in observational data*. Tech. rep. Working Paper, Harvard University Department of Economics.
- Reuben, Ernesto, Paola Sapienza, and Luigi Zingales (2014). “How stereotypes impair women’s careers in science”. In: *Proceedings of the National Academy of Sciences* 111.12, pp. 4403–4408.
- Riach, Peter A and Judith Rich (2002). “Field experiments of discrimination in the market place”. In: *The economic journal* 112.483, F480–F518.
- Ross, Stephen L and John Yinger (1999). “The default approach to studying mortgage discrimination: A rebuttal”. In: *Mortgage Lending Discrimination: A Review of Existing Evidence, edited by Margery A. Turner and Felicity Skidmore*. Urban Institute, Washington DC.
- Wozniak, David and Timothy MacNeill (2020). “Racial discrimination in the lab: Evidence of statistical and taste-based discrimination”. In: *Journal of Behavioral and Experimental Economics* 85, p. 101512.

Supplementary Appendix for “A Robust Test of Prejudice for Discrimination Experiments”

A Incorrect Prior Beliefs

In what follows, we relax the assumption of a correct prior belief to an assumption that the prior belief is interior: $\mu_g \in (0, 1)$. The following result jointly identifies the beliefs and tastes consistent with the data. We refer to $P_g(d = 1|s)$ as the decision probability for group g conditional on state s .

Proposition 3 (Identification with Possibly Incorrect Prior Beliefs). *For each group g , suppose that $P_g(d, s) \in (0, 1)$ and that the DM behaves according to our model with a (possibly incorrect) prior belief μ_g . Then for each group g , the prior and threshold are jointly bounded by the likelihood ratios of conditional decision probabilities across states:*

$$\frac{P_g(d = 0|s = 1)}{P_g(d = 0|s = 0)} \leq \frac{1 - \mu_g}{\mu_g} \frac{t_g}{1 - t_g} \leq \frac{P_g(d = 1|s = 1)}{P_g(d = 1|s = 0)} \quad (12)$$

One takeaway from Proposition 3 is that $t_g = \mu_g$ is always an empirically consistent threshold, so without further data or assumptions, it is impossible to discern whether disparities in the DM’s decisions across groups are the result of taste-based prejudice or incorrect prior beliefs.

Nevertheless, Proposition 3 has useful applications. First, the result restricts the set of thresholds consistent with a set of prior beliefs. Using this, the researcher may experimentally elicit prior beliefs, as in Bohren, Haggag, Imas, and Pope (2019), and then restrict the set of consistent thresholds. Second, the result bounds the set of prior beliefs consistent with a threshold. Thus the researcher can determine the set of prior beliefs such that decisions can be rationalized without prejudice. Next we consider these applications empirically.

A.1 Empirical Application

The stereotype that men outperform women in math and science-related tasks is a primary motivation for the experiment of Reuben, Sapienza, and Zingales (2014), and they find evidence of such stereotyping by eliciting expectations of performance and implicit association biases.²⁶ This raises a natural question for our test: what (perhaps incorrect) prior beliefs

²⁶Their tables S8 and S9 document systemic overestimation in employers’ reported expectations that the male candidate will outperform the female candidate. While they may suggest incorrect beliefs, these reports of whether the posterior belief exceeds 50% are unfortunately not sufficient to recover prior or posterior beliefs. We are grateful to an anonymous referee for pushing us to elaborate on this point, as it also underscores the importance of eliciting *probabilistic* beliefs in experiments that wish to apply our tests.

would still imply the existence of taste-based prejudice?

Applying Proposition 3 to initial hiring decisions in the “Decision Then Cheap Talk” treatment, we find that our previous evidence of prejudice against women from initial hiring decisions (before employers received cheap talk information) is robust to any prior belief that a female candidate outperforms a male candidate above 54.0% (s.e. = 1.89%). We can similarly deduce that for any prior belief about women’s ability between 41.5% and 54.0%, the data can be rationalized without taste-based prejudice. Likewise, for any prior belief below 41.5%, we would conclude that there is simultaneously both evidence of bias *against women* in prior beliefs and evidence of prejudice *against men* in tastes.

We also apply Proposition 3 to the “Decision Then Past Performance” treatment.²⁷ In this treatment, our test with correct prior beliefs (based on Theorem 1) does not provide evidence of prejudice in initial hiring decision because women in the treatment are on average less productive than men (46.6% vs. 53.4%). However, based on Proposition 3 we can identify the threshold prior belief above which the observed initial hiring decisions would imply prejudice against women. The threshold belief (54.2%, s.e. = 2.1%) in the “Decision Then Past Performance” treatment is notably similar to the threshold belief in the “Decision Then Cheap Talk” treatment, even though the experimental distributions of outcomes differ. The threshold prior beliefs for these treatments are slightly higher than the 53.1% probability that women are more productive than men across all treatments. If employer beliefs agreed with this overall percentage, then we would come close to finding evidence of prejudice in both treatments.

B Robustness of Selection Motive Results

We now discuss the empirical robustness and interpretation of our finding of negative returns to resume quality for African-American men in the data of Bertrand and Mullainathan (2004). The finding is primarily driven by a difference in the Chicago labor market. In Chicago, the callback rate of 12.9% for low-quality resumes with male African-American names is the highest among all intersectional groups and resume qualities, while the callback rate of 3.5% for high-quality resumes with male African-American names is the lowest among all groups and qualities. The null hypothesis that the callback rate in Chicago for resumes with African-American names is weakly increasing in quality is rejected at the 95% level of confidence ($p = 0.013$). Even in the other market, Boston, the effect of quality on callback

²⁷Recall that employers in this treatment received information about performance on a previous task – instead of cheap talk – before making their second hiring decision. This distinction is known to participants before they make their initial decisions, which motivated our decision to consider these treatments separately. The results of this section are qualitatively the same if we pool initial decisions (before receiving additional information) across the two treatments.

rates for resumes with African-American names is weakly negative, albeit statistically indistinguishable from zero.²⁸ In each city, the estimated returns to quality are negative only for African-American men.

We additionally consider whether the effect is a consequence of job type, since over 93% of male resumes (but only half of female resumes) are sent to sales jobs. Indeed, for both White and African-American female resumes sent to sales jobs, the effect of resume quality is indistinguishable from zero.²⁹ In other words, economically meaningful returns to resume quality for women are limited to administrative job roles. The returns to resume quality for males in sales roles remain qualitatively similar to before, given the similarity of samples. For African-American men, we again find significant evidence of negative returns to quality (one-sided $p = 0.048$). While we fail to reject that the returns for White men in sales roles are also weakly negative, the evidence is suggestive of positive returns. Namely, for White men in sales roles, the callback rate for low-quality resumes is 7.9%, the callback rate for high-quality resumes is 10.4%, and the one-tail p -value is 0.16. Interpreting this as evidence of positive returns yields an intriguing conclusion: for sales roles, only White men benefit from higher resume quality, women are unaffected, and African-American men are harmed.

We also find that this result is robust to job type. 36.1% of resumes for African-American men were sent to retail sales jobs, 32.6% were sent to sales representative positions, and 25.7% were sent to sales manager positions. For all three job types, we find evidence of negative returns to quality. For retail sales jobs, high quality resumes had a 7.0% callback rate, while low quality resumes had a 12.2% callback rate. For sales representative positions, the respective rates are 1.1% and 3.3%, and for sale manager positions, they are 2.7% and 6.0%.

C Proofs

We begin by stating and proving two simple lemmas about beliefs and decision-making that are invoked in the text and subsequent results.

Lemma 1. *Suppose posterior beliefs are consistent with Bayes' Rule (1). Then higher posteriors are more likely in higher states, in the sense of first order stochastic dominance:*

$$\pi_g(\gamma|s = 1) \succeq_{FOSD} \pi_g(\gamma|s = 0)$$

²⁸For African-American men in Boston, the callback rate for low-quality resumes is 4.8%, the callback rate for high-quality resumes is 4.7%, and the one-tail p -value is 0.47.

²⁹For African-American women in sales roles, the callback rate for low-quality resumes is 7.0%, the callback rate for high-quality resumes is 6.7%, and the one-tail p -value is 0.44. For White women in sales roles, the respective numbers are 8.2%, 8.5%, and $p = 0.45$. In contrast, for women in administrative roles the increase in each callback rate across resume quality is between 2 and 3 percentage points and is statistically significant at the 95% level of confidence.

Proof of Lemma 1. By definition of first-order stochastic dominance, it suffices to show that:

$$\sum_{\gamma \leq p} \pi_g(\gamma|s=0) \geq \sum_{\gamma \leq p} \pi_g(\gamma|s=1) \quad \text{for all } p \in [0, 1]. \quad (13)$$

The result is trivial for $\mu_g \in \{0, 1\}$, in which case the only reached posterior is the prior. Henceforth we assume $\mu_g \in (0, 1)$. Rearranging (1) yields:

$$\gamma[1 - \mu_g]\pi_g(\gamma|s=0) = [1 - \gamma]\mu_g\pi_g(\gamma|s=1) \quad (14)$$

for every reached posterior γ . The term $\gamma[1 - \mu_g]$ is increasing in γ and the term $[1 - \gamma]\mu_g$ is decreasing in γ , and the two terms are equal at $\gamma = \mu_g$. Therefore we have $\pi_g(\gamma|s=0) \gtrless \pi_g(\gamma|s=1)$ for $\mu_g \gtrless \gamma$. Summing over reached posteriors $\gamma \leq \mu_g$ yields the desired inequality (13) for $p \leq \mu_g$:

$$\sum_{\gamma \leq p} \pi_g(\gamma|s=0) \geq \sum_{\gamma \leq p} \pi_g(\gamma|s=1) \quad \text{for } p \in [0, \mu_g]$$

For $p \geq \mu_g$ we instead have:

$$\sum_{\gamma > p} \pi_g(\gamma|s=0) \leq \sum_{\gamma > p} \pi_g(\gamma|s=1) \quad \text{for } p \in [\mu_g, 1]$$

Substituting:

$$\sum_{\gamma > p} \pi_g(\gamma|s) = 1 - \sum_{\gamma \leq p} \pi_g(\gamma|s)$$

and simplifying yields the desired inequality (13) for the remaining case $p \geq \mu_g$. □

Lemma 2. *Suppose decisions maximize expected utility according to Bernoulli utility function (3). Then the posterior-conditional decision probability $\sigma_g(d=1|\gamma)$ is nondecreasing in the posterior belief γ .*

Proof of Lemma 2. Expected utility maximization and the functional form (3) imply that for every posterior γ :

$$\sigma(d=1|\gamma)[\gamma - t_g] \geq 0 \quad \text{and} \quad \sigma(d=0|\gamma)[\gamma - t_g] \leq 0$$

The optimality condition for $d=1$ requires that $\sigma(d=1|\gamma) = 0$ if $\gamma < t_g$. Analogously, the optimality condition for $d=0$ requires that $\sigma(d=0|\gamma) = 0$ if $\gamma > t_g$. Because $\sigma(d=1|\gamma) = 1 - \sigma(d=0|\gamma)$, combining implies:

$$\sigma(d=1|\gamma) = \begin{cases} 0 & \text{if } \gamma < t_g \\ 1 & \text{if } \gamma > t_g \end{cases}$$

Since $\sigma(d=1|t_g) \in [0, 1]$, this yields the desired result. □

Proof of Theorem 1. Expected utility maximization and the functional form (3) imply that for every posterior γ :

$$\sigma(d = 1|\gamma)[\gamma - t_g] \geq 0.$$

For every posterior γ reached with positive probability, plugging in the Bayesian consistency condition (1) and rearranging yields:

$$\mu_g \pi_g(\gamma|s = 1) \sigma(d = 1|\gamma) \geq t_g \pi_g(\gamma) \sigma(d = 1|\gamma)$$

Plugging in correct beliefs $\mu_g = P_g(s = 1)$, summing over reached posteriors, and replacing from (2) yields:

$$P_g(d = 1, s = 1) \geq t_g P_g(d = 1).$$

Dividing by $P_g(d = 1) > 0$ and expressing in terms of the conditional probability yields:

$$P_g(s = 1|d = 1) \geq t_g. \tag{15}$$

An analogous argument beginning from the expected utility maximization of $d = 0$ yields:

$$P_g(s = 1|d = 0) \leq t_g. \tag{16}$$

Combining implies:

$$P_g(s = 1|d = 0) \leq t_g \leq P_g(s = 1|d = 1).$$

Comparing across groups, (6) implies that:

$$t_w \geq P_w(s = 1|d = 0) > P_m(s = 1|d = 1) \geq t_m.$$

which rejects the null hypothesis that $t_w \leq t_m$. □

Proof of Proposition 3. As in the proof of Theorem 1, expected utility maximization implies that:

$$\sigma(d = 1|\gamma)[\gamma - t_g] \geq 0.$$

For every posterior γ reached with positive probability, plugging in the Bayesian consistency condition (1) and rearranging yields:

$$\mu_g [1 - t_g] \pi_g(\gamma|s = 1) \sigma(d = 1|\gamma) \geq [1 - \mu_g] t_g \pi_g(\gamma|s = 0) \sigma(d = 1|\gamma)$$

Summing over reached posteriors and invoking (2) in conditional form yields:

$$\mu_g [1 - t_g] P_g(d = 1|s = 1) \geq (1 - \mu_g) t_g P_g(d = 1|s = 0). \tag{17}$$

By assumption that $P_g(d, s) \in (0, 1)$, we have $P_g(d|s) \in (0, 1)$ for all d, s . Also by assumption $\mu_g \in (0, 1)$. Finally our assumptions imply that $t_g < 1$, else the preceding inequality (17)

combined with $\mu_g \in (0, 1)$ would imply that $P_g(d = 1|s = 0) = 0$, a contradiction to the preceding statement that $P_g(d|s) \in (0, 1)$. Thus rearranging (17) yields:

$$\frac{1 - \mu_g}{\mu_g} \frac{t_g}{1 - t_g} \leq \frac{P_g(d = 1|s = 1)}{P_g(d = 1|s = 0)} \quad (18)$$

An analogous argument beginning from the expected utility maximization of $d = 0$ yields:

$$\frac{1 - \mu_g}{\mu_g} \frac{t_g}{1 - t_g} \geq \frac{P_g(d = 0|s = 1)}{P_g(d = 0|s = 0)} \quad (19)$$

Combining implies:

$$\frac{P_g(d = 0|s = 1)}{P_g(d = 0|s = 0)} \leq \frac{1 - \mu_g}{\mu_g} \frac{t_g}{1 - t_g} \leq \frac{P_g(d = 1|s = 1)}{P_g(d = 1|s = 0)}$$

which proves the desired bounds. \square

Proof of Proposition 1. To make use of our existing results, we prove the contrapositive of the orderings in the case of selection against the state (the arguments for positive selection follow analogously). Namely, we begin by showing that if the DM selects for the state, then:

$$P_g(d = 1|s = 0) \leq P_g(d = 1|s = 1). \quad (20)$$

From the bounds in Proposition 3, we have:

$$\frac{P_g(d = 0|s = 1)}{P_g(d = 0|s = 0)} \leq \frac{P_g(d = 1|s = 1)}{P_g(d = 1|s = 0)} \quad (21)$$

Substituting $P_g(d = 0|s) = 1 - P_g(d = 1|s)$ and simplifying yields the desired ordering (20) for conditional decision probabilities. To obtain the desired ordering for conditional outcome probabilities, we substitute according to Bayes' Rule $P_g(d|s) = P_g(s|d)P_g(d)/P_g(s)$ in (21). Analogously to the argument for decision probabilities, substituting $P_g(s = 0|d) = 1 - P_g(s = 1|d)$ and simplifying yields the desired order for outcome probabilities. Finally, the test for prejudice in selection motive is immediate from comparing the identified selection motive across groups. \square

Proof of Proposition 2. As in the proof of Proposition 1, we prove the contrapositive of the orderings in the case of selection against the state. Namely, it suffices to show that if the DM selects for the imperfectly observed state s , then:

$$P_g(d = 1|\hat{s} = 0) \leq P_g(d = 1|\hat{s} = 1). \quad (22)$$

The remaining results then follow as in Proposition 1. To show (22), first note that $\sigma(d = 1|\gamma)$ is a nondecreasing function of γ by Lemma 2. Since $\sigma(d = 1|\gamma)$ is a nondecreasing function of γ , the assumed FOSD ranking implies that:

$$\sum_{\gamma} \hat{\pi}_g(\gamma|\hat{s} = 0)\sigma(d = 1|\gamma) \leq \sum_{\gamma} \hat{\pi}_g(\gamma|\hat{s} = 1)\sigma(d = 1|\gamma).$$

But this is exactly:

$$P_g(d = 1|\hat{s} = 0) \leq P_g(d = 1|\hat{s} = 1)$$

which is the desired result.

□

Fix (and suppress the notation for) a group g and suppose that the DM's decision rule $\sigma(d|\gamma)$ is consistent with expected utility maximization for some Bernoulli utility function with $u(d, s) \neq u(1-d, s)$ for some state s . EU maximization implies that for any posterior γ and decision d such that $\sigma(d|\gamma) > 0$,

$$\gamma u(d, 1) + (1 - \gamma)u(d, 0) \geq \gamma u(1 - d, 1) + (1 - \gamma)u(1 - d, 0)$$

where $\sigma(d|\gamma) = 1$ if the condition holds strictly. Rearranging terms, it must be that:

$$\gamma \left[\left(u(d, 1) - u(1 - d, 1) \right) + \left(u(1 - d, 0) - u(d, 0) \right) \right] \geq \left[u(1 - d, 0) - u(d, 0) \right]$$

If the bracketed term on the LHS is equal to zero, then by assumption that the utility function is non-trivial and by EU maximization it must be that the term on the RHS is strictly negative. Thus the EU maximization condition holds strictly for all posteriors, so that $\sigma(d|\gamma) = 1$ for all $\gamma \in [0, 1]$. A parametrized utility function with threshold $t = 0$ and $a = 1 - d$ rationalizes such behavior. If the bracketed term on the LHS is nonzero, dividing yields a posterior cutoff representation. Letting:

$$a = \mathbb{I} \left\{ \left[u(d, 1) - u(1 - d, 1) \right] + \left[u(1 - d, 0) - u(d, 0) \right] > 0 \right\}$$

and setting t equal to the cutoff (or if the cutoff falls outside the interval $[0, 1]$, the nearest endpoint) yields the desired representation.

D Supplementary Tables

Table 1: Conditional Productivity Rates by Gender and Hire Status

Gender	First Decision				Second Decision			
	Female		Male		Female		Male	
	No	Yes	No	Yes	No	Yes	No	Yes
Hired								
Treatment: "No Information Then Cheap Talk"								
Productivity	52.20	64.37	35.63	47.80	42.62	84.88	15.12	57.38
Rate (%)	(2.93)	(4.71)	(4.71)	(2.93)	(3.66)	(3.87)	(3.87)	(3.66)
N	182	87	87	182	183	86	86	183
Treatment: "No Information Then Past Performance"								
Productivity	42.48	54.12	45.88	57.52	15.86	94.62	5.38	84.14
Rate (%)	(2.32)	(4.96)	(4.96)	(2.32)	(3.17)	(2.24)	(2.24)	(3.17)
N	153	85	85	153	145	93	93	145
Treatments: Pooled "No Information Then {Cheap Talk, Past Performance}"								
Productivity	47.76	59.30	40.70	52.24				
Rate (%)	(1.94)	(3.47)	(3.47)	(1.94)				
N	335	172	172	335				
Treatment: "Cheap Talk"								
Productivity					43.40	92.59	7.41	56.60
Rate (%)					(5.36)	(3.61)	(3.61)	(5.36)
N					106	54	54	106
Treatment: "Past Performance"								
Productivity					23.26	84.68	15.32	76.74
Rate (%)					(3.64)	(4.23)	(4.23)	(3.64)
N					129	111	111	129

Notes: This table provides productivity rates (outcome probabilities) and standard errors by applicant gender and hire status across treatment subsamples in the dataset of Reuben, Sapienza, and Zingales (2014). As in their analysis, standard errors are computed from a probit regression with random effects and clustering at the employer level. Since employers were incentivized to choose the higher-performing candidate in each pair, a group's productivity rate is defined as the percentage of mixed-gender pairs in which candidates of that group were higher-performing.

Table 2: Callback Rates by Race, Gender, and Quality

Race Gender Quality	Applicant Characteristics							
	Black		White		Black		White	
	Females		Females		Males		Males	
	Low	High	Low	High	Low	High	Low	High
Panel A: All Observations								
Callback Rate (%)	5.84 (0.77)	7.41 (0.85)	8.75 (0.93)	11.03 (1.03)	7.38 (1.59)	4.32 (1.22)	7.69 (1.58)	10.03 (1.77)
N	941	945	926	934	271	278	286	289
Panel B: Chicago								
Callback Rate (%)	4.44 (0.85)	5.53 (0.94)	6.64 (1.03)	8.45 (1.14)	12.94 (3.64)	3.53 (2.00)	10.84 (3.42)	12.22 (3.45)
N	585	597	587	592	85	85	83	90
Panel C: Boston								
Callback Rate (%)	8.15 (1.45)	10.63 (1.65)	12.39 (1.79)	15.50 (1.96)	4.84 (1.57)	4.66 (1.52)	6.40 (1.72)	9.05 (2.04)
N	356	348	339	342	186	193	203	199
Panel D: Sales								
Callback Rate (%)	6.99 (1.55)	6.67 (1.56)	8.20 (1.72)	8.54 (1.78)	7.17 (1.63)	3.83 (1.19)	7.87 (1.65)	10.37 (1.86)
N	272	255	256	246	251	261	267	270
Panel E: Admin								
Callback Rate (%)	5.38 (0.87)	7.68 (1.01)	8.96 (11.0)	11.92 (1.24)	10.00 (6.71)	11.76 (7.82)	5.26 (5.13)	5.26 (5.13)
N	669	690	670	688	20	17	19	19

Notes: This table provides callback rates (decision probabilities) and standard errors by applicant race, gender, and resume quality across city and job occupation subsamples in the dataset of Bertrand and Mullainathan (2004). Similarly to their Table 5, standard errors are corrected for clustering at the employment-ad level in a probit regression of the callback dummy on a full interaction of race, gender, and resume quality.

Supplementary Appendix References

- Bertrand, Marianne and Sendhil Mullainathan (2004). “Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination”. In: *American economic review* 94.4, pp. 991–1013.
- Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope (2019). *Inaccurate statistical discrimination*. Tech. rep. National Bureau of Economic Research.
- Reuben, Ernesto, Paola Sapienza, and Luigi Zingales (2014). “How stereotypes impair women’s careers in science”. In: *Proceedings of the National Academy of Sciences* 111.12, pp. 4403–4408.