http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 68, No. 6, June 2022, pp. 4527–4536
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

A Robust Test of Prejudice for Discrimination Experiments

Daniel Martin,[a] **Philip Marx**[b]

[a] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208; [b] Department of Economics, Louisiana State University, Baton Rouge, Louisiana 70803

Contact: d-martin@kellogg.northwestern.edu, https://orcid.org/0000-0001-6483-3923 (DM); philiplmarx@gmail.com (PM)

Received: July 14, 2021
Revised: October 20, 2021
Accepted: December 7, 2021
Published Online in Articles in Advance: March 30, 2022

https://doi.org/10.1287/mnsc.2022.4396

Abstract. A large experimental literature is 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 learn 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 two influential labor market experiments.

History: This paper was accepted by Yan Chen, behavioral economics and decision analysis.
Supplemental Material: The online appendix and data are available at https://doi.org/10.1287/mnsc.2022.4396.

Keywords: discrimination • prejudice • experiments • inattention • labor markets

1. Introduction

Discrimination is a pressing issue for society and the management and regulation of firms, and a large experimental literature is devoted to its study. Experiments—in both the laboratory and the field—provide 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 et al. (2014) find that employers in a laboratory 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 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 to call back that candidate? This identification challenge is especially pronounced in settings in which discrimination is impacted by selective attention that depends on group identity. For example, Bertrand and Mullainathan (2004, p. 1011) 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 to which decision makers attend, it becomes even harder to determine what factors enter into the decision maker's choices.

We overcome this identification challenge by deriving conditions on outcomes that reveal prejudice regardless of what decision makers learn about individuals in each demographic 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* the decision maker's 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 the decision maker 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 laboratory experiment of Reuben et al. (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 the "decision then cheap talk" treatment (in which initial hiring decisions are 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 the 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 the threshold belief for hiring men, indicating that the employer is prejudiced against women. This conclusion holds regardless of what employers learn about male and female candidates based on their appearance.[4]

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 antidiscrimination 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) is not 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 glean from the resumes they receive. 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 et al. (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) shows that a decision maker is prejudiced if there 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 the hired applicants of another group. However, a limitation of the Becker test is that it is often difficult to identify marginal decisions, and it is shown that the test can produce misleading conclusions about prejudice if it is applied to average (inframarginal) outcomes (see Ross and Yinger 1999, Ayres 2002). In a groundbreaking paper, Knowles et al. (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 in which 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 et al. (2018) and Marx (2022) 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 et al. 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 et al. (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 the individual's 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, $\pi_g(\gamma \mid s)$, with the unconditional distribution of posteriors denoted by $\pi_g(\gamma) = \mu_g \pi_g(\gamma \mid s = 1) + (1 - \mu_g)\pi_g(\gamma \mid s = 0)$. The prior and information structure each may vary by group. However, the DM's beliefs are internally consistent with Bayes' rule:

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

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 $\sigma_g(d \mid \gamma)$. The joint probability of deciding d in state s is, thus,

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

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

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

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

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

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 in which $P_g(d,s) \in (0,1)$ 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

Our test for prejudice bounds the threshold t_g by the conditional outcome probabilities given by $P(s = 1 \mid d)$ 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:*

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

Hence, there is evidence of prejudice against group w if

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

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

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 are 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 on which firms, regulators, and researchers may wish to focus 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 in the online appendix). 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 et al. (2014) investigate 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. Because 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.

Their study finds 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

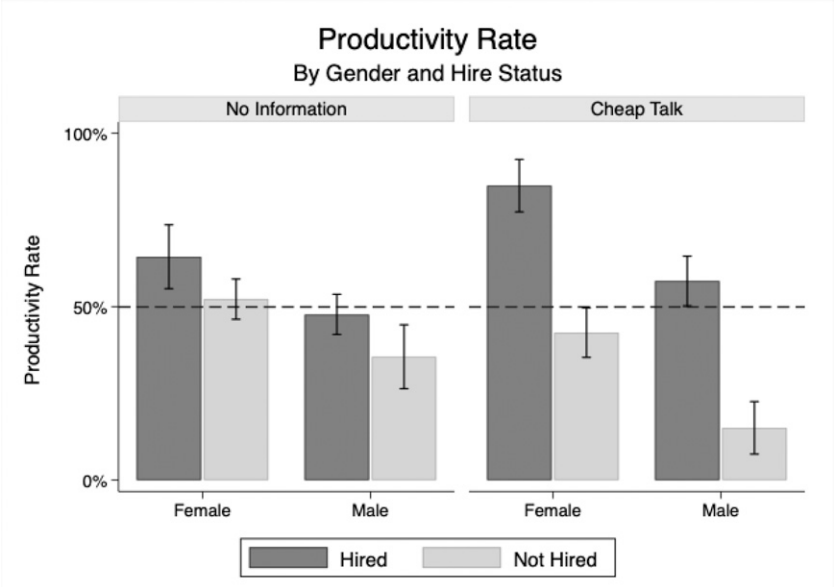
discrimination. To answer this question, we apply our robust outcome test established in Theorem 1. We begin by focusing attention on the experiment's "decision then cheap talk" treatment because our test provides evidence of prejudice in this treatment and decisions in this treatment provide a useful 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, whereas 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. Additionally, Table 1 in the online appendix provides a tabular summary of productivity rates and their standard errors across treatments. 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% versus 57.4%). As a result, whereas hired women are significantly more productive than hired men (84.9% versus 57.4%),[14] 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. 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

Figure 1. Test for Prejudice in the "Decision Then Cheap Talk" Treatment of Reuben et al. (2014)



Notes. As in their analysis, standard errors are computed from a probit regression with random effects and clustering at the employer level. Because 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).

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 et al. (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" and the "decision then past performance" treatments 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 the online appendix.

4. Selection Motive

So far, we assume 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 state for group w and *against* the state for group m.[16] Formally, this means

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

For example, a prejudiced employer who has to comply with antidiscrimination laws may do so in letter but not in spirit by calling back (or hiring) White applicants who are more likely to be productive and 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*

$$P_g(s = 1 \mid d = 0) < P_g(s = 1 \mid d = 1) \tag{8}$$

or decision probabilities

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

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:

$$P_m(s = 1 \mid d = 1) < P_m(s = 1 \mid d = 0)$$
$$\text{and}$$
$$P_w(s = 1 \mid d = 1) > P_w(s = 1 \mid d = 0). \tag{10}$$

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 $\hat{\pi}_g(\gamma \mid \hat{s})$ 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 $\hat{s}$ that correlate with the true productivity or qualifications s for which employers want to select. 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):*

$$\hat{\pi}_g(\cdot \mid \hat{s} = 1) \geq_{FOSD} \hat{\pi}_g(\cdot \mid \hat{s} = 0). \tag{11}$$

Then, the selection motive is identified as in Proposition 1 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).

4.1. Empirical Application

In an influential study, Bertrand and Mullainathan (2004) randomly assign names that strongly signal race and gender to fictitious resumes and find 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

finds that the returns to resume quality were lower for candidates with African American names.[17]

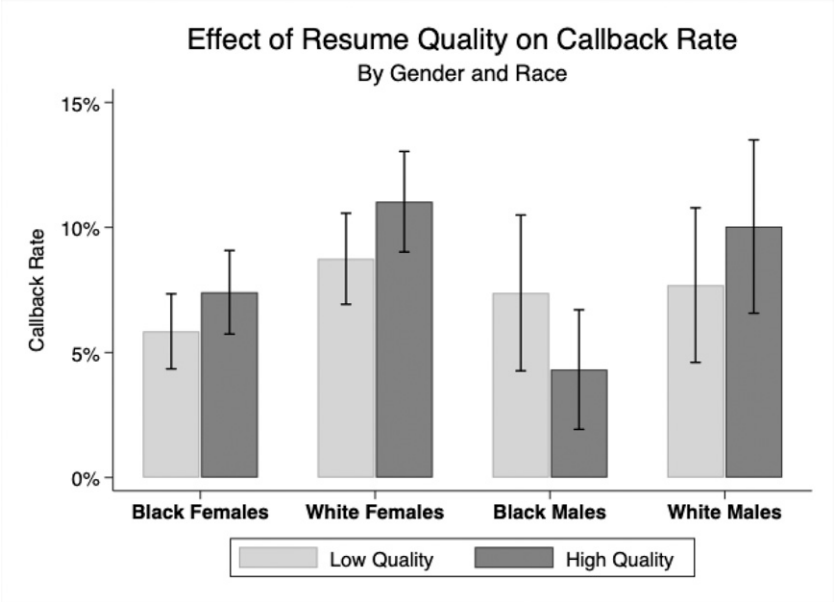
As is well-known, data on decisions alone cannot identify whether a difference in decisions across groups is the result of preference-based prejudice, information-based statistical discrimination, or both. However, Proposition 2 provides a test of preference-based prejudice in selection motive if we 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.

Figure 2 plots the callback rates across resume quality for each intersectional group. Additionally, Table 2 in the online appendix provides a tabular summary of callback rates and their standard errors across intersectional groups, cities, and job types. 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 difference 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

Figure 2. Mean Callback Rates by Subjective Resume Quality Across Race–Gender Pairs in the Study of Bertrand and Mullainathan (2004)



Notes. Similar 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.

White applicants yet—perhaps to abide by antidiscrimination 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 underprivileged.[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, 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 in which 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, by which the DM is conservative in updating. In that case, if unhired women are more productive than hired men, the highest (incorrect) posterior belief at which women are not hired is still above the lowest (incorrect) posterior belief at which hired men are hired.[24] This is also 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 et al. (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, the beliefs can be distorted by stereotyping (e.g., Bordalo et al. 2016) or because the experimental distribution of the outcomes deviates from the population in ways of which decision makers are unaware. 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 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 online appendix, we provide and empirically apply a result (Proposition 3 in the online appendix) that expresses joint bounds on a decision maker's preferences and prior beliefs. These bounds can be either 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 the online appendix

Acknowledgments

The authors 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.

Endnotes

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

[2] Bartoš et al. (2016) develop a related model of discrimination based on selective attention and use it to explain empirical disparities in labor and housing markets.

[3] Our test can be extended beyond binary decisions and states by applying the "no improving action switches" 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.

[4] This 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 the online appendix, we elaborate on the role of prior beliefs in our test.

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

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

[7] This allows us to use our test to infer prejudice in correspondence studies that exogenously vary observable nondemographic characteristics that correlate with quality. For a review of correspondence studies, see Quillian et al. (2017), Baert (2018), and Gaddis (2018).

[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 prior beliefs in Section 5.

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

[11] 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. In total, the data analyzed from the experiment consists of 932 employer decisions over 76 mixed-gender candidate pairs.

[12] A table that details the outcome probabilities for all treatments and decisions in the experiment is available in the online appendix.

[13] 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 the 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.

[14] Comparing these hired outcome probabilities, the hit rate test of Knowles et al. (2001) 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.

[15] In 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 in the online appendix.

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

[17] 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 a 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.

[18] 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 ($p = 0.047$).

[19] In the online appendix, we provide a robustness analysis for this result.

[20] We are grateful to Roland Bénabou for this suggestion.

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

[22] We are grateful to Sigrid Suetens for this suggestion.

[23] See Bohren et al. (2019) for a review of the literature on incorrect statistical discrimination and Bursztyn and Yang (2021) for a recent meta-analysis of field experiments documenting a broad pattern of group-based misperceptions in beliefs.

[24] See Benjamin (2019) for a review of belief biases and Albrecht et al. (2013) for experimental evidence of discrimination being driven by conservatism in belief updating.

[25] The indistinguishability between prior beliefs and taste thresholds has been discussed previously in the context of healthcare (Chandra and Staiger 2010, Abaluck et al. 2016) and is also formalized in Arnold et al. (2018).

References

Abaluck J, Agha L, Kabrhel C, Raja A, Venkatesh A (2016) The determinants of productivity in medical testing: Intensity and allocation of care. *Amer. Econom. Rev.* 106(12):3730–3764.

Albrecht K, Von Essen E, Parys J, Szech N (2013) Updating, self-confidence, and discrimination. *Eur. Econom. Rev.* 60:144–169.

Anderson L, Fryer R, Holt C (2006) Discrimination: Experimental evidence from psychology and economics. Rodgers WM, ed. *Handbook on the Economics of Discrimination* (Elger, Cheltenham, UK), 97–118.

Anwar S, Fang H (2006) An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *Amer. Econom. Rev.* 96(1):127–151.

Arnold D, Dobbie W, Yang CS (2018) Racial bias in bail decisions. *Quart. J. Econom.* 133(4):1885–1932.

Arrow K (1971) The theory of discrimination, Technical report, Princeton University, Princeton, NJ.

Ayres I (2002) Outcome tests of racial disparities in police practices. *Justice Res. Policy* 4(1–2):131–142.

Baert S (2018) Hiring discrimination: An overview of (almost) all correspondence experiments since 2005. Gaddis SM, ed. *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (Springer International, Cham, Switzerland), 63–77.

Bartoš V, Bauer M, Chytilová J, Matějka F (2016) Attention discrimination: Theory and field experiments with monitoring information acquisition. *Amer. Econom. Rev.* 106(6):1437–1475.

Becker GS (1957) *The Economics of Discrimination* (University of Chicago Press, Chicago).

Benjamin DJ (2019) Errors in probabilistic reasoning and judgment biases. Bernheim DB, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics: Foundations and Applications 2* (Elsevier, Amsterdam), 69–186.

Bertrand M, Duflo E (2017) Field experiments on discrimination. Banerjee VB, Duflo E, eds. *Handbook of Economic Field Experiments* (Elsevier, Amsterdam), 309–393.

Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Amer. Econom. Rev.* 94(4):991–1013.

Bohren JA, Haggag K, Imas A, Pope DG (2019) Inaccurate statistical discrimination. Technical report, National Bureau of Economic Research, Cambridge, MA.

Bordalo P, Coffman K, Gennaioli N, Shleifer A (2016) Stereotypes. *Quart. J. Econom.* 131(4):1753–1794.

Bursztyn L, Yang DY (2021) Misperceptions about others. Technical report, National Bureau of Economic Research, Cambridge, MA.

Caplin A, Martin D (2015) A testable theory of imperfect perception. *Econom. J. (London)* 125(582):184–202.

Chandra A, Staiger DO (2010) Identifying provider prejudice in healthcare. Technical report, National Bureau of Economic Research, Cambridge, MA.

Frandsen BR, Lefgren LJ, Leslie EC (2019) Judging judge fixed effects. Technical report, National Bureau of Economic Research, Cambridge, MA.

Gaddis SM (2018) *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, vol. 14 (Springer International, Cham, Switzerland).

Gelbach JB (2021) Testing economic models of discrimination in criminal justice. Preprint, submitted February 18, https://dx.doi.org/10.2139/ssrn.3784953.

Knowles J, Persico N, Todd P (2001) Racial bias in motor vehicle searches: Theory and evidence. *J. Political Econom.* 109(1):203–229.

Lane T (2016) Discrimination in the laboratory: A meta-analysis of economics experiments. *Eur. Econom. Rev.* 90:375–402.

Marx P (2022) An absolute test of racial prejudice. *J. Law, Econom., Organ.* 38(1):42–91.

Neumark D (2018) Experimental research on labor market discrimination. *J. Econom. Literature* 56(3):799–866.

Phelps ES (1972) The statistical theory of racism and sexism. *Amer. Econom. Rev.* 62(4):659–661.

Quillian L, Pager D, Hexel O, Midtbøen AH (2017) Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proc. Natl. Acad. Sci. USA* 114(41):10870–10875.

Rambachan A (2021) Identifying prediction mistakes in observational data. Working paper, Harvard University Department of Economics, Cambridge, MA.

Reuben E, Sapienza P, Zingales L (2014) How stereotypes impair women's careers in science. *Proc. Natl. Acad. Sci. USA.* 111(12): 4403–4408.

Riach PA, Rich J (2002) Field experiments of discrimination in the market place. *Econom. J.* 112(483):F480–F518.

Ross SL, Yinger J (1999) The default approach to studying mortgage discrimination: A rebuttal. Turner MA, Skidmore F, eds. *Mortgage Lending Discrimination: A Review of Existing Evidence* (Urban Institute, Washington, DC), 107–127.

Wozniak D, MacNeill T (2020) Racial discrimination in the lab: Evidence of statistical and taste-based discrimination. *J. Behav. Experiment. Econom.* 85:101512.