A Robust Test of Prejudice for Discrimination Experiments

Daniel Martin* and Philip Marx†

May 8, 2021

Abstract

Lab and field experiments have proven to be an important source of empirical evidence on discrimination. We show that if average outcomes in a discrimination experiment satisfy simple conditions, then this provides evidence that decision-makers are prejudiced – regardless of what they learned about individuals in each demographic group before making their decisions. We demonstrate our robust test of prejudice using the lab experiment of Reuben, Sapienza, and Zingales (2014) and the field experiment of Bertrand and Mullainathan (2004).

Key words: Discrimination, prejudice, experiments
JEL codes: D60, D83, D91

*Kellogg School of Management, Northwestern University.
†Department of Economics, Louisiana State University.

1 Introduction

Discrimination is a pressing social issue, 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, it is important to know what drives such discrimination. The economic literature distinguishes primarily between preference- and belief-based channels of discrimination.[2] *Prejudice* (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 challenge is especially pronounced in settings where discrimination is impacted by selective attention that depends on group identity (Bartoš, Bauer, Chytilová, and Matějka 2016). 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. Thus they may never see the skills of African-American candidates and this could explain why these skills are not rewarded."

We overcome this identification challenge by deriving conditions on outcomes that reveal prejudice regardless what decision-makers have learned about individuals in each demographic group. In other words, we provide a test for prejudice that is *robust* to any form of private learning. The key to our test is to compare outcomes across decisions and groups,

[1] Anderson, Fryer, and Holt (2006) and Lane (2016) review lab experiments on discrimination, and Riach and Rich (2002), Bertrand and Duflo (2017), and Neumark (2018) review field experiments on discrimination.

[2] See Heckman (1998), Persico (2009), Charles and Guryan (2011), Bohren, Haggag, Imas, and Pope (2019), and Lang and Kahn-Lang Spitzer (2020) for reviews of this literature.

which is often possible in experiments.[3] For example, our test identifies prejudice if *unhired* women are more productive than *hired* men.

Because of its robustness, our test can only be used to provide evidence of prejudice, not evidence that a decision-maker's choices are free of prejudice. 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 can be seen as having low power, but providing strong evidence of prejudice.

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 test provides robust 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 when it holds, 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.

We provide two extensions of our test. Our first extension is to allow for incorrect prior beliefs. A decision-maker's prior can be incorrect for a number reasons. For instance, their beliefs can be distorted by stereotyping (e.g., Coffman 2014) or because the experimental distribution of the outcomes deviates from the population in ways they are not aware of.[4] In this extension, we provide joint bounds on a decision-maker's preferences and prior beliefs.

[3]It is worth noting that outcomes are not always observed in experiments. For example, the quality of math questions studied in Bohren, Imas, and Rosenberg (2019) is subjective.

[4]See Bohren, Haggag, Imas, and Pope (2019) for a review of incorrect statistical discrimination.

These can either be combined with experimentally elicited information on prior beliefs to recover a test for prejudice or used to determine the set of prior beliefs that would imply a decision-maker is prejudiced.

Our second extension is to consider 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. 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.[5] 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.[6]

We demonstrate this second extension 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 robust 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. 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.

Our paper contributes to the literature on discrimination in three main ways. First, because our test is both general and simple, it can be applied widely, which we demonstrate using well-known discrimination experiments by Bertrand and Mullainathan (2004) and Reuben, Sapienza, and Zingales (2014). In both experiments, we show that there is evidence of prejudice which is robust to any form of private learning. In addition, our analysis provides new insights from these experiments. For instance, we show evidence of intersectional prejudice in Bertrand and Mullainathan (2004), which, to the best of our knowledge, has not been documented previously.

Second, by leveraging data on outcomes that is often available in experiments, we are able to offer an outcome-based test that does not require observing marginal decisions. In the first outcome test, Becker (1957) showed that a decision-maker is prejudiced if there

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

[6]This 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 Baert (2018).

are differences in outcomes across groups at the margin. For instance, his test identifies prejudice against 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 (inframarginal) 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. However, experimental data on outcomes allows us to simultaneously test (and reject) this implication, as well as offer a new test that does not suffer from the inframarginality problem of the Becker outcome test.

Third, by leveraging experimental data on outcomes we are also able to offer a test of prejudice which is more robust than other outcomes tests that circumvent the inframarginality problem. 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. Arnold, Dobbie, and Yang (2018) and Marx (2020) concurrently develop more powerful tests that jointly use information on decisions and outcomes.[7] 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 with differences in preferences. However, such assumptions have been questioned in some settings, such as the context of judicial decision-making (Frandsen, Lefgren, and Leslie 2019; Gelbach 2021), which suggests that robustness along this dimension could be valuable.

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 then provides a demonstration using the experiment of Reuben, Sapienza, and Zingales (2014). Section 4 provides our first extension to incorrect beliefs and a related demonstration of this extension. Section 5 provides our second extension to prejudice in selection motive and demonstrates this extension using the experiment of Bertrand and Mullainathan (2004). In the Appendix, we provide mathematical proofs.

[7] It is noteworthy that our test coincides with that of Marx (2020) in the case of "identification at infinity" (Heckman 1990): namely, when one of the decision-makers always makes the same decision (treats) and thus identifies the unconditional distribution of outcomes.

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 ($m$inority or $w$hite) or of different gender ($m$en or $w$omen),[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, we assume that the DM's beliefs are internally consistent with Bayes' Rule:

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

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:

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

We assume that the decision rule σ_g maximizes expected utility based on a possibly group-dependent and non-trivial 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 parametrize the utility function as:

$$u_g(d, s) = d[s - t_g] \tag{3}$$

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

[9] We consider the case of incorrect beliefs in Section 4.

[10] We consider the alternate case and study prejudicial disparities in this selection motive in Section 5.

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 $d = 1$.[11] We say that the DM exhibits *prejudice* against group w if:

$$t_w > t_m. \tag{4}$$

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

The 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 robust 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:*

$$P_g(s=1|d=0) \leq t_g \leq P_g(s=1|d=1). \tag{5}$$

Hence there is evidence of prejudice against group w if:

$$P_w(s=1|d=0) > P_m(s=1|d=1). \tag{6}$$

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. Thus, a necessary condition for our test to uncover any evidence of prejudice is that the outcome probabilities differ by group. Moreover, to uncover evidence of prejudice against a group the outcome probabilities must be higher for that group. For example, a necessary condition for uncovering

[11] Frankel and Kamenica (2018) call such problems *simple* decision problems, and show that every valid measure of information or uncertainty can be expressed as arising from a measure over such problems.

evidence of prejudice against women in the hiring context is that women 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.

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.

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.[12]

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

[12]Within 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. Employers were incentivized to hire the better-performing candidate in the subsequent task, who we label as having higher productivity. In total, the data analyzed from the experiment consists of 932 employer decisions over 76 mixed-gender candidate pairs.

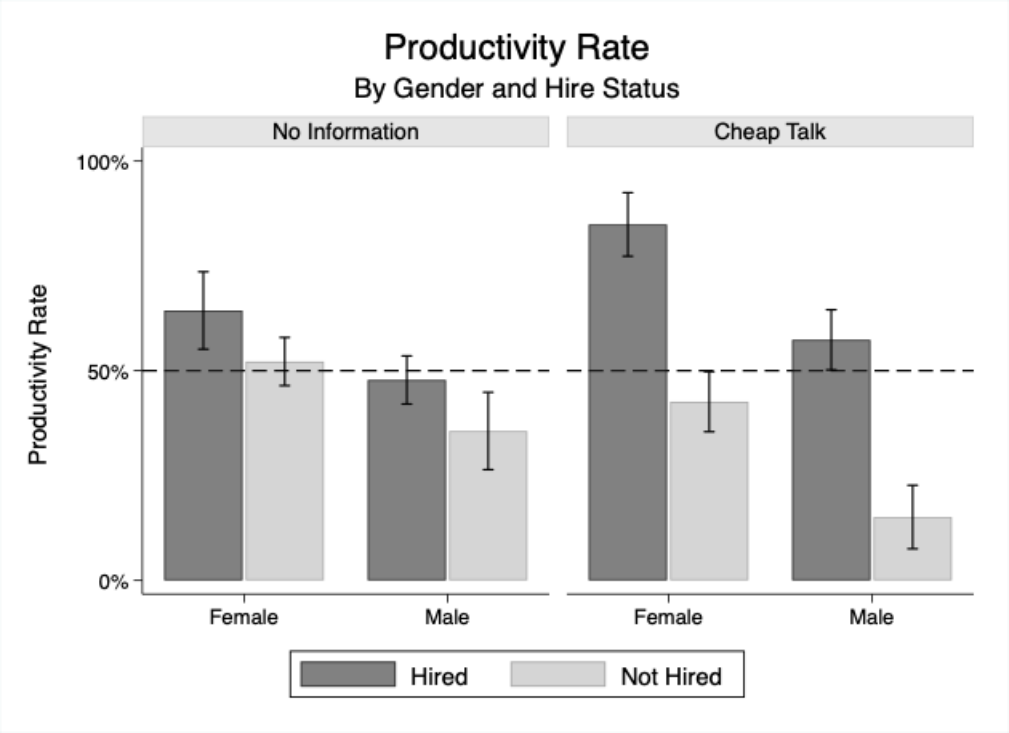

Figure 1: This figure illustrates our test for prejudice in the "Decision Then Cheap Talk" treatment of Reuben, Sapienza, and Zingales (2014). As in their analysis, standard errors are computed from a probit regression with random effects and clustering at the employer level. For initial decisions (before 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).

by statistical discrimination (with correct beliefs).[13] To answer this question, we apply our robust outcome test, which is established in Theorem 1. We begin by focusing attention on the experiment's "Decision Then Cheap Talk" treatment because our test provides robust 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

[13] In Subsection 4.1 we also discuss the case where prior beliefs may differ from the experimentally observed mean outcome.

are more productive (52.2%) than hired men (47.8%).[14] However, 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 ($p = 0.28$).[15] A necessary condition for even our suggestive finding of prejudice is that women are more productive than men, which is the case in this treatment (56.1% vs. 43.7%).

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%),[16] observed outcomes can 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. However, it is worth nothing that just because our test does not provide evidence of prejudice, this does not mean that employers are no longer prejudiced after receiving cheap talk information.

We conclude by relating our test to the experimental design. First, 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. Second, our test 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. Despite this, our robust test was able to provide evidence of prejudice in their experiment because women did perform slightly better than men at the task. That our test provides any evidence of prejudice is particularly stark because the performance rates are very similar across genders.

[14] A table that details the outcome probabilities for all treatments and decisions in the experiment is available in the Appendix.

[15] The 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.

[16] Comparing these hired outcome probabilities, the hit rate test of Knowles, Persico, and Todd (2001) would find significant evidence of prejudice against women, but their test is invalid in this context because of selection: hired employees are much more productive than unhired employees.

4 Incorrect Beliefs

The preceding tests for prejudice assumed that the DM's prior beliefs coincided with the observed distribution of the state for both groups. This can be violated because of stereotyping or because the distribution of the state in the experiment does not match the DM's expectations. The latter can happen, for instance, if the DM does not appreciate selection into the experiment. This can also happen in field studies when random assignment of quality by demographic group is independent of the distribution in the field (as is often the case in correspondence studies).

In what follows, we address these cases by relaxing the assumption of correct beliefs. Instead we will just assume that the prior be 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 1 (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:*

$$\frac{P_g(d=0|s=1)}{P_g(d=0|s=0)} \leq \frac{1-\mu_g}{\mu_g}\frac{t_g}{1-t_g} \leq \frac{P_g(d=1|s=1)}{P_g(d=1|s=0)} \tag{7}$$

One takeaway from Proposition 1 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.[17]

Nevertheless, Proposition 1 has useful applications. First, this 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, this 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.

4.1 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

[17] The indistinguishability between prior beliefs and 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 Yang (2018) and Bohren, Haggag, Imas, and Pope (2019).

evidence of such stereotyping by eliciting estimates of performance and implicit association biases.[18] This raises a natural question for our test: what (perhaps incorrect) prior beliefs would still imply the existence of taste-based prejudice?

Applying Proposition 1 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.

It is also insightful to apply Proposition 1 to the "Decision Then Past Performance" treatment, which differs from the "Decision Then Cheap Talk" treatment in that employers received information about performance on a previous task before making their second hiring decision.[19] 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 are on average less productive than men in this treatment (46.6% vs. 53.4%). However, based on Proposition 1 we can identify the threshold prior belief above which the observed initial hiring decisions would imply prejudice against women. Notably, the threshold belief (54.2%, s.e. = 2.1%) is very similar to the one in the "Decision Then Cheap Talk" treatment, even though the experimental distributions of outcomes differ. Finally, we observe that 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.

5 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 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

[18] Unfortunately, these point estimates are not sufficient to recover prior beliefs.

[19] This distinction is known participants before they make their initial decisions, which motivated our decision to consider these treatments separately. However, the remaining results are qualitatively the same if we pool initial decisions (before receiving additional information) across the two treatments.

state for group w and *against* the state for group m.[20] Formally, this means:

$$u_w(d,s) = d[s - t_w] \quad \text{and} \quad u_m(d,s) = -d[s - t_m]. \tag{8}$$

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. Importantly, 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, even in the case of incorrect beliefs.

Proposition 2. *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, selection for the state is identified by a strict ordering of conditional outcome probabilities:*

$$P_g(s = 1|d = 0) < P_g(s = 1|d = 1) \tag{9}$$

or decision probabilities:

$$P_g(d = 0|s = 1) < P_g(d = 1|s = 1). \tag{10}$$

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:

$$\begin{gathered} P_m(s = 1|d = 1) < P_m(s = 1|d = 0) \\ \textit{and} \\ P_w(s = 1|d = 1) > P_w(s = 1|d = 0). \end{gathered} \tag{11}$$

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 correspondence CV study, the researcher may devise "good" and "bad" resumes $\hat{s}$ which presumably 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.

[20] Motivated by our subsequent application, we use m and w in this section to denote *m*inority and *w*hite applicants, respectively.

Proposition 3. *Suppose the signal realization $\hat{s} = 1$ leads to stochastically higher posterior beliefs over the state (distributions of posteriors across signals are first-order stochastically ordered):*

$$\hat{\pi}_g(\cdot|\hat{s} = 1) \succeq_{FOSD} \hat{\pi}_g(\cdot|\hat{s} = 0). \tag{12}$$

Then the selection motive is identified as in Proposition 2 upon replacing the true but unobserved stated 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).

5.1 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.[21]

As is well-known, data on decisions alone cannot identify whether differences in decisions across groups is the result of preference-based prejudice, information-based statistical discrimination, or both. However, Proposition 3 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). Namely, 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.

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

[21] More 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.

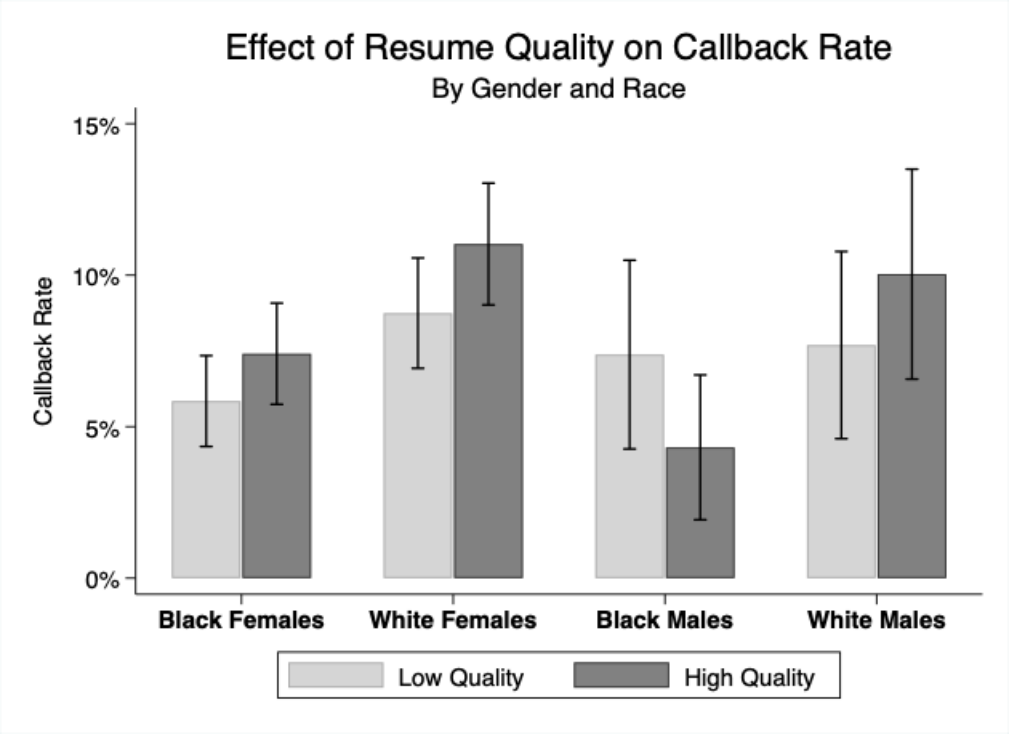

Figure 2: This figure plots the mean callback rates by subjective resume quality across race-gender pairs in the study of Bertrand and Mullainathan (2004). Similarly to their Table 5, standard errors are corrected for clustering at the employment-ad level in a probit regression of the callback dummy 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.

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 ($p = 0.063$).[22] 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 an artefact 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 robust evidence of preference-based prejudice in the selection motive.

We now discuss the empirical robustness and interpretation of this result. First, 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 rates for resumes with African-American names is weakly negative, albeit statistically indistinguishable from zero.[23] It is noteworthy that 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.[24] 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

[22] The 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 ($p = 0.047$).

[23] For 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.

[24] For 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.

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.

Appendix

A 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:*

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

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

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

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:

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

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 $\pi_g(\gamma|s=0) \gtreqless \pi_g(\gamma|s=1)$ for $\mu_g \gtreqless \gamma$. 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:

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

In that case, 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 1. 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:

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

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

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

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. □

Proof of Proposition 2. 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:

$$P_g(d=1|s=0) \leq P_g(d=1|s=1). \tag{20}$$

From the bounds in Proposition 1, we have:

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

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. □

Proof of Proposition 3. As in the proof of Proposition 2, 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:

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

The remaining results then follow as in Proposition 2. 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. □

B Tables

Table 1: Conditional Productivity Rates by Gender and Hire Status

	First Decision				Second Decision			
Gender	Female		Male		Female		Male	
Hired	No	Yes	No	Yes	No	Yes	No	Yes
	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
	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.

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

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

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.

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.

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.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg (2019). "The dynamics of discrimination: Theory and evidence". In: *American economic review* 109.10, pp. 3395–3436.

Chandra, Amitabh and Douglas O Staiger (2010). *Identifying provider prejudice in healthcare*. Tech. rep. National Bureau of Economic Research.

Charles, Kerwin Kofi and Jonathan Guryan (2011). "Studying discrimination: Fundamental challenges and recent progress". In: *Annu. Rev. Econ.* 3.1, pp. 479–511.

Coffman, Katherine Baldiga (2014). "Evidence on self-stereotyping and the contribution of ideas". In: *The Quarterly Journal of Economics* 129.4, pp. 1625–1660.

Frandsen, Brigham R, Lars J Lefgren, and Emily C Leslie (2019). *Judging judge fixed effects*. Tech. rep. National Bureau of Economic Research.

Frankel, Alexander and Emir Kamenica (2018). "Quantifying information and uncertainty". In: *American Economic Review*.

Gelbach, Jonah B (2021). "Testing Economic Models of Discrimination in Criminal Justice". In: *Available at SSRN 3784953*.

Heckman, James (1990). "Varieties of selection bias". In: *The American Economic Review* 80.2, pp. 313–318.

— (1998). "Detecting discrimination". In: *Journal of economic perspectives* 12.2, pp. 101–116.

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.

Lang, Kevin and Ariella Kahn-Lang Spitzer (2020). "Race Discrimination: An Economic Perspective". In: *Journal of Economic Perspectives* 34.2, pp. 68–89.

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.

Persico, Nicola (2009). "Racial profiling? Detecting bias using statistical evidence". In: *Annu. Rev. Econ.* 1.1, pp. 229–254.

Phelps, Edmund S (1972). "The statistical theory of racism and sexism". In: *The american economic review* 62.4, pp. 659–661.

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*.