INACCURATE STATISTICAL DISCRIMINATION: AN IDENTIFICATION PROBLEM

J. Aislinn Bohren, Kareem Haggag, Alex Imas, and Devin G. Pope*

Abstract—We study inaccurate beliefs as a source of discrimination. Economists typically characterize discrimination as stemming from a taste-based (preference) or accurate statistical (belief-based) source. Although individuals may have inaccurate beliefs about how relevant characteristics (e.g., productivity, signals) are correlated with group identity, fewer than 7% of empirical discrimination papers in economics consider the possibility of such *inaccurate statistical discrimination.* Using theory and a labor market experiment, we show that failing to account for inaccurate beliefs leads to a misclassification of source. We outline three methods to identify source: varying observed signals, belief elicitation, and an intervention to target inaccurate beliefs.

I. Introduction

DISCRIMINATION based on group identity has been shown to be prevalent in many important settings, including labor markets, housing markets, credit markets, and online consumer markets (see Charles & Guryan, 2011 and Bertrand & Duflo, 2017 for reviews). Economists studying direct discrimination—that is, the causal link between group identity and differential treatment—often also seek to identify its source.[1] Sources are typically categorized into one of two forms. In the case of taste-based discrimination (Becker, 1957), an individual has animus toward members of a particular group and discriminates against them because he receives disutility from providing services to or interacting with members of this group. In the case of accurate statistical discrimination (Phelps, 1972; Aigner & Cain, 1977), differential treatment occurs because productivity is unobserved and a particular group's distribution of productivity or signal of productivity is perceived to differ, relative to an alternative group.[2]

Although statistical discrimination is typically assumed to be driven by rational expectations, a large literature in psychology and economics has shown that people's beliefs are often incorrect (e.g., Kravitz & Platania, 1993; Fiske, 1998). This motivates the topic of the current paper, which studies the role of inaccurate beliefs about the distribution of productivity or signal of productivity in driving discrimination. Using a theoretical framework and an experimental labor market setting, we demonstrate the importance of accounting for inaccurate beliefs when classifying the source of discrimination. We show that inaccurate beliefs give rise to similar patterns in the data as taste-based sources; in turn, commonly used methods cannot disentangle inaccurate beliefs from preferences as a driver of discrimination. Moreover, failure to account for inaccurate beliefs can lead to a potential misclassification of source. We outline three alternative identification methods: eliciting beliefs, varying the number of signal draws, and providing direct information about the productivity distribution.

We start with a systematic review of the discrimination literature. Although a large plurality of empirical discrimination papers (61.9%) attempt to differentiate between taste-based versus statistical sources, only a small proportion (10.5%) discuss the possibility of inaccurate beliefs. Yet identifying the source of discrimination is important for a myriad of reasons: designing an effective policy intervention to reduce discrimination crucially depends on its driver,[3] welfare and efficiency analyses differ with the source, and the extent to which competitive markets will eliminate discrimination depends on the source (see Fang & Moro, 2011 for review). Additionally, when discrimination stems from inaccurate beliefs or preferences, it can lead to further discrimination by *other* people (or algorithms) who learn from the decisions of the discriminators but are unaware of their bias (Bohren et al., 2022).

To formalize how the possibility of inaccurate beliefs impacts identification, we first develop a theoretical framework for modeling inaccurate statistical discrimination. Consider an evaluator who observes the group identity of a worker as well as a signal about her productivity, then decides whether to hire the worker. *Direct discrimination* occurs when two workers who generate identical signals are evaluated differently based on their group identity. This discrimination can stem from either belief-based partiality—where evaluators have group-dependent beliefs about the productivity and/or the signal distributions—or preference-based

Received for publication August 5, 2021. Revision accepted for publication December 30, 2022. Editor: Raymond Fisman.

*Bohren: University of Pennsylvania; Haggag: UCLA; Imas and Pope: University of Chicago.

We thank Steven Durlauf, Hanming Fang, Alex Frankel, Emir Kamenica, Emily Nix, and seminar participants at Harvard Business School, Harvard Kennedy School, the SaMMF Discrimination in Labor Markets Workshop, Stanford University, UCLA, University of Chicago, University of Melbourne, University of Pennsylvania, University of Southern California, University of Sydney, University of Virginia, Cambridge University, and the Virtual Market Design Seminar for helpful comments and suggestions. Cuimin Ba, Byunghoon Kim, and Jihong Song provided excellent research assistance. Bohren gratefully acknowledges financial support from NSF grant SES-1851629. The experiment received IRB approval at CMU.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_01367.

[1] Bohren et al. (2022) considers the role of both direct and indirect, that is, *systemic* sources of discrimination in generating group-based disparities. The current paper focuses on the former channel.

[2] Differences in the productivity distribution may be due to exogenous differences (Phelps, 1972) or part of a self-fulfilling equilibrium (Arrow, 1973).

[3] For example, if discrimination stems from inaccurate beliefs, an effective policy response could be providing individuals with information about the correct distributions, whereas such a policy would have no effect when discrimination stems from the other two sources. See, for example, Jensen (2010) in the case of inaccurate beliefs about the returns to education, or Bursztyn et al. (2020) in the case of inaccurate beliefs about the beliefs of others, that is, pluralistic ignorance.

The Review of Economics and Statistics, May 2025, 107(3): 1–16

https://doi.org/10.1162/rest_a_01367

partiality—where evaluators have group-dependent preferences over hiring workers of a given expected productivity. The former is typically referred to as statistical discrimination, whereas the latter is referred to as taste-based discrimination, prejudice, or animus. We expand this standard framework by considering both accurate and inaccurate beliefs about the productivity and signal distributions.

We first characterize the set of preferences and beliefs that result in equivalent discrimination—that is, the same group-dependent hiring decisions. It is readily apparent that a continuum of preference and belief profiles can give rise to equivalent discrimination.[4] Therefore, identifying the level of discrimination does not identify its source. In fact, it does not even rule out any form of inaccurate beliefs (e.g., inaccurate beliefs about the signal precision versus average productivity). Therefore, additional data are necessary to isolate the source.

It may be that a researcher also has access to the true productivity and signal distributions. In such cases, studies have used a method often referred to as an outcomes-based test that compares evaluation decisions to the true distributions.[5] For example, a researcher may compare differences in lending rates between two groups to differences in their loan default rates. When maintaining the assumption that evaluators have accurate beliefs, this method pins down the source of discrimination. However, identification depends critically on this assumption: without it, we show that the only source that can be ruled out is accurate statistical discrimination, that is, an evaluator with accurate beliefs and no preference partiality.[6] Moreover, erroneously assuming that an evaluator has accurate beliefs leads a researcher to *mistakenly* attribute the share of discrimination arising from inaccurate beliefs to preferences. For example, suppose a researcher finds evidence for racial discrimination in lending decisions. If the researcher observes that Black and white borrowers have an identical loan default rate (productivity) and signal of likely default distributions, and assumes that loan officers have correct beliefs about these distributions, then she will conclude that the source of the observed discrimination must be preference-based. However, an alternative explanation is that loan officers have incorrect beliefs, which leads to inaccurate statistical discrimination. Without further data, it is impossible to distinguish between these explanations.

We next outline alternative methods for identifying the source of discrimination. One method, as proposed by Manski (2004), is to directly collect data on the subjective beliefs of evaluators. Combined with observing the evaluation decisions and signals, this identifies preferences. Data on the true distributions are also required to determine whether beliefs are accurate.[7] In many settings, eliciting beliefs will not be feasible. An alternative method is to manipulate the signal precision by varying the number of signal draws observed by evaluators. For example, one could vary the number of recommendation letters for a job candidate or the number of reviews on a platform such as TaskRabbit. We demonstrate that this method can partially identify the source of discrimination: it identifies the extent of preference-based partiality but cannot distinguish between different forms of belief-based partiality (i.e., different beliefs about average productivity versus signal precision). Importantly, this method requires multiple signals from the same domain (e.g., reviews from the same population of evaluators); if the signals are from different domains (e.g., SAT scores and education history), then the identification problem persists.[8]

We next demonstrate the identification issue and alternative methods in a stylized hiring experiment. Participants are recruited and assigned to the role of either "worker" or "employer." Workers created profiles that included a variety of characteristics, such as their country of origin (United States vs. India), gender, and age, along with other information such as their beverage and movie preferences. They then completed a series of logical reasoning questions. Employers were shown the profiles of twenty workers and asked the maximum wage they would be willing to pay to hire each worker. The employer's payoff depended on her offered wage and how many questions the worker answered correctly.

We find that employers discriminate based on the worker's country of origin and gender: Americans and females received systematically lower wage offers than Indians and males. According to the standard classification, the observed discrimination is generated by two potential sources. Employers may offer lower wages to American and female workers because they believed that members of those groups answer fewer questions correctly on average than Indian and male workers. They lack information on the productivity of any given worker, and so employers used these group

[4]Manski (2004) first illustrated that observed choice behavior could be consistent with multiple sets of preferences and subjective beliefs, and hence, identification of preferences from choice data required strong assumptions, such as rational expectations. He proposed that data on expectations could be used to validate or relax the rational expectations assumption.

[5]This commonly used method has been employed in many domains, including lending, policing, and bail decisions (Knowles et al., 2001; Antonovics & Knight, 2009; Pope & Sydnor, 2011; Arnold et al., 2022).

[6]Recent work by Arnold et al. (2018) and Grau and Vergara (2021) consider a different type of outcome-based test that does not assume that the researcher observes the decision maker's signals. Using IV and marginal treatment effect (MTE) methods, these tests can also reject accurate statistical discrimination (Hull, 2021), but the identification problem outlined in the current paper remains.

[7]A growing empirical literature elicits expectations to separate preferences from subjective expectations, including birth control choices (Delavande, 2008), college major choices (Wiswall & Zafar, 2015), and secondary education choices (Giustinelli, 2016).

[8]Another approach that we do not study in this paper derives predictions from a specific structural model of biased beliefs and takes these predictions to the data. For example, Arnold et al. (2018) compare the distributions of pretrial misconduct of marginal Black and white defendants. They argue that the distributional differences are consistent with bail judges holding incorrect stereotypes about the release risk of Black defendants. Bohren et al. (2019) model how discrimination evolves across evaluation rounds in a social learning setting, and argue that the observed dynamics are consistent with discrimination driven by inaccurate belief partiality but inconsistent with accurate belief partiality or preference partiality.

statistics to inform their compensation decisions. Alternatively, employers may be prejudiced toward members of the discriminated group and offered them lower wages because they did not want to reward them.

As discussed above, outcomes-based tests are often used to distinguish between these sources by comparing the compensation decisions to the "ground truth"—the true performance distributions by group. Our experiment allows us to measure the "ground truth" by comparing the number of questions answered correctly across the various groups. We find that, if anything, Americans slightly outperform Indians on the task (although the difference is not statistically significant), and females perform less well than males. Under the assumption of accurate beliefs, we would conclude that the discrimination against Americans is due to preference partiality. Further, because the level of discrimination against females is substantially smaller than the actual gap in performance, this approach would conclude that evaluators have preference partiality against men.

However, an alternative explanation is that individuals have no preference partiality toward or against a particular group but, rather, have inaccurate beliefs about the respective performance distributions. To identify this channel, we elicited the beliefs of employers and compared them to the "ground truth." Consistent with inaccurate statistical discrimination, employers mistakenly predicted that American workers perform much worse than their Indian counterparts, and that female workers only slightly underperform relative to males. Accounting for these inaccurate beliefs substantially changes the inferred source of discrimination. What was originally classified as preference-based discrimination *in favor of* Indians is mostly explained by mistaken beliefs—if anything, the preference-based channel goes slightly *against* Indian workers. Similarly, a large portion of the gender gap in wages can be explained by inaccurate statistical discrimination.

The line between inaccurate beliefs and animus may sometimes be blurry. For example, individuals may develop inaccurate beliefs *because* they have animus against members of a particular group. We propose that these channels can be separately identified through the provision of information about the relevant distributions. Specifically, if agents are provided with credible information on how the productivity or signal distributions vary by group, those with inaccurate beliefs should update their beliefs and adjust their behavior accordingly. However, if mistaken beliefs merely mask an underlying animus, then agents are unlikely to change their behavior in response to such information. We implement this method in our experiment by providing employers with information on average performance by gender, nationality, and age. After receiving this information, participants were asked to make wage offers to ten additional workers. We find that employers significantly changed their wage offers in the direction consistent with correcting their beliefs. This methodology is portable outside of our stylized experimental setting as a way to identify animus-driven inaccurate beliefs versus inaccurate beliefs stemming from inexperience or a lack of information.

The paper proceeds as follows. Section II presents a review of the economics literature on discrimination, and section III outlines our theoretical framework and results. Section IV illustrates these findings in a stylized hiring experiment. Section V concludes.

II. Survey of the Literature

We conducted a systematic survey of the empirical economics literature on discrimination to determine (1) how often papers seek to distinguish between taste-based and belief-based (statistical) sources of discrimination and (2) how often papers seek to distinguish between accurate and inaccurate beliefs for belief-based sources of discrimination. We find that few papers consider mistaken beliefs when attempting to isolate its source.[9]

Our survey focused on papers that tested for evidence of discrimination and published in ten top economics journals between 1990 and 2018 (the exact methodology and inclusion criteria are outlined in online appendix C). Of the 105 papers that met our inclusion criteria, most found evidence of discrimination: 102 out of 105 papers, or 97.1%, documented evidence for discrimination against at least one group that was considered in the paper. The majority of papers (61.9%) discussed the source of discrimination as being driven by either preferences (taste-based) or beliefs (statistical), and nearly half of the papers (46.7%) attempted to formally distinguish between these two sources. However, very few papers discussed the possibility that beliefs may be inaccurate (10.5%), and fewer still tested whether beliefs were accurate or inaccurate (6.7%).[10] Despite the lack of discussion and explicit tests, we would argue that inaccurate statistical discrimination is a reasonable alternative interpretation in nearly all of these cases. See table C1 in online appendix C for these and other summary statistics from our survey.

III. A Model of Discrimination with Inaccurate Beliefs

In this section, we model discrimination with inaccurate beliefs in the context of a simple hiring decision. An evaluator learns about a worker's productivity from a signal, then decides whether to hire the worker. Inaccurate beliefs refer to the evaluator's misperception of how the distribution

[9] Although our survey focuses on the empirical literature, the potential for inaccurate beliefs has also been discussed in theoretical discrimination research (Arrow, 1973, 1998; Schwartzstein, 2014). Arrow (1973) notes that employers may be more willing to accept subjective probabilities that accord with their actions. Arrow (1998) writes "the discussion of statistical discrimination so far assumes that the employers or creditors use all the information available throughout the economy…. But of course this is not so." Neither paper formally models inaccurate beliefs.

[10] The seven papers that tested for inaccurate beliefs are Fershtman and Gneezy (2001), List (2004), Mobius and Rosenblat (2006), Beaman et al. (2009), Agan and Starr (2017), Arnold et al. (2018), and Hedegaard and Tyran (2018). We discuss the methods and findings of these papers further in online appendix C.

of productivity or the signal distribution varies by group identity. We use this model to explore how a researcher can identify the source of discrimination. We first show that many different preferences and beliefs generate an identical pattern of discrimination, creating an identification challenge. We then show that, when allowing for inaccurate beliefs, the commonly used outcomes-based method can reject the possibility of accurate statistical discrimination, but it cannot separate whether discrimination stems from preferences or inaccurate beliefs. Further, erroneously assuming accurate beliefs and using this method leads to a misclassification of source. We conclude by outlining two alternative methods—eliciting beliefs and multiple informational treatments—to identify source. A reader who prefers to skip the formal presentation of the theory can jump to the empirics in section IV. All proofs from this section are in online appendix A.

A. Model

Workers. Consider a worker who has observable group identity $g \in \{M, F\}$ and unobservable productivity a drawn from normal distribution $N(\mu_g, 1/\tau_g)$, with mean productivity $\mu_g \in \mathbb{R}$ and concentration of productivity $\tau_g > 0$. The worker completes a task, such as an interview or test, that generates a signal of productivity $s = a + \epsilon$, where $\epsilon \sim N(0, 1/\eta_g)$ with signal precision $\eta_g > 0$.[11] Without loss of generality, we focus on discrimination against workers from group F.

Evaluators. An evaluator decides whether to hire the worker, $v \in \{0, 1\}$, where 1 corresponds to hire and 0 corresponds to do not hire. Before making this decision, the evaluator observes the worker's group identity g and realized signal s.

We model inaccurate beliefs as a misspecified model of the group-specific productivity and signal distributions. Namely, the evaluator holds subjective beliefs $\hat{\mu}_g \in \mathbb{R}$ and $\hat{\tau}_g > 0$ about the mean and concentration of productivity for group g, and subjective belief $\hat{\eta}_g > 0$ about the precision of the signal for group g. Inaccurate beliefs correspond to the case in which these subjective distributions differ from the true distributions.[12] The evaluator uses Bayes's rule with respect to these subjective distributions to form a posterior belief about the worker's productivity. From Bohren and Hauser (2021), this misspecified model framework can capture a variety of biases and heuristics in belief formation that have been documented in the literature, including non-Bayesian updating rules.

[11] To distinguish the two variance parameters, we refer to τ_g as the concentration of productivity and to η_g as the signal precision.

[12] An additional form of inaccurate beliefs that we do not consider is the possibility that an evaluator believes that the mean of the signal differs by group identity. For example, all signals for group F are inflated by a constant $b > 0$ that is, $s = a + b + \epsilon$, and therefore, the evaluator discounts a signal to $s - b$ for group F.

The evaluator hires the worker if her subjective posterior belief about expected productivity is above a group-specific hiring threshold $u_g \in \mathbb{R}$. This threshold is a reduced form representation of how the evaluator's preferences depend on productivity and group identity.[13] We refer to the evaluator's preferences and subjective beliefs as her type, denoted by $\theta \equiv (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g\in\{M,F\}}$. Let $v(s, g, \theta) \equiv \mathbb{1}\{\hat{E}_\theta[a|s, g] \geq u_g\}$ denote the optimal hiring decision by an evaluator of type θ who observes a worker from group g with realized signal s, where $\hat{E}_\theta$ denotes the expectation taken with respect to θ's subjective beliefs.

We next categorize different forms of preferences and beliefs. We use the term *partiality* to refer to properties of these model primitives. An evaluator with *preference partiality* sets different expected productivity thresholds for hiring workers from groups F and M.

Definition 1 (Preference Partiality). *An evaluator has preference partiality against group F if $u_F > u_M$, preference partiality against group M if $u_M > u_F$, and preference neutrality if $u_F = u_M$.*

Preference partiality leads the evaluator to make different hiring decisions when she has the same posterior belief about the expected productivity of a worker from each group. An evaluator with *belief partiality* has different subjective beliefs about the productivity and/or signal distributions for each group.

Definition 2 (Belief Partiality). *An evaluator has* belief partiality *if* $(\hat{\mu}_F, \hat{\tau}_F, \hat{\eta}_F) \neq (\hat{\mu}_M, \hat{\tau}_M, \hat{\eta}_M)$ *and* belief neutrality *if* $(\hat{\mu}_F, \hat{\tau}_F, \hat{\eta}_F) = (\hat{\mu}_M, \hat{\tau}_M, \hat{\eta}_M)$. *This belief partiality stems from (i) lower expected productivity if $\hat{\mu}_F < \hat{\mu}_M$; (ii) lower (higher) concentration if $\hat{\tau}_F < \hat{\tau}_M$ ($\hat{\tau}_F > \hat{\tau}_M$); and (iii) lower (higher) signal precision if $\hat{\eta}_F < \hat{\eta}_M$ ($\hat{\eta}_F > \hat{\eta}_M$). Belief partiality is* accurate *if* $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g) = (\mu_g, \tau_g, \eta_g)$ *for $g \in \{M, F\}$ and otherwise is* inaccurate.

Discrimination. Following the definition proposed in Bohren et al. (2022), we focus on *direct* discrimination, which is based on the difference in the hiring decision for a worker from group M versus F with the same realized signal. Let

$$D(s, \theta) \equiv v(s, M, \theta) - v(s, F, \theta) \tag{1}$$

denote this difference for an evaluator of type θ who observes realized signal s. *Direct discrimination* occurs at s when $D(s, \theta) \neq 0$; it occurs against group F if $D(s, \theta) > 0$ and against group M if $D(s, \theta) < 0$. There is *no*

[13] The microfoundation for this reduced form is as follows. If the evaluator hires the worker, she earns a payoff that is linear in productivity and also depends on group identity, $m_g a + b_g$, where $m_g > 0$ is a group-specific marginal value of productivity and $b_g \in \mathbb{R}$ is a group-specific taste parameter. If she does not hire the worker, she earns outside option $\underline{u}$. The evaluator maximizes her expected payoff. She hires the worker if and only if $\hat{E}[m_g a + b_g|s, g] > \underline{u}$, or $\hat{E}[a|s, g] > (\underline{u} - b_g)/m_g \equiv u_g$, where $\hat{E}$ denotes the expectation with respect to the evaluator's subjective beliefs. Therefore, u_g is a reduced form representation of the evaluator's payoff.

discrimination if $D(s, \theta) = 0$ for all $s \in \mathbb{R}$. When different sets of beliefs and preferences give rise to the same discriminatory behavior at all signals, we refer to this as *equivalent discrimination*.

Definition 3 (Equivalent Discrimination). *Two evaluators of types* θ *and* θ' *exhibit* equivalent discrimination *if* $D(s, \theta) = D(s, \theta')$ *for all* $s \in \mathbb{R}$.

While partiality refers to evaluators' preferences and beliefs, *discrimination* is a property of behavior and a consequence of these primitives. Identifying the source of discrimination refers to determining which form(s) of partiality generate the observed discrimination. Using this terminology, what the literature often refers to as taste-based discrimination corresponds to differential treatment stemming from preference partiality, whereas what is often referred to as statistical discrimination corresponds to differential treatment stemming from belief partiality. We define *inaccurate statistical discrimination* as differential treatment stemming from inaccurate belief partiality.

Discussion of model. We focus on binary evaluations for a population of workers with normally distributed productivity and signals. This simple setup allows us to illustrate how inaccurate beliefs impact discrimination in a tractable and succinct way. Our set-up easily extends to alternative forms of evaluations (e.g., selecting a continuous wage offer as in the experiment in section IV or a rating from a nonbinary discrete set) and to other distributions.

In terms of identifying the source of discrimination, we focus on sources of *direct* discrimination, where workers from different groups receive differential treatment conditional on generating the same information. A broader definition of discrimination, termed *total discrimination*, considers differential treatment conditional on underlying productivity *a* or some other qualification. This broader definition encompasses both direct and indirect, or *systemic*, discrimination (Bohren et al., 2022). Although it is likely that inaccurate beliefs present an identification challenge for identifying the source of systemic discrimination as well, a formal analysis of this is beyond the scope of the current paper.

B. Optimal Hiring Rule and Equivalent Discrimination

We next derive how the optimal hiring rule depends on preferences and beliefs. Given signal s and group identity g, the evaluator's posterior belief about productivity is normally distributed with mean $\hat{\mu}_g(s, \theta) \equiv (\hat{\tau}_g \hat{\mu}_g + \hat{\eta}_g s)/(\hat{\tau}_g + \hat{\eta}_g)$ and variance $1/(\hat{\tau}_g + \hat{\eta}_g)$. Since the posterior mean is monotonic with respect to s, the optimal hiring rule can be represented as a cutoff with respect to the signal.

Lemma 1 (Optimal Hiring Rule). *A type* θ *evaluator hires a worker from group* g *who generates signal* s*, that is,* $v(s, g, \theta) = 1$*, if and only the signal is weakly greater than*

$$\overline{s}(\theta, g) \equiv \left(\frac{\hat{\tau}_g + \hat{\eta}_g}{\hat{\eta}_g}\right) u_g - \frac{\hat{\tau}_g}{\hat{\eta}_g}\hat{\mu}_g. \tag{2}$$

The signal required to hire a worker is increasing in the evaluator's preference u_g and decreasing in the prior belief about average productivity $\hat{\mu}_g$. When $\hat{\mu}_g < u_g$, it is increasing in the concentration of productivity $\hat{\tau}_g$ and decreasing in the signal precision $\hat{\eta}_g$. In this case the evaluator seeks workers perceived to be in the top tail of the productivity distribution. Therefore, a higher signal realization is required to offset a concentrated productivity distribution. In contrast, the evaluator is willing to hire at lower signal realizations when the signal is more precise. These comparative statics reverse when $\hat{\mu}_g > u_g$ and the evaluator seeks to avoid workers perceived to be in the bottom tail.

We use lemma 1 to derive the sets of beliefs and preferences that give rise to equivalent discrimination. An evaluator of type θ discriminates against group F if she sets a higher hiring rule for group F, $\overline{s}(\theta, F) > \overline{s}(\theta, M)$. Types exhibit equivalent discrimination when they have preferences and beliefs that lead to the same pair of hiring rules.

Lemma 2 (Equivalent Discrimination). *For any constants* $(s_M, s_F) \in \mathbb{R}^2$ *with* $s_F > s_M$*, the set of types*

$$\{\theta \mid \overline{s}(\theta, M) = s_M \text{ and } \overline{s}(\theta, F) = s_F\} \tag{3}$$

exhibit equivalent discrimination against group F*. For each* $(s_M, s_F) \in \mathbb{R}^2$ *such that* $s_M = s_F$*, the set of types that satisfy equation (3) exhibit no discrimination.*

A given pattern of discrimination can stem from both preference and belief partiality against group F, belief partiality that is somewhat offset by more favorable preferences, or vice versa. For example, an evaluator with mild preference partiality and extreme belief partiality can exhibit equivalent discrimination to an evaluator with more extreme preference partiality and mild belief partiality.

We can represent the sets of types that exhibit equivalent discrimination as a pair of level sets parameterized by $(s_M, s_F) \in \mathbb{R}^2$, which we refer to as an *isodiscrimination curve*. Figure A1 in online appendix A illustrates an isodiscrimination curve in two dimensions. Fixing the other parameters, panel a plots the continuum of preference parameters and subjective average productivities for group F that lead to a given pair of hiring rules.

C. Identifying the Source of Discrimination

Researchers are often interested in identifying the *source* of discrimination, that is, the form of partiality that generates the observed discriminatory behavior.[14] Manski (2004) first observed that choice behavior could be consistent with

[14] A property is *identified* if it can be backed out from available data, or more formally, if there exists an injective relationship between the observed data and the property (Haavelmo, 1944).

multiple sets of preferences and subjective beliefs and, hence, lead to difficulties when using choice data to identify preferences and beliefs. We explore how the possibility of inaccurate beliefs impacts such identification in relation to discrimination.

To proceed, we assume that the researcher observes the group identity g, realized signal s, and hiring decision v for each worker, and that the data set includes a sufficiently rich set of workers such that the hiring rule for each group can be identified from these data—that is, the pair of signal cutoffs, which we denote by $(s_M, s_F) \in \mathbb{R}^2$.[15]

An identification challenge. It is well known that measuring discrimination (e.g., the extent to which workers who generate similar signals receive different evaluations) cannot be used to distinguish between preference-based partiality and accurate belief-based partiality about the productivity distribution (see, e.g., Bertrand & Mullainathan, 2004). The same insight extends to inaccurate beliefs about the distribution of productivity and accurate or inaccurate beliefs about the signal distribution. To formalize this insight, we show that for any pair of hiring rules, each form of partiality in isolation can generate the given pattern of discrimination. Therefore, observing the hiring rule for each group does not rule out either preference-based partiality or any of the forms of belief-based partiality.

Proposition 1 (Equivalent Sources). *For any pair of hiring rules $(s_M, s_F) \in \mathbb{R}^2$ with $s_F > s_M$, a continuum of types exhibit equivalent discrimination, including the following:*

1. *A type with* preference partiality *and belief neutrality, $u_F > u_M$ and $(\hat{\mu}_F, \hat{\tau}_F, \hat{\eta}_F) = (\hat{\mu}_M, \hat{\tau}_M, \hat{\eta}_M)$*
2. *A type with preference neutrality and* belief partiality due to lower expected productivity, *$\hat{\mu}_F < \hat{\mu}_M$ and $(u_F, \hat{\tau}_F, \hat{\eta}_F) = (u_M, \hat{\tau}_M, \hat{\eta}_M)$*
3. *A type with preference neutrality and* belief partiality due to higher concentration of productivity, *$\hat{\tau}_F > \hat{\tau}_M$ and $(u_F, \hat{\mu}_F, \hat{\eta}_F) = (u_M, \hat{\mu}_M, \hat{\eta}_M)$, and also such a type with belief partiality due to lower concentration of productivity*
4. *A type with preference neutrality and* belief partiality due to higher signal precision, *$\hat{\eta}_F > \hat{\eta}_M$ and $(u_F, \hat{\mu}_F, \hat{\tau}_F) = (u_M, \hat{\mu}_M, \hat{\tau}_M)$, and also such a type with belief partiality due to lower signal precision.*

From proposition 1, when all other parameters are equal, a higher preference parameter or a lower subjective average productivity for group F relative to M generates discrimination against group F. For the other parameters, all else equal, a higher subjective concentration of productivity or a lower subjective signal precision for group F generates discrimination against F for types with $\hat{\mu}_g < u_g$, and the opposite holds for types with $\hat{\mu}_g > u_g$. This stems from the comparative static in equation (2) for the variance parameter of interest: as discussed following lemma 1, how the parameter impacts the signal thresholds, and, therefore, the pattern of discrimination, depends on $\hat{\mu}_g$ and u_g. Figure A1 in online appendix A illustrates the evaluator types constructed in proposition 1.

We next discuss methods that seek to separate preference and belief-based sources.

Outcomes-based test. A common method used to identify the source of discrimination under the assumption of accurate beliefs is to compare evaluations to the outcome distribution for each group. In the current framework, the outcomes-based test corresponds to comparing hiring rules to the true productivity and signal distributions. Clearly this requires the researcher to identify the true productivity and signal distributions, in addition to the hiring rules:

> *Suppose the researcher can identify the hiring rules $(s_M, s_F) \in \mathbb{R}^2$ and the true productivity and signal distributions (μ_g, τ_g, η_g) for each group $g \in \{M, F\}$.*

Under the assumption of accurate beliefs, the outcomes-based test identifies the evaluator's preference parameters (u_F, u_M) and, therefore, the source(s) of discrimination. See lemma 3 in online appendix A for a formal statement of this insight and figure A2 in online appendix A for an illustration of the unique preference parameters that are consistent with a given set of hiring rules and true distributions.

We next explore how erroneously assuming accurate beliefs impacts the conclusions of an outcomes-based test. To do so, we first define how inaccurate beliefs impact discrimination. We say a type's inaccurate beliefs increase discrimination if the type sets a higher hiring rule for group F and a lower hiring rule for group M, relative to the type with accurate beliefs and the same preferences, and similarly for decreasing discrimination.

Definition 4 (Increasing and Decreasing Discrimination). *Suppose type θ^* has accurate beliefs and type θ has the same preferences as θ^* but inaccurate beliefs. Then θ's inaccurate beliefs* increase *discrimination against group F if, relative to θ^*, $\overline{s}(\theta, F) \geq \overline{s}(\theta^*, F)$ and $\overline{s}(\theta, M) \leq \overline{s}(\theta^*, M)$, with one inequality strict. The definition of decreasing discrimination is analogous with the inequalities reversed.*

The validity of the accurate beliefs assumption is crucial: when the researcher erroneously assumes accurate beliefs and uses an outcomes-based test, she mistakenly attributes discrimination stemming from inaccurate beliefs to a

[15] In practice, observing signals directly may not be possible. An alternative method is a *correspondence study*, which randomly assigns group identity and signals to a set of fictitious workers, then elicits hiring decisions (e.g., the classic resume study of Bertrand & Mullainathan, 2004). This ensures that workers from each group in the fictitious sample have the same distribution over signals, and therefore, any differences in hiring can be causally attributed to group identity. An audit study uses a similar randomized procedure to identify discrimination—experimental confederates with different group identities interact with evaluators while following the same script.

preference-based source. Depending on whether the inaccurate beliefs increase or decrease discrimination, the misidentified preference parameters will over- or underestimate the level of preference partiality.

Proposition 2 (Misclassification of Source). *Suppose a researcher identifies the hiring rules* (s_M, s_F) *and true distributions* (μ_g, τ_g, η_g) *for* $g \in \{M, F\}$. *If a researcher incorrectly assumes an evaluator has accurate beliefs and uses an outcomes-based test to identify the source of discrimination, then for a generic set of types and true distributions, the researcher misidentifies the evaluator's type. If inaccurate beliefs increase discrimination against group F, then the researcher overestimates the evaluator's preference partiality against group F, while if inaccurate beliefs decrease discrimination, then the researcher underestimates preference partiality.*

As illustrated in figure A2 in online appendix A, if the evaluator believes that the average productivity for group F is $\hat{\mu}_F = 3$ when in fact it is $\mu_F = 4.5$, then incorrectly assuming accurate beliefs will lead a researcher to conclude that $u_F = 6.3$ and $u_M = 6$ when, in actuality, the evaluator has preference neutrality, $u_F = u_M = 6$. Therefore, the researcher attributes discrimination stemming from this inaccurate belief to preference partiality.

While erroneously assuming accurate beliefs leads to a misclassification of source, the outcomes-based test no longer identifies the source when one relaxes this assumption. From equation (2), it is clear that when beliefs may be inaccurate, identifying the true belief parameters does not identify the evaluator's preferences. It can only be used to potentially rule out *accurate statistical discrimination*—that is, discrimination stemming from accurate beliefs and preference neutrality. The following result establishes when the observed pattern of discrimination is inconsistent with accurate statistical discrimination.

Proposition 3 (Rejecting Accurate Statistical Discrimination). *Suppose a researcher identifies the hiring rules* (s_M, s_F) *and true distributions* (μ_g, τ_g, η_g) *for* $g \in \{M, F\}$. *If*

$$\frac{\tau_M \mu_M + \eta_M s_M}{\tau_M + \eta_M} \neq \frac{\tau_F \mu_F + \eta_F s_F}{\tau_F + \eta_F}, \tag{4}$$

then the evaluator is not an accurate statistical discriminator.

Accurate statistical discrimination is of particular interest because it is often viewed as efficient from an informational perspective and has been used to justify social stereotyping and rationalize discriminatory behavior (Tilcsik, 2021).[16] When equation (4) holds, the researcher can reject this explanation and conclude that the observed discrimination either stems from animus toward a group or inaccurate beliefs about them. For example, in figure A2 (online appendix A), the accurate statistical discriminator type does not lie on the isodiscrimination curve and, therefore, is not consistent with the observed hiring rules.[17]

Of course, when the observed pattern of discrimination is consistent with accurate statistical discrimination, this does not identify accurate statistical discrimination: other preferences and beliefs can also generate the observed behavior. Even in this case, it is still important to identify the source of discrimination. Although a type with inaccurate beliefs may exhibit equivalent behavior to an accurate statistical discriminator for the current hiring decision, these inaccurate beliefs may affect the worker in future evaluations in ways that differ from accurate statistical discrimination. For example, consider an evaluator who overestimates the difference in average productivity between groups, but this is offset by preferences that favor the disadvantaged group for entry-level positions. Then if the evaluator feels compelled to favor only the disadvantaged group for entry-level hiring, these inaccurate beliefs will lead to persistently lower rates of promotion and advancement for the disadvantaged group.

Given the difficulty of using an outcomes-based test to identify the source of discrimination, we next explore two alternative methods.

Eliciting beliefs. If it is possible to collect data on the evaluator's subjective beliefs, then comparing hiring decisions to these beliefs can identify the source of discrimination.[18]

> *Suppose the researcher can identify the hiring rules* $(s_M, s_F) \in \mathbb{R}^2$ *and the subjective productivity and signal distributions* $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$ *for each group* $g \in \{M, F\}$.

One way to identify subjective beliefs would be to directly elicit them from evaluators.

Similar to the outcomes-based test under the assumption of accurate beliefs, this method identifies the evaluator's preferences and, therefore, the source of discrimination.

Proposition 4 (Identifying Preferences from Subjective Beliefs). *Suppose a researcher identifies the hiring rules*

[16] The main argument is that, if an evaluator is applying differential treatment to groups when underlying differences do exist, then this evaluator is simply using information in an optimal way and engaging in profit-maximizing behavior.

[17] Prior work has highlighted additional identification challenges for outcomes-based tests, including the problems of inframarginality (Ayres, 2002; Simoiu et al., 2017) and relying on administrative data that may condition on a posttreatment outcome (Knox et al., 2020). In contemporaneous theoretical work, Hull (2021) shows that outcome-based tests that use IV and MTE methods (e.g., Arnold et al., 2018; Grau & Vergara, 2021) can distinguish between accurate statistical discrimination and other sources. These marginal outcome tests do not require the researcher to observe the decision maker's signal and do not suffer from inframarginality or selection problems by definition. However, they still require additional assumptions in order to separate taste-based discrimination from inaccurate beliefs (see, e.g., the structural model developed in Arnold et al., 2018).

[18] Manski (2004) first proposed combining data on expectations with choice data to identify preferences without assuming rational expectations.

(s_M, s_F) and subjective beliefs $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$ for $g \in \{M, F\}$. This identifies the preference parameters (u_M, u_F), and therefore, the evaluator's type.

Importantly, observing subjective beliefs does not identify whether they are accurate—additional data, such as outcomes, are necessary to determine this.[19]

In practice, this method will be difficult in many settings—both due to the complexity and reliability of methods for eliciting beliefs about higher moments and due to the feasibility of collecting such information (e.g., it may not be possible to collect beliefs in certain settings such as on an online platform). The next method provides an alternative, simpler way to partially identify the source of discrimination.

Manipulating information. Suppose it is possible to manipulate the amount of information presented to evaluators. For example, one could compare discrimination in a treatment in which only one customer review is revealed to a treatment in which five customer reviews are revealed. In the current framework, we model this as varying the number of signal draws x that the evaluator observes for a worker.

> *Suppose the researcher can identify the hiring rules $(s_M^i, s_F^i) \in \mathbb{R}^2$ for multiple informational treatments i with x_i signal draws.*

If an evaluator believes that one draw of the signal has precision $\hat{\eta}_g$, then she believes that observing $x \geq 1$ conditionally independent draws of this signal has precision $x\hat{\eta}_g$. The characterization of the optimal hiring rules and set of types that generate equivalent discrimination following x draws is identical to the case of one draw, substituting $x\hat{\eta}_g$ for $\hat{\eta}_g$.

We next establish that manipulating the number of signal draws can separate preference partiality from belief partiality, but it cannot separate the different forms of belief partiality. Proposition 5 establishes that for any two informational treatments, there is a unique pair of preference parameters that yield equivalent discrimination. However, there are a continuum of types with the same pair of preference parameters and distinct belief parameters that exhibit equivalent discrimination across both informational treatments. Moreover, this set of types also exhibits equivalent discrimination across all informational treatments. Therefore, identifying the hiring rules for at least two informational treatments identifies the evaluator's preferences u_g but does not identify beliefs $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$.

[19] An alternative methodology involves eliciting beliefs about group performance and comparing evaluations when the same groups are identified either using labels subject to stereotypes (e.g., gender) or not (e.g., birth month) (Coffman et al., 2021). The performance distributions are the same regardless of the label, and so any differences in evaluations between the two treatments can be assigned to tastes rather than beliefs. As the authors note, creating equivalent evaluation settings for both types of labels requires that the methodology be implemented in a controlled laboratory environment.

Proposition 5 (Identifying Preferences from Manipulating Information). *Suppose a researcher identifies the hiring rules (s_M, s_F) and (s'_M, s'_F) for two informational treatments corresponding to an evaluator observing either $x \geq 1$ or $x' \neq x$ signal draws for each worker. This identifies the preference parameters,*

$$u_g = \frac{xs_g - x's'_g}{x - x'}, \tag{5}$$

for $g \in \{M, F\}$, but does not identify beliefs $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$. Additional informational treatments provide no further identification of beliefs.

Figure A3 in online appendix A illustrates this result. Only types with preference parameter $u_F = 6.2$ exhibit the observed discrimination for both informational treatments; the types with other preference parameters that lie on the same isodiscrimination curve for one signal draw do not lie on the same isodiscrimination curve for two draws.

A crucial requirement is that multiple signals are drawn from the same distribution, so that the evaluator has the same belief about the signal distribution for each draw. For example, suppose each signal draw is a past review, such as on Airbnb, and the worker is rated multiple times by evaluators from the same population. Then a researcher could manipulate information by varying the number of ratings that are visible to evaluators. Another example is settings where a worker receives multiple simultaneous signals, such as reviews for a grant proposal or recommendations for employment from colleagues with similar qualifications. Then a researcher could manipulate information by varying the number of signals that are shown to the evaluating committee. In contrast, varying observation of signals from different domains (e.g., comparing discrimination when education is observed to discrimination when education and SAT score are observed) does not identify preferences because the evaluator may have a different subjective signal distribution for the signal from each domain.

Taken together, the proposed belief elicitation and information manipulation methods can be used to separate preference and belief partiality. If it is possible to elicit an evaluator's beliefs for all parameters of the relevant distributions, then it is possible to fully identify the evaluator's type. If not, then the information manipulation method provides an alternative, simpler way to identify preferences and "aggregate" belief partiality—although it comes at the cost of not being able to separate the different ways that beliefs may be inaccurate.

IV. Identifying the Source of Discrimination in a Hiring Experiment

We next employ a stylized experimental setting to demonstrate how assuming accurate beliefs can lead to erroneous conclusions about the source of discrimination and to illustrate how the belief elicitation method outlined above can

solve it. The experiment allows us to observe the actual distribution of productivity by group and, therefore, to perform the accounting exercise employed in outcomes-based tests and also to elicit beliefs about relevant characteristics. We show that average beliefs about productivity are incorrect, thereby violating the accurate beliefs assumption, and that ignoring these inaccurate beliefs leads to a false identification of the source of discrimination. We also demonstrate how providing information about the true group-specific average productivities can be used both to separate inaccurate beliefs from underlying animus and to correct these inaccurate beliefs. Participants adjust their behavior significantly in the direction of the information, suggesting that at least some of the observed discrimination is driven by inaccurate beliefs rather than animus.

A. *Experimental Design*

This subsection provides a summary of the preregistered experimental design.[20] Two minor differences exist between the preregistration plan and the actual study. First, we preregistered that we would recruit 400 U.S. employers in the hiring task survey, but then decided to target an additional 200 Indian employers so that we could examine in-group/outgroup evaluations. Second, we did not preregister sample restrictions because of completing the task too quickly or slowly. We dropped twelve subjects in the work task survey and five in the hiring task survey because of these restrictions. The full surveys are in the Replication Data. We recruited two samples of subjects on Amazon Mechanical Turk ("MTurk"; participants) to complete either a work task (Survey 1) or a hiring task (Survey 2), which we next describe in detail.

Survey 1 (work task). We recruited 589 subjects from MTurk on February 23, 2018, for the first survey.[21] The survey was posted with the title "Math Questions and Demographics" and the description "A 20-minute task of answering math questions." We paid $2 (i.e., a projected $6/hour wage) and recruited a subject pool of 392 from the United States and 197 from India, all of whom had completed at least 500 prior tasks and had an 80% or higher approval rate for these tasks.[22] After starting the survey, subjects were informed that they would first answer demographic questions and then answer fifty multiple choice math questions. They were told that their performance would not affect their payment and were asked not to use a calculator or any outside help, but just to do their best. This was followed by seven questions that provided the information used for their profiles in the second survey: favorite color, favorite movie, coffee versus tea preference, age, gender, favorite subject in high school, and favorite sport. The math test included a mix of arithmetic (e.g., "$5 * 6 * 7 =$?"), algebra (e.g., "If $(y + 9) * (y^2 - 121) = 0$, then which of the following cannot be y?"), and more conceptual questions (e.g., "Which of the following is not a prime number?"). Finally, subjects were thanked for their participation and informed that they may receive a small bonus based on a different experiment, for reasons unrelated to their performance on the task. We describe the basis for such bonuses in the description of Survey 2.

The purpose of the first survey was to create a bank of "workers" who could be hired by the "employers" in the second survey. This novel design has several advantages over the existing paradigms for studying discrimination in the field. First, in contrast to correspondence studies, we did not employ deception at any point—all profiles shown to employers corresponded to actual workers who would in fact be paid as described in the following paragraph. However, similar to a correspondence study, we were able to control the information seen by an employer about a prospective worker by constructing worker profiles that included information that is ostensibly relevant for animus and/or beliefs about productivity (e.g., age, gender, and nationality), as well as other nontarget information (e.g., tea preference). The nontarget information ensures that the relevant demographics are not the only salient information provided to the employer (this mimics the additional—ostensibly less decision-relevant—information contained on a CV). Finally, instead of the coarse measures of discrimination used in many other studies (e.g., callback or stop rates), we elicit relatively continuous and precise measures of productivity and discrimination that are tightly linked. The downside to such a design is that the target characteristics and productivity may be correlated with the nontarget information and thus may inform their decisions.[23]

Survey 2 (employer tasks). We recruited 577 different MTurk subjects on February 26, 2018. We used the same hiring criteria as the first survey (392 from United States, 185

[20] The experiment was preregistered on AsPredicted (https://aspredicted.org; #8678).

[21] We received 604 responses in total, but dropped twelve responses that corresponded to the top 1% (<227 seconds) and bottom 1% (>3,274 seconds) in terms of survey duration. Of the remaining 592 responses, we dropped three whose Qualtrics survey responses could not be matched to their MTurk records.

[22] This geographic restriction is based on the addresses MTurkers used to register on Amazon. The survey was posted as two tasks on MTurk, with one only eligible for Indian workers and the other for U.S. workers.

[23] Although we chose items that we intended to be less informative for the task at hand, favorite high school subject was in fact both relevant for performance (those who mentioned math performed roughly 3.2 points higher on the math test) and anticipated in wage offers (they were offered wages 5 cents higher on average). The other items were less relevant for both scores and wage offers. In online appendix B, we show that nationality and gender remain significant when controlling for binary versions of these profile attributes. Although our design has some advantages (e.g., no deception), by not randomizing profile attributes we are capturing a broader "bundle of sticks" (Sen & Wasow, 2016) than in a standard correspondence or audit study. Incentivized resume rating (Kessler et al., 2019) is an alternative design that has no deception while maintaining randomized attributes across groups.

from India, $\geq$ 80% approval rate).[24] The survey was posted with the title "20-Minute Survey about Decision-Making" and the description "20-Minute Survey about Decision-Making." We paid $2 (i.e., a projected $6/hour wage). Subjects were first asked to report their gender, age, and education level. Subjects were then presented with the first hiring task portion of the survey.

First hiring task. We informed subjects that we had previously paid other subjects ("workers") to answer fifty math questions, showed them five examples of the math questions, and told them that, on average, participants answered 36.95 out of fifty questions correctly. They were then told that they would act as an employer and hire one of these workers by stating a wage (paid as a bonus to the worker). In return, they would receive a payment based on how many questions their hired worker answered correctly. This was followed by a more detailed description of the assignment. Each "employer" would view twenty profiles of potential workers and state the highest wage (between 0 cents to 50 cents) they were willing to pay to each worker. The employer would be paid 1 cent for each question answered correctly by the hired worker. We next described the mechanism (Becker-Degroot-Marschak) used to assign payment. We would randomly select a profile from the twenty potential workers. We would then draw a random number from 0 to 50. If the wage the employer stated for the worker was equal or greater than that number, then the worker would receive the random number as a bonus and the employer would receive a "profit" equal to the worker's performance minus the random number. If instead the employer stated a wage for the worker that was lower than the random number, then neither the worker nor the employer would receive a payment.

To ensure comprehension, we showed subjects an example profile (see figure B1 in online appendix B) and stated wage. We gave examples of actual performance and randomly generated numbers that would produce positive profit, negative profit, and no hiring. Having highlighted the possibility of negative profit, we then noted that all employers would automatically be paid a $0.50 bonus in addition to any money made through the hiring task, so that no employers would owe money. Finally, we ran a comprehension check with the same example profile, a specific wage (43), a random number (18), and an actual performance (10). We required the employer to correctly state how many cents they would have to pay the worker (18) and how many cents the employer would be paid before subtracting off the amount they would pay the worker (10).[25] Finally, employers were presented with a second wage (15) and answered the same questions. They were then presented with twenty profiles, each randomly selected with replacement from the bank of 589 profiles produced by the first survey.

Belief elicitation task. Next, subjects were randomly assigned to one of two different conditions: an incentivized or unincentivized belief elicitation. Across both conditions, subjects were reminded that the full sample answered 36.95 out of fifty questions correctly. They were then asked to answer six questions of the form "On average, how many math questions out of 50 do you think X answered correctly?" where X corresponded to the groups "women," "men," "people from the United States," "people from India," "people below or at the age of 33," and "people above the age of 33."[26] In the incentivized condition, prior to the six questions, subjects were told that they could earn a significant bonus for an accurate prediction. One of the six questions would be randomly selected, and they would be paid $5 minus their deviation from the question (bounded below by $0). For example, if they answered 40 and the true average was 37, they would receive a $2 bonus. Finally, they were asked to "please answer the questions as carefully as possible so that you can potentially win a large bonus."

Information intervention and second hiring task. After completing the belief elicitation, subjects were shown the correct answer for all six groups: women (35.28), men (38.32), people from the United States (37.14), people from India (36.58), people below or at the age of 33 (37.10), and people above the age of 33 (36.79). As discussed in section I, providing accurate information about group-level statistics is one potential method for differentiating between inaccurate beliefs and "animus-driven" beliefs. Although the former should shift in the direction of the information, the latter are unlikely to be moved because the errors are due to noninformational factors. Following this information, we stated, "Now that you have learned those facts, we would like you to work on 10 more profiles." We noted that, as in the first hiring task, we would randomly select one profile and a number, and pay bonus and wages accordingly (with an additional $0.50 automatic bonus to ensure no negative payments). After employers reviewed the ten additional worker profiles, we thanked them for their participation, noted that we would calculate bonuses and pay them within a week, and allowed subjects the option to leave comments.

Summary statistics. Table B1 in the online appendix provides summary statistics for the full sample of subjects that completed surveys 1 and 2 (column 1), as well as these statistics for each of the six demographic groups used in the second survey. On average, the work task (survey 1) took subjects 19 minutes to complete, and the hiring task took 23

[24] We recruited 587 subjects in total, but dropped seven whose surveys were completed in under 300 seconds and three whose stimuli (the profiles they evaluated) could not be matched to the first survey.

[25] Entering an incorrect an answer would generate a pop-up with "Wrong Answer" and restrict the individual from moving to the next page.

[26] We elicited only beliefs about the first moment of the performance distribution. Although participants may also have inaccurate beliefs about other statistics, demonstrating a difference in subjective versus actual means is sufficient to falsify the assumption that beliefs are correct, which was the primary goal of the illustrative experiment.

TABLE 1.—WAGES AND "PRODUCTIVITIES," BY EMPLOYEE CHARACTERISTICS (HIRING TASK 1)

	Group 1 (1)	Group 2 (2)	Diff. (3)	*p* value (4)	No. obs. G1 (5)	No. obs. G2 (6)
Panel A: Employers' wage WTP, by employee characteristics						
Gender (1 = Male, 2 = Female)	31.90	30.85	1.05	0.01	6,306	5,234
	(12.07)	(12.23)				
Country (1 = United States, 2 = India)	30.71	32.85	−2.14	0.00	7,700	3,840
	(12.20)	(11.95)				
Age (1 = Under 33, 2 = Over 33)	31.67	31.14	0.54	0.17	6,139	5,401
	(12.00)	(12.33)				
Panel B: Employee productivity, by employee characteristics						
Gender (1 = Male, 2 = Female)	38.30	34.98	3.32	0.00	6,306	5,234
	(8.55)	(8.73)				
Country (1 = United States, 2 = India)	37.01	36.36	0.65	0.41	7,700	3,840
	(8.93)	(8.49)				
Age (1 = Under 33, 2 = Over 33)	36.96	36.60	0.37	0.63	6,139	5,401
	(8.62)	(8.98)				

Standard deviations in parentheses. One observation per worker-employer combination. Column 4 shows the *p* value from a regression of the outcome on a dummy variable for group membership, with standard errors two-way clustered by employer and worker.

minutes. There is variation in this timing across groups. Subjects from the United States took an average of 19 minutes to complete the hiring task, whereas subjects from India took 31.60 minutes, a difference also reflected in their median times (15.8 vs. 25.6). Another large difference between the U.S. and India samples is the average age of participants; the average Indian subject in the work task is approximately 8 years younger than the average U.S. subject. This gap shrinks to 4 years for the hiring task. The Indian sample also skews more male than the U.S. sample (68.5% vs. 48.2% and 76.8% vs. 51.4% for survey 1 and 2, respectively) and is more likely to have a college education or above (90.3% vs. 56% in survey 2; the question was not asked in survey 1). Although we primarily focus on simple comparisons between each demographic group, these observed differences motivate our use of multivariate regressions in robustness tests presented in online appendix B.

Connection to theoretical framework from section III. In the experiment, productivity a corresponds to the worker's performance on the math test. The experimental design simplifies the theoretical framework by eliminating the signal of productivity—evaluators observe group identity but no performance signal. It also has a richer action space: subjects choose a wage between 0 and 50 cents, whereas the theoretical framework is based on a binary hiring decision. Given the induced payoffs in the experiment, the optimal action depends on the subjective average productivity but not the subjective variance of productivity. The analysis from section III easily extends to this alternative action space and decision rule. Given that there is only scope for belief-based partiality due to differences in subjective average productivity, in analyzing the experimental data, we focus on comparing average wages to measure discrimination, and we elicit beliefs about average productivity to determine whether beliefs are inaccurate.

B. *Experimental Results*

A necessary prerequisite to study the source of discrimination is to find a context and a population in which discrimination occurs. Ex ante, it was not obvious that our stylized hiring experiment would satisfy this requirement. The employers knew that they were being observed as part of a research study and the relevant group information was represented abstractly (e.g., written text) rather than viscerally (e.g., a picture). All of these factors may attenuate the influence of animus.[27]

Despite these attenuating factors, we did find evidence of discrimination with respect to two out of three group identities: gender and nationality. Panel A of table 1 presents the differences in average wages paid by employers to worker profiles from each group. With respect to gender, male profiles were paid on average 31.90 cents, whereas female profiles were paid 30.85 cents, a significant 3.4% difference ($p < 0.01$). With respect to nationality, profiles from India were favored, earning an average of 32.85 cents, whereas profiles from the United States earned 30.71 cents, a significant 7.0% difference ($p < 0.01$). Finally, we found no statistically significant evidence of age discrimination: subjects at or below age 33 were paid an average of 31.67 cents, and those above age 33 were paid 31.14 cents, a 1.7% difference ($p = 0.17$). Table B2 in the online appendix demonstrates that these results are relatively similar in a multiple regression framework with employer fixed effects, though adding additional some profile characteristics does attenuate differences (notably, "favorite high school subject," which is both predictive of productivity and wages and correlated with gender and nationality).

[27]For example, Bar and Zussman (2019) argue that a lack of interaction may attenuate the extent of taste-based discrimination in driving test examinations.

To examine the possibility of in-group bias, we run similar regressions controlling for the employer belonging to the group of interest (e.g., female) and the interaction of the two indicators to measure in-group bias (see table B3 in the online appendix). We find that the interaction is insignificant for gender and marginally significant for nationality, although in the direction of favoring the out-group. For age, we find a significant interaction effect. This suggests that the null effect in table B2 masks in-group bias by both older and younger employers. Antonovics and Knight (2009) use a similar set of regressions to test for taste-based discrimination. This specification is motivated by the assumption that animus varies between groups (i.e., there is less animus toward one's in-group than out-group), but that beliefs are similar across groups (because they are taking a "standard model of statistical discrimination" as the benchmark and note that "these beliefs must be correct in equilibrium"). Table B4 in the online appendix tests this assumption in our experimental environment. We find that beliefs about the gender performance gap are identical among both female and male employers. However, for nationality, we find significant differences. Americans hold beliefs that favor the out-group and Indians hold beliefs favoring the in-group—while both groups believe Indians will outperform Americans, the latter group predicts a larger gap.

Having demonstrated moderate levels of discrimination in hiring, we now examine the "ground truth" in actual productivity differences between groups. The typical outcomes-based test of statistical discrimination requires mapping disparities between groups in the evaluators' relevant decision (e.g., the wages offered to employees) to disparities in an outcome in the evaluators' objective function (e.g., the employees' productivity).[28] In our context, this requires mapping disparities in the employers' stated wages to disparities in group-specific productivity differences, that is, the number of questions answered correctly. The commonly used outcome method compares disparities in wages to disparities in performance to measure the relative role of (accurate) statistical versus taste-based discrimination (in the context of our framework, accurate belief-based versus preference-based partiality). For simplicity, we will refer to both disparities as measured in "points."

Panel B of table 1 shows the average number of correct answers by each sub-group (see figure B2 in the online appendix for probability density functions). As shown in panel A of table 1, the gap in average wages for men and women was lower than the gap in average performance (1.05 points vs. 3.32 points).[29] Therefore, if we used the standard outcome method to separate statistical and taste-based discrimination, we would conclude that the entire 1.05 point disparity in wages is due to (accurate) statistical discrimination—the remaining 2.27 point difference in performance would be attributed to taste-based discrimination against men. Turning to nationality-based discrimination, there was a wage gap of −2.14 points in favor of Indians, compared to a performance gap of 0.65 points in favor of Americans. Under the standard approach, we would conclude that the −2.14 point disparity in wages, when compared to the +0.65 point difference in performance, suggests taste-based discrimination against Americans.[30]

We now proceed to examine whether inaccurate beliefs can explain the disparities in compensation. As an initial check to see whether employers' decisions were guided by the elicited beliefs, we correlate wages with their beliefs about group-specific productivities. We find positive correlations for all six groups of workers (Female: 0.12, Male: 0.12, India: 0.15, United States: 0.12, Over 33: 0.12, Under 33: 0.10). Given that we elicited beliefs after the hiring task, it is possible that part of these correlations are due to rationalization (e.g., an individual first discriminates against women when setting wages, then chooses beliefs to justify this decision) or audience effects (e.g., an individual falsely reports beliefs that justify the discriminatory decision to the experimenter). To test for this, we provided half of the employees with large incentives for belief accuracy. In table B5 in the online appendix, we show that beliefs are nearly identical across both incentive conditions, with none of the six comparisons being significantly different from one another. Together these findings suggest that the employers' group-specific performance predictions provide meaningful information about their true beliefs.

In table 2 we present employer beliefs about the group-specific average performance, which can be compared directly to the actual group-specific performance reported in panel B of table 1. Predictions about performance are lower than actual performance for all six groups. This overall underestimation is consistent with risk aversion (recall that employers face the potential of a negative payment, taken from their $0.50 bonus, if they overestimate performance). Consistent with this, gaps in beliefs about performance are larger than gaps in wage payments. Using employers' actual beliefs to identify the source of discrimination leads to substantially

[28] Translating the two measures may require strong modeling assumptions (e.g., whether there is heterogeneity in the search costs faced by evaluators). For discussions of these assumptions in the context of the hit-rate tests, see Dharmapala and Ross (2004), Anwar and Fang (2006), and Antonovics and Knight (2009).

[29] We calculate productivity differences using the full sample of profiles observed in hiring task 1. This is a weighted sample of the original population of 577 workers (because each of the 589 employers saw independent random samples of twenty of the 577 workers). Because of the random variation in the profiles observed, the group-level averages slightly differ from those found in table B1. For example, the male-female performance gap is 3.04 points in table B1 and 3.32 points in this weighted sample. Note that the averages in table B1 are the basis for the informational intervention.

[30] Although we document significant discrimination by gender (i.e., men are paid more than women), the outcome method reveals that the performance gap exceeds the pay gap. This leads to the conclusion that there is taste-based discrimination *against* men. Although the literature often equates taste-based discrimination with animus or prejudice, this link may be inappropriate when discrimination manifests as an equalizing action. For example, people may be equalizing wages between two groups despite differences in productivity due to fairness concerns. We discuss the implications of this distinction further in the conclusion.

TABLE 2.—BELIEFS ABOUT PRODUCTIVITY BY EMPLOYEE CHARACTERISTICS

	Group 1 (1)	Group 2 (2)	Diff. (3)	p value (4)
Gender (1 = Male, 2 = Female)	34.04	32.14	1.89	0.00
	(8.26)	(8.41)		
Country (1 = United States, 2 = India)	32.08	34.80	−2.72	0.00
	(8.56)	(9.44)		
Age (1 = Under 33, 2 = Over 33)	33.41	31.57	1.84	0.00
	(8.97)	(9.00)		

Standard deviations in parentheses. One observation per employer combination. Column 4 shows the p value from one-sample t-tests for the equality of columns 1 and 2. Number of observations = 577.

TABLE 3.—EFFECT OF INFORMATION: DIFFERENCE IN DIFFERENCES BY HIRING TASK

	(1)	(2)	(3)	(4)	(5)
Post-Info	1.53***	1.60***	1.06***	1.97***	2.33***
	(0.31)	(0.27)	(0.31)	(0.39)	(0.34)
Female	−1.05***			−0.66*	−0.80**
	(0.38)			(0.37)	(0.33)
Female	−0.64*			−0.89**	−1.01***
× post-info	(0.38)			(0.38)	(0.29)
Indian		2.14***		2.01***	2.02***
		(0.41)		(0.43)	(0.38)
Indian		−1.07**		−1.20***	−1.65***
× post-info		(0.43)		(0.44)	(0.33)
Over 33			−0.54	0.06	0.29
			(0.39)	(0.39)	(0.35)
Over 33			0.41	0.12	−0.21
× post-info			(0.42)	(0.42)	(0.31)
N	17,310	17,310	17,310	17,310	17,310
R^2	0.01	0.01	0.00	0.01	0.48
DepVarMean	31.90	30.71	31.67	30.71	30.71
Employer FE?	No	No	No	No	Yes

Standard errors in parentheses, two-way clustered by employer and worker. "DepVarMean" is the mean of the dependent variable (wage WTP) in the omitted group (e.g., Male Workers in Hiring Task 1 for column 1). "Post-Info" is an indicator for whether a profile came in the second hiring task (i.e., profiles 21–30 of the thirty total profiles evaluated). The observed performance (trivia score) averages for the sample of profiles observed in Hiring Task 2 are 38.13 (Male), 35.13 (Female), 36.95 (U.S.), 36.53 (India), 36.84 (Under 33), 36.77 (Over 33), 36.81 (Prefer Coffee), 36.79 (Prefer Tea). $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

different conclusions than the outcomes-based method outlined above. Looking at nationality, the wage gap is −2.14 points and the performance gap is +0.65 points; the gap in beliefs is −2.72 points. Thus, the *entire* wage gap can be explained by inaccurate beliefs. In contrast to the outcome method which infers taste-based discrimination in favor of Indian workers, the remaining 0.58 point difference between the belief and wage gaps suggests prejudice *against* them. Looking at gender, the wage gap is 1.05 points, the performance gap is 3.32 points, and the belief gap is 1.89 points. The majority of the wage gap can be explained by inaccurate beliefs: the residual attributed to preference-based sources shrinks from 2.27 to 0.84 points. Finally, despite the minimal gap in wages and performance based on age, employers believed that young workers will significantly outperform older ones. This suggests some preference-based partiality against younger workers. Together these results highlight that a failure to account for inaccurate statistical discrimination may lead to the wrong conclusion on the source of treatment disparities.[31]

To identify whether the observed disparate treatment was driven by inaccurate statistical discrimination or animus-driven beliefs, we examined how behavior would respond to an informational intervention. Table 3 compares the differences between the two hiring rounds ("Post-Info"), the differences between wages assigned to profiles of each demographic group (e.g., "Female"), and the difference-in-differences (e.g., "Female X Post-Info"). The coefficients on "Post-Info" suggests substantial belief updating across all demographic groups, partially correcting the large level differences in the first hiring task between wages and actual group-specific productivity (a gap of roughly 5 points on average). The effect of the informational intervention on hiring decisions suggests that the majority of initial discrimination was driven by inaccurate beliefs rather than accurate statistical or preference-based sources.[32]

V. Conclusion

The study of discrimination and its motives has a rich history in economics. Separating out statistical and taste-based drivers of discrimination is important, but as our literature survey illustrates, most of the empirical literature thus far has relied on the assumption of accurate beliefs. We have many reasons to suspect that beliefs may not always be accurate. Our theoretical framework and stylized experiment outline the identification problem inherent in distinguishing between different drivers of discrimination when allowing for inaccurate beliefs, as well as the pitfalls of ignoring this possibility. We also illustrate potential methodologies for identification.

The results of our experimental information intervention have policy implications for reducing discrimination. However, some important caveats must be kept in mind when considering how this type of intervention would be implemented outside of the lab. First, such an intervention is likely feasible only in contexts where the underlying target outcome (e.g., productivity) is reliably measured and reflects the appropriate counterfactual outcome for all groups. To the first point, the accuracy of this measurement may differ by group (e.g., police officers have been shown to be more likely to discount the recorded speed of a white driver

[31] In table B6 in the online appendix, we show that the differences in beliefs are quite similar after trimming the top and bottom 5% of the distributions of belief differences by each group. Consistent with figure B3 in the online appendix, differences in beliefs about group productivities are driven by a large mass of employers with biased beliefs rather than a few employers with extreme beliefs.

[32] Several caveats should be noted when interpreting these results. Beliefs were not measured a second time. Additionally, experimenter demand may have played a role, though recent work suggests that this factor is likely small (De Quidt et al., 2018). Finally, the change in wages could reflect an experience effect between assigning wages in the first and second hiring task. To investigate this channel, we perform a test comparing the average wages assigned in the first ten profiles and the second ten profiles during the initial task. We do not find evidence for an experience effect (36.86 vs. 36.72; $p = 0.39$). Although we cannot fully rule out all these possible confounds, we view the information intervention as a proof of concept for the type of methodology that can be used as both an intervention for correcting beliefs and identifying belief-based discrimination from preference-based motives (e.g., animus-driven beliefs).

than a minority driver [Goncalves & Mello, 2019). To the latter point, in some settings discrimination at (often unobserved) intermediate stages renders final productivity measures unreliable because of behavioral responses. For example, minority pitchers correctly anticipate discrimination by umpires and modify their behavior, resulting in a downward bias for performance measures (Parsons et al., 2011). Studies have also documented that bias at intermediate stages can skew final productivity measures among grocery store workers (Glover et al., 2017) and academic economists (Hengel, 2019). It is also important to consider the underlying psychology of how people will respond to the information. Selection decisions such as hiring are rarely unidimensional. Drawing attention to a (smaller than expected) productivity gap could correct beliefs, while nonetheless increasing discrimination if it increases the salience of the gap. These concerns highlight the need for research on informational interventions in field contexts.

Throughout the paper, we document discrimination in wages by gender (i.e., men paid more than women). Carrying out the standard outcomes-based method reveals that the gap in performance exceeds the gap in pay. This leads to the conclusion that there is preference-based partiality against the group that received higher wages—male workers. Although taste-based discrimination is often used as a synonym for animus or prejudice against a group, this link seems misplaced when discrimination manifests as an equalizing action (e.g., equalizing wages). For example, people may treat groups similarly regardless of actual or believed differences due to fairness concerns. Additionally, often one finds an equity-efficiency trade-off to discrimination, such that even in the absence of legal or social sanctions, an employer may wish to equalize wages across groups (for a theoretical discussion of these trade-offs in the context of racial profiling, see Durlauf, 2005). Such a concern may be especially pronounced for wages, where even abstracting away from demographic groups, evidence suggests that fairness norms may contribute to observed wage compression (e.g., Breza et al., 2018).

Just as determining the nature of belief-based discrimination has implications for policy, the same may be true for preference-based discrimination. For example, if the basis for preference-based discrimination is animus or prejudice, then a policy that increases contact between groups may reduce disparities (Dobbie & Fryer, 2015; Paluck et al., 2018; Rao, 2019). By contrast, if the behavior is instead sanction- or value-oriented, then such interventions will likely have little impact. Although it is difficult to imagine a simple elicitation that would allow for a parsimonious quantitative decomposition of "tastes," survey measures may be able to make some headway in this endeavor. Such a decomposition is outside of the scope of this paper, but future work along these lines would enrich our understanding of discrimination and help in the development of tools used to identify it and design policy.

Finally, our findings speak to the continued need for innovative methods to model and measure belief-based discrimination (e.g., Bartoš et al., 2016; Bordalo et al., 2016), as such methods may be able to help identify inaccurate beliefs. Two broad causes may lead to inaccurate beliefs that drive discrimination. First, research in psychology and economics has shown that heuristics and biases may generate beliefs that are systematically incorrect, leading to inaccurate stereotypes about certain groups.[33] Second, inaccurate beliefs may arise because of a lack of information—the relevant information necessary to form correct beliefs may not be available to a decision maker. For example, an employer may have an unbiased prior belief about the productivity distributions of two groups but lack information about how selection into the job application process differs across groups, leading to inaccurate beliefs about productivity differences in the realized applicant pool. Failing to account for selection effects can also be a form of bias, as in Hübert and Little (2020) in the case of discrimination in policing. Learning will eventually mitigate inaccurate beliefs in some settings. But in other situations, there will be little or no feedback on the decisions being made, leading to learning traps in which inaccurate beliefs persist in the long run.[34] Further, learning may not lead to correct long-run beliefs if information is filtered through a misspecified model of the world.[35]

As research begins to identify situations where inaccurate beliefs are a driving factor for discrimination, future work will hopefully also begin to develop policy interventions that are able to effectively correct beliefs and thereby reduce discrimination.

[33] See Schneider et al. (1979), Judd and Park (1993), Hilton and Hippel (1996), and Fiske (1998) for review. Bordalo et al. (2016) model inaccurate stereotype formation based on the representativeness heuristic where evaluators overweight the prevalence of characteristics that differ most between groups. Biased beliefs can also arise in a dynamic learning setting when individuals use updating rules that depend on group identity (Albrecht et al., 2013), have selective attention (Schwartzstein, 2014), or have incorrect models of how others evaluate workers (Bohren et al., 2019).

[34] For example, if employers face a trade-off between learning about the productivity distribution of groups or maximizing cost effectiveness in hiring, this can prevent full learning even though the employers are not inherently biased (Lepage, 2020).

[35] For example, confirmation bias (Rabin & Schrag, 1999), overreaction (Epstein et al., 2010), and misattribution of reference dependence (Bushong & Gagnon-Bartsch, 2022) all lead to incorrect learning in the long run. In a social learning setting, the presence of biased agents can also lead to incorrect long-run beliefs for unbiased agents who are unaware of their bias (Bohren & Hauser, 2021).

REFERENCES

Agan, A., and S. Starr, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *Quarterly Journal of Economics* 133:1 (2017), 191–235. 10.1093/qje/qjx028

Aigner, D. J., and G. G. Cain, "Statistical Theories of Discrimination in Labor Markets," *ILR Review* 30:2 (1977), 175–187. 10.1177/001979397703000204

Albrecht, K., E. V. Essen, J. Parys, and N. Szech, "Updating, Self-Confidence, and Discrimination," *European Economic Review* 60 (2013), 144–169. 10.1016/j.euroecorev.2013.02.002

Antonovics, K., and B. G. Knight, "A New Look at Racial Profiling: Evidence from the Boston Police Department," this REVIEW 91:1 (2009), 163–177.
Anwar, S., and H. Fang, "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence," *American Economic Review* 96:1 (2006), 127–151. 10.1257/000282806776157579
Arnold, D., W. Dobbie, and C. Yang, "Racial Bias in Bail Decisions," *Quarterly Journal of Economics* 133 (2018), 1885–1932. 10.1093/qje/qjy012
Arnold, D., W. Dobbie, and P. Hull, "Measuring Racial Discrimination in Bail Decisions," *American Economic Review* 112:9 (2022), 2992–3038. 10.1257/aer.20201653
Arrow, K. J., "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees (eds.), *Discrimination in Labor Markets* (Princeton, NJ: Princeton University Press, 1973).
——— "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives* 12:2 (1998), 91–100. 10.1257/jep.12.2.91
Ayres, I., "Outcome Tests of Racial Disparities in Police Practices," *Justice Research and Policy* 4:1–2 (2002), 131–142. 10.3818/JRP.4.1.2002.131
Bar, R., and A. Zussman, "Identity and Bias: Insights from Driving Tests," *Economic Journal* 130 (2019), 1–23. 10.1093/ej/uez048
Bartoš, V., M. Bauer, J. Chytilová, and F. Matêjka, "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition," *American Economic Review* 106:6 (June 2016), 1437–1475.
Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova, "Powerful Women: Does Exposure Reduce Bias?" *Quarterly Journal of Economics* 124:4 (2009), 1497–1540. 10.1162/qjec.2009.124.4.1497
Becker, G., *The Economics of Discrimination* (Chicago: University of Chicago Press, 1957).
Bertrand, M., and E. Duflo, "Field Experiments on Discrimination" (pp. 309–393), in *Handbook of Field Experiments*, Vol. 1 (North-Holland, 2017). 10.1016/bs.hefe.2016.08.004
Bertrand, M., and S. Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94:4 (2004), 991–1013. 10.1257/0002828042002561
Bohren, J. A., and D. N. Hauser, "Learning With Heterogeneous Misspecified Models: Characterization and Robustness," *Econometrica* 89:6 (2021), 3025–3077. 10.3982/ECTA15318
Bohren, J. A., P. Hull, and A. Imas, "Systemic Discrimination: Theory and Measurement" (2022).
Bohren, J. A., A. Imas, and M. Rosenberg, "The Dynamics of Discrimination: Theory and Evidence," working paper (2019).
Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer, "Stereotypes," *Quarterly Journal of Economics* 131 (2016), 1753–1794. 10.1093/qje/qjw029
Breza, E., S. Kaur, and Y. Shamdasani, "The Morale Effects of Pay Inequality," *Quarterly Journal of Economics* 133:2 (2018), 611–663. 10.1093/qje/qjx041
Bursztyn, L., A. L. González, and D. Yanagizawa-Drott, "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia," *American Economic Review* 110 (2020), 2997–3029. 10.1257/aer.20180975
Bushong, B., and T. Gagnon-Bartsch, "Learning with Misattribution of Reference Dependence," *Journal of Economic Theory* 203 (2022), 1–45.
Charles, K. K., and J. Guryan, "Studying Discrimination: Fundamental Challenges and Recent Progress," *Annual Review of Economics* 3 (2011), 479–511. 10.1146/annurev.economics.102308.124448
Coffman, K. B., C. L. Exley, and M. Niederle, "The Role of Beliefs in Driving Gender Discrimination," *Management Science* 67 (2021), 3551–3569. 10.1287/mnsc.2020.3660
De Quidt, J., J. Haushofer, and C. Roth, "Measuring and Bounding Experimenter Demand," *American Economic Review* 108:11 (2018), 3266–3302. 10.1257/aer.20171330
Delavande, A., "Pill, Patch, or Shot? Subjective Expectations and Birth Control Choice," *International Economic Review* 49:3 (2008), 999–1042. 10.1111/j.1468-2354.2008.00504.x
Dharmapala, D., and S. L. Ross, "Racial Bias in Motor Vehicle Searches: Additional Theory and Evidence," *Contributions to Economic Analysis and Policy* 3:1 (2004), 89–111.
Dobbie, W., and R. G. Fryer, "The Impact of Youth Service on Future Outcomes: Evidence from Teach for America," *B.E. Journal of Economic Analysis and Policy* 15:3 (2015), 1031–1066. 10.1515/bejeap-2014-0187
Durlauf, S. N., "Racial Profiling as a Public Policy Question: Efficiency, Equity, and Ambiguity," *American Economic Review* 95:2 (2005), 132–136. 10.1257/000282805774669646
Epstein, L. G., J. Noor, and A. Sandroni, "Non-Bayesian Learning," *B.E. Journal of Theoretical Economics* 10:1 (2010), 0000102202193517041623. 10.2202/1935-1704.1623
Fang, H., and A. Moro, "Theories of Statistical Discrimination and Affirmative Action: A Survey" (pp. 133–200), in *Handbook of Social Economics*, Vol. 1 (Elsevier, 2011). 10.1016/B978-0-444-53187-2.00005-X
Fershtman, C., and U. Gneezy, "Discrimination in a Segmented Society: An Experimental Approach," *Quarterly Journal of Economics* 116 (February 2001), 351–377. 10.1162/003355301556338
Fiske, S. T., "Stereotyping, Prejudice, and Discrimination" (pp. 357–411), in *The Handbook of Social Psychology*, Vol. 2 (McGraw-Hill, 1998).
Giustinelli, P., "Group Decision Making with Uncertain Outcomes: Unpacking Child-Parent Choice of the High School Track," *International Economic Review* 57:2 (2016), 573–602. 10.1111/iere.12168
Glover, D., A. Pallais, and W. Pariente, "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores," *Quarterly Journal of Economics* 132 (July 2017), 1219–1260. 10.1093/qje/qjx006
Goncalves, F., and S. Mello, "A Few Bad Apples? Racial Bias in Policing," *American Economic Review* 111:5 (2019), 1406–1441. 10.1257/aer.20181607
Grau, N., and D. Vergara, "An Observational Implementation of the Outcome Test with an Application to Ethnic Prejudice in Pretrial Detentions," University of Chile working paper wp514 (2021).
Haavelmo, T., "The Probability Approach in Econometrics," *Supplement to Econometrica* 12 (1944), 1–115.
Hedegaard, M. S., and J.-R. Tyran, "The Price of Prejudice," *American Economic Journal: Applied Economics* 10:1 (2018), 40–63. 10.1257/app.20150241
Hengel, E., "Publishing While Female," working paper (December 2019).
Hilton, J. L., and W. V. Hippel, "Stereotypes," *Annual Review of Psychology* 47 (1996), 237–271. 10.1146/annurev.psych.47.1.237
Hübert, R., and A. T. Little, "A Behavioral Theory of Discrimination in Policing," working paper (2020).
Hull, P., "What Marginal Outcome Tests Can Tell Us about Racially Biased Decision-Making," NBER working paper 28503 (2021).
Jensen, R., "The (Perceived) Returns to Education and the Demand for Schooling," *Quarterly Journal of Economics* 125 (May 2010), 515–548. 10.1162/qjec.2010.125.2.515
Judd, C. M., and B. Park, "Definition and Assessment of Accuracy in Social Stereotypes," *Psychological Review* 100:1 (1993), 109–128. 10.1037/0033-295X.100.1.109
Kessler, J. B., C. Low, and C. D. Sullivan, "Incentivized Resume Rating: Eliciting Employer Preferences without Deception," *American Economic Review* 109:11 (2019), 3173–3174. 10.1257/aer.20181714
Knowles, J., N. Persico, and P. Todd, "Racial Bias in Motor Vehicle Searches: Theory and Evidence," *Journal of Political Economy* 109:1 (2001), 203–229. 10.1086/318603
Knox, D. E., W. Lowe, and J. Mummolo, "Administrative Records Mask Racially Biased Policing," *American Political Science Review* 114 (2020), 619–637. 10.1017/S0003055420000039
Kravitz, D. A., and J. Platania, "Attitudes and Beliefs about Affirmative Action: Effects of Target and of Respondent Sex and Ethnicity," *Journal of Applied Psychology* 78:6 (1993), 928–938. 10.1037/0021-9010.78.6.928
Lepage, L.-P., "Endogenous Learning and the Persistence of Employer Biases in the Labor Market," mimeo (2020).
List, J. A., "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *Quarterly Journal of Economics* 119 (February 2004), 49–89. 10.1162/003355304772839524
Manski, C. F., "Measuring Expectations," *Econometrica* 72:5 (2004), 1329–1376. 10.1111/j.1468-0262.2004.00537.x
Mobius, M., and T. Rosenblat, "Why Beauty Matters," *American Economic Review* 96:1 (2006), 222–235. 10.1257/000282806776157615

Paluck, E. L., S. Green, and D. P. Green, "The Contact Hypothesis Reevaluated," *Behavioural Public Policy* 3 (2018), 129–158. 10.1017/bpp.2018.25

Parsons, C. A., J. Sulaeman, M. C. Yates, and D. S. Hamermesh, "Strike Three: Discrimination, Incentives, and Evaluation," *American Economic Review* 101 (June 2011), 1410–1435. 10.1257/aer.101.4.1410

Phelps, E. S., "The Statistical Theory of Racism and Sexism," *American Economic Review* 62:4 (1972), 659–661.

Pope, D. G., and J. R. Sydnor, "What's in a Picture? Evidence of Discrimination from Prosper.com," *Journal of Human Resources* 46:1 (2011), 53–92. 10.1353/jhr.2011.0025

Rabin, M., and J. L. Schrag, "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics* 114:1 (1999), 37–82. 10.1162/003355399555945

Rao, G., "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools," *American Economic Review* 109:3 (2019), 774–809. 10.1257/aer.20180044

Schneider, D., A. Hastorf, and P. Ellsworth, *Person Perception* (Reading, MA: Addison-Wesley, 1979).

Schwartzstein, J., "Selective Attention and Learning," *Journal of the European Economic Association* 12:6 (2014), 1423–1452. 10.1111/jeea.12104

Sen, M., and O. Wasow, "Race as a Bundle of Sticks: Designs That Estimate Effects of Seemingly Immutable Characteristics," *Annual Review of Political Science* 19 (2016), 499–522. 10.1146/annurev-polisci-032015-010015

Simoiu, C., S. Corbett-Davies, and S. Goel, "The Problem of Infra-Marginality in Outcome Tests for Discrimination," *Annals of Applied Statistics* 11:3 (2017), 1193–1216. 10.1214/17-AOAS1058

Tilcsik, A., "Statistical Discrimination and the Rationalization of Stereotypes," *American Sociological Review* 86:1 (2021), 93–122. 10.1177/0003122420969399

Wiswall, M., and B. Zafar, "Determinants of College Major Choice: Identification Using an Information Experiment," *Review of Economic Studies* 82:2 (2015), 791–824. 10.1093/restud/rdu044