DISCUSSION PAPER SERIES

DP13790
(v. 3)

Inaccurate Statistical Discrimination: An Identification Problem

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

INDUSTRIAL ORGANIZATION

LABOUR ECONOMICS

ISSN 0265-8003

Inaccurate Statistical Discrimination: An Identification Problem

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

Discussion Paper DP13790
First Published 11 June 2019
This Revision 23 March 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Industrial Organization
- Labour Economics

Inaccurate Statistical Discrimination: An Identification Problem

Abstract

Discrimination, defined as differential treatment by group identity, is widely studied in economics. Its source is often categorized as taste-based or statistical (belief-based)---a valuable distinction for policy design and welfare analysis. However, in many situations individuals may have inaccurate beliefs about the relevant characteristics of different groups. This paper demonstrates that this possibility creates an identification problem when isolating the source of discrimination. A review of the empirical discrimination literature in economics reveals that a small minority of papers---fewer than 7%---consider inaccurate beliefs. We show both theoretically and experimentally that, if not accounted for, such inaccurate statistical discrimination will be misclassified as taste-based. We then examine three alternative methodologies for differentiating between different sources of discrimination: varying the amount of information presented to evaluators, eliciting their beliefs, and presenting them with accurate information. Importantly, the latter can be used to differentiate whether inaccurate beliefs are due to a lack of information or motivated factors.

JEL Classification: D90, J71

Keywords: discrimination, Inaccurate Beliefs, model misspecification

Aislinn Bohren - abohren@sas.upenn.edu
University of Pennsylvania and CEPR

Kareem Haggag - kareem.haggag@cmu.edu
Carnegie Mellon University

Alex Imas - Alex.Imas@chicagobooth.edu
University of Chicago

Devin G. Pope - devin.pope@chicagobooth.edu
University of Chicago

Acknowledgements
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 and Jihong Song provided excellent research assistance. Bohren gratefully acknowledges financial support from NSF grant SES-1851629. The experiment received IRB approval at CMU.

Inaccurate Statistical Discrimination: An Identification Problem

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

March 20, 2021

Abstract

Discrimination, defined as differential treatment by group identity, is widely studied in economics. Its source is often categorized as taste-based or statistical (belief-based)—a valuable distinction for policy design and welfare analysis. However, in many situations individuals may have inaccurate beliefs about the relevant characteristics of different groups. This paper demonstrates that this possibility creates an identification problem when isolating the source of discrimination. A review of the empirical discrimination literature in economics reveals that a small minority of papers—fewer than 7%—consider inaccurate beliefs. We show both theoretically and experimentally that, if not accounted for, such *inaccurate statistical discrimination* will be misclassified as taste-based. We then examine three alternative methodologies for differentiating between different sources of discrimination: varying the amount of information presented to evaluators, eliciting their beliefs, and presenting them with accurate information. Importantly, the latter can be used to differentiate whether inaccurate beliefs are due to a lack of information or motivated factors.

KEYWORDS: Discrimination, Inaccurate beliefs, Model misspecification

*Bohren: University of Pennsylvania, abohren@gmail.com, Haggag: Social and Decision Sciences, Carnegie Mellon University, kareem.haggag@cmu.edu. Imas: Booth School of Business, University of Chicago, aimas@uchicago.edu. Pope: Booth School of Business, University of Chicago, devin.pope@chicagobooth.edu. 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 and Jihong Song provided excellent research assistance. Bohren gratefully acknowledges financial support from NSF grant SES-1851629. The experiment received IRB approval at CMU.

1 Introduction

Economists define discrimination as differential treatment of otherwise identical individuals from different social groups (i.e. race, gender, age, etc.). Discrimination has been shown to be prevalent in labor markets, housing markets, credit markets, and online consumer markets among others. A variety of empirical techniques have been used to document discrimination, including audit studies, correspondence studies, and outcomes-based tests (for recent reviews of the discrimination literature, see Bertrand and Duflo (2017) and Charles and Guryan (2011)).

In addition to establishing a causal link between group identity and differential treatment, economists also often attempt to identify its source. Sources are typically categorized into one of two types. In the case of taste-based discrimination (Becker, 1957), an individual or firm has animus towards members of a particular group, and may therefore choose to discriminate against them because he receives disutility from providing services to or interacting with members of that group. In the case of statistical (belief-based) discrimination, differential treatment may occur because productivity is unobserved and a particular group is *correctly* perceived to have a lower average productivity. This difference in productivity may be due to exogenous differences (Phelps, 1972) or as part of a self-fulfilling equilibrium (Arrow, 1973). Alternatively, the group's productivity may be perceived to have a different variance (Aigner and Cain, 1977) or information about productivity may be viewed as more or less informative.

Distinguishing between these forms of discrimination is important for several reasons. First, designing an effective policy intervention to reduce discrimination crucially depends on the source of discrimination. Second, welfare and efficiency analyses differ as a function of the source. For example, because statistical discrimination is typically assumed to be driven by rational expectations, it is sometimes referred to as "efficient discrimination" in that it is the optimal response to a signal-extraction problem. Finally, the extent to which competitive markets will eliminate discrimination depends on its source (see Fang and Moro (2011) for review).

However, a large literature in psychology and economics has shown that people's beliefs can be incorrect in a wide variety of settings. This paper outlines the implication of incorrect beliefs for identifying the source of discrimination. We refer to belief-based discrimination that stems from incorrect beliefs as *inaccurate statistical discrimination.* Just as it is important to distinguish between taste-based and accurate statistical discrimination for policy design and welfare analysis, we show that it is also critically important to separate inaccurate from accurate statistical discrimination. For example, if discrimination stems from inaccurate beliefs, an effective policy response could be providing individuals with information about the correct distributions.[1] Importantly, as we formally show in Section 3, allowing for beliefs to be inaccurate generates an identification problem for methods commonly used to isolate the source of discrimination, such as outcomes-based tests.

Two broad sources may lead to inaccurate beliefs. First, research in psychology has shown that heuristics and biases may generate beliefs that are systematically incorrect, leading to inaccurate stereotypes about certain groups (Schneider, Hastorf, and Ellsworth (1979); Judd and Park (1993); Hilton and Hippel (1996); see Fiske (1998) for review). Bordalo, Coffman, Gennaioli, and Shleifer (2016) present a formal model for inaccurate stereotype formation based on the representativeness heuristic. There, evaluators overweight the prevalence of characteristics that differ most between groups, and end up believing that the 'representative' type is more prevalent than it actually is.[2] Biased beliefs can also arise in a dynamic learning setting when individuals either have incorrect models of how others evaluate workers (Bohren, Imas, and Rosenberg, 2019) or have different updating rules that depend on group identity (Albrecht, Von Essen, Parys, and Szech, 2013).

Second, inaccurate beliefs may simply be due to a lack of information. A fully rational actor may lack the relevant information necessary to form correct beliefs. For example, an employer may have an unbiased prior belief about the average productivity of individuals from two different social groups, but be unaware that there is positive selection into the job application process for members of one group. This employer will be less likely to hire members of a particular group because she has an inaccurate belief about productivity differences due to a lack of information about the selection process.

Regardless of their source, inaccurate beliefs have been shown to exist in a variety of important domains, including the value of human capital formation (Jensen, 2010), the prevalence of affirmative action (Kravitz and Platania, 1993), and the extent of wealth inequality in the US (Norton and Ariely, 2011). Learning will mitigate inaccurate beliefs in some settings. But in other situations, there will be little or no feedback on the deci-

[1] See for example, Jensen (2010) in the case of inaccurate beliefs about the returns to education, or Bursztyn, González, and Yanagizawa-Drott (2018) in the case of inaccurate beliefs about the beliefs of others, i.e. pluralistic ignorance

[2] For example, because there are more older people in Florida than the rest of the country, a person may incorrectly believe that most Floridians are old.

sions being made, leading to learning traps in which inaccurate beliefs persist. Inaccurate beliefs may persist in the long run even when employers are perfectly Bayesian if, for example, they face a trade off between learning about the productivity of groups through hiring or maximizing cost-effectiveness (Lepage, 2020). Further, as shown theoretically by Gagnon-Bartsch, Rabin, and Schwartzstein (2018), learning may not correct beliefs if information is filtered through an incorrect model of the world. Importantly, besides affecting the level of discrimination of the person holding them, inaccurate beliefs can also have negative spillovers on the discrimination of *other* evaluators if people learn from both signals and the actions of others (Hübert and Little, 2020).

Although the potential of inaccurate beliefs has been discussed in theoretical work, the vast majority of empirical tests rely on the assumption that beliefs are correct.[3] To examine the extent to which inaccurate beliefs have been considered as a source of discrimination in empirical work, we conducted an in-depth review of the economics literature on discrimination, spanning ten journals and nearly three decades (1990-2018) of papers. Section 2 presents our findings. The vast majority of papers found evidence for discrimination (97.1%) and a large plurality (61.9%) differentiated between taste-based versus statistical motives when discussing its source. However, only a small proportion (10.5%) make the distinction between accurate and inaccurate beliefs, and a smaller fraction still (6.7%) incorporates this distinction into the analysis.

Motivated by these findings, we formally demonstrate how the possibility of inaccurate beliefs generates an identification problem for attempts to isolate the source of discrimination. This analysis shows that failing to allow for mistaken beliefs can lead to a misclassification of belief-based discrimination as stemming from taste-based sources. We first distinguish between *discrimination*, which is a property of behavior, and *partiality* which is a primitive of the model (i.e. preferences and beliefs). In the context of the labor market, discrimination occurs when two workers generate identical signals but are evaluated differently based on their group identity. Partiality refers to the source of this disparity based on model primitives: belief-based partiality corresponds to evaluators having different beliefs about decision-relevant attributes (e.g. productivity) depending on group identity, while preference-based partiality corresponds to evaluators setting

[3]For example, Arrow (1973) notes that employers may be more willing to accept subjective probabilities that accord with their actions. Similarly, Arrow (1998) writes "the discussion of statistical discrimination so far assumes that the employers or creditors use all the information available throughout the economy. In Bayesian terms, the posterior information is sufficiently rich to make the contribution of the prior minimal. But of course this is not so." Notably, in neither paper is the potential of inaccurate beliefs addressed formally.

different hiring thresholds based on group identity. The former is typically referred to as taste-based discrimination, prejudice, or animus, while the latter is referred to as statistical discrimination when beliefs about the distribution of the job-relevant factor are correct. We expand the standard approach by considering both accurate belief-based partiality (traditional statistical discrimination) and *inaccurate* belief-based partiality (inaccurate statistical discrimination).

To facilitate the analysis, we introduce the concept of *isodiscrimination curves* as the set of preferences and beliefs that lead to equivalent discrimination. Similar to indifference curves in utility theory, there is a collection of isodiscrimination curves that each correspond to a level of discrimination on a given set of signals. Notably, it is readily apparent that the same level of discrimination can stem from a continuum of preference and belief combinations. Isodiscrimination curves provide a tractable way to examine what a given data set can and cannot identify.

We first show that collecting data on evaluation decisions, e.g. hiring, as a function of group identity and signals can be used to identify the isodiscrimination curve and rule out the case of no discrimination. Consider the correspondence study method, which uses an experimental technique where evaluators receive identical signals from people of different group identities. Perhaps the most famous example is the resume study by Bertrand and Mullainathan (2004), where experimenters sent out identical resumes to potential employees—where the only thing that differed was whether the group identity associated with the applicant's name—and looked at call-back rates as a measure of discrimination. Depending on the richness of the data, this technique can either identify an upper and lower bound on the isodiscrimination curve or the specific isodiscrimination curve. However, even if the latter case is possible, this method does not identify the source or even rule out any form of partiality. Let belief and preference partiality (neutrality) refer to an evaluator whose beliefs or preferences differ (do not differ) by group identity. In the case of the correspondence study method, for any evaluator with preference partiality and belief neutrality, there exists a continuum of evaluators with belief-based partiality and lower (or no) preference-based partiality that would exhibit the same level of observed discrimination. This includes belief-based partiality about the mean of the relevant distribution, its variance, or the precision of signals of the relevant characteristic. In turn, additional data is necessary to isolate the source of discrimination.

In some cases, researchers have access to not only the evaluation decisions, but also to the underlying outcome distributions. In these cases, studies attempt to identify the source of discrimination through a technique often referred to as an outcomes-based test, which compares evaluation decisions to the true underlying distribution of the relevant characteristics for these decisions. This commonly-used technique has been employed in attempts to identify the source of discrimination in many domains, including lending, policing and bail decisions (Pope and Sydnor, 2011; Knowles, Persico, and Todd, 2001; Antonovics and Knight, 2009). For example, a researcher may compare differences in lending rates between two groups to differences in their loan default rates. Under the commonly-made assumption of accurate beliefs, we show that this technique isolates both the isodiscrimination curve and the exact source of discrimination. However, identification depends critically on the assumption that evaluators have accurate beliefs. Without this assumption, a *continuum* of preference and belief combinations can generate equivalent discrimination regardless of the underlying outcome distribution. As we formally demonstrate, the only case that can sometimes be ruled out with outcome data is accurate statistical discrimination, i.e. an evaluator with accurate beliefs and preference neutrality.[4]

Further, erroneously assuming that an evaluator has accurate beliefs when using the outcomes-based method leads a researcher to mistakenly attribute the share of discrimination arising from inaccurate beliefs to preferences. Depending on whether the inaccurate beliefs increase or decrease discrimination, the researcher will over- or underestimate the degree to which an evaluator has preferences that favor one of the groups. As an example of this identification issue, consider a study that measures productivity outcomes and finds evidence for discrimination in the treatment of two groups. If the researcher observes that both groups have identical distributions of productivity, she concludes that because accurate statistical discrimination cannot explain behavior, the source of the observed discrimination must be preference-based. However, an alternative explanation is that evaluators have incorrect beliefs and use these beliefs to engage in inaccurate statistical discrimination. Without further data, it is impossible to distinguish between preference channels and incorrect beliefs. Alternatively, consider a study where discrimination is documented and the underlying distributions do differ.

[4]Other papers have highlighted identification challenges for outcomes-based tests, including the problem of infra-marginality (Ayres, 2002; Simoiu, Corbett-Davies, and Goel, 2017), as well as issues related to relying on administrative data that may condition on a post-treatment outcome (Knox, Lowe, and Mummolo, 2020). We raise a complementary concern that remains even if these other issues are solved.

The study claims no evidence for prejudice as group differences can explain differential treatment. As we show, any departures from correct beliefs imply that prejudice *does* play a role in the observed discrimination.

Finally, we outline what type of data can be used to overcome the identification problem. One method, which we illustrate in Section 4, is to collect data on the subjective beliefs of evaluators. Combined with observing the evaluation decisions and signals, this allows for identification of preference and belief partiality. Data on the true outcome distributions is required to determine whether beliefs are accurate. Another method is varying the precision of information by increasing the number of signal draws supplied to evaluators. 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. differential means and variances of the relevant characteristic, or differential signal precisions). Notably, this method requires multiple signals from the same domain (e.g. number of positive reviews); if the multiple signals are from different domains (e.g. SAT scores and education history), then the identification problem persists.[5]

We then use a stylized experimental setting to demonstrate the potential pitfalls of the identification problem and propose a portable method of addressing it. Participants are recruited to take part in a hiring experiment, where some are assigned the role of "worker" and others the role of "employer". Workers begin the first stage of the experiment by answering a series of questions. Employers are then shown profiles of 20 potential employees (workers from the initial stage) and asked the maximum wage they would be willing to pay to hire each. The profiles include a variety of worker-specific characteristics, such as their country of origin (US vs. India), gender, and age, interspersed amongst other information such as their beverage and movie preferences. Importantly, profiles do not include any information about performance on the question task. One worker profile is then selected. If the offered wage is above a randomly determined threshold, then the employer hires the worker; the worker earns a bonus and the employer is paid proportional to how many questions the worker answered correctly. If the offered wage is below the threshold, the worker is not hired.

[5]Another approach derives predictions from a specific structural model of biased beliefs and takes these predictions to the data. Arnold, Dobbie, and Yang (2018b) compare the distributions of pre-trial misconduct of marginal black and white defendants. They argue that the distributional differences are consistent with bail judges holding incorrect stereotypes, as modeled by Bordalo et al. (2016), about the release risk of black defendants.

We find that employers discriminate based on worker characteristics. Americans and females receive systematically lower wage offers than Indians and males. We find little evidence for age discrimination. 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 believe that members of those groups answer fewer questions correctly on average than Indian and male workers. Since they lack information on the productivity of any given worker, employers use these group statistics to inform their compensation decisions. Alternatively, employers may be prejudiced towards members of the discriminated group and offer them lower wages because they do not want to reward them.

As discussed above, outcomes-based tests are often used to distinguish between statistical and taste-based discrimination by comparing the distributions of compensation decisions to the "ground truth"—the true distributions of performance across groups. If the difference in the group-specific performance distributions is similar to the difference in wages, then this is used as evidence that evaluators are statistically discriminating, i.e. differential treatment is due to belief-based partiality. Otherwise, discrimination is categorized as preference based. 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 perform directionally better than Indians on the task (though the difference is not statistically significant), while females perform less well than males. Under the assumption of accurate beliefs, we would conclude that the source of discrimination against Americans is preference based. Further, because the level of discrimination against women is substantially smaller than the actual gap in performance, this approach would conclude that there was preference based discrimination against men.

However, an alternative explanation is that individuals have no preference-based motives towards 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" distributions. 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. Specifically, if agents are provided with credible information on the relevant distributions, those with inaccurate beliefs should adjust their behavior accordingly. However, if mistaken beliefs merely mask an underlying animus, then agents are unlikely to change their behavior in response.

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 10 additional workers. We find that employers significantly changed their wage offers in the direction consistent with correcting their inaccurate beliefs. This methodology is portable outside of our stylized experimental setting as a way to identify animus-driven inaccurate beliefs.

The paper proceeds as follows. Section 2 presents a review of the economics literature on discrimination, demonstrating that few papers consider mistaken beliefs when attempting to isolate its source. Section 3 formally outlines how a failure to account for the possibility of inaccurate beliefs leads to an identification problem. Section 4 illustrates a potential methodology for overcoming this identification problem through a stylized experiment. Section 5 concludes.

2 Survey of the Literature

We conducted a systematic survey of the economics literature on discrimination in order to determine: (1) how often papers seek to distinguish between taste-based and belief-based (statistical) sources of discrimination; (2) how often papers seek to distinguish between accurate and inaccurate beliefs for belief-based sources of discrimination. Table 1 tabulates the 105 papers published in 10 top economics journals between 1990 and 2018 that test for evidence of discrimination. Most papers that met our inclusion criteria (outlined below) 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

Table 1. Summary of Literature Review on Discrimination

	All: 1990 - 2018		**Recent: 2014 - 2018**	
	# Papers	*% Total*	*# Papers*	*% Total*
Papers meeting inclusion criteria	105	100.0%	31	100.0%
Evidence of discrimination	102	97.1%	31	100.0%
Discuss taste-based versus statistical source	65	61.9%	23	74.2%
Test for taste-based versus statistical source	49	46.7%	16	51.6%
Discuss accurate versus inaccurate beliefs	11	10.5%	5	16.1%
Test for inaccurate beliefs	7	6.7%	3	9.7%
Measure beliefs	7	6.7%	3	9.7%

of the papers (46.7%) attempted to distinguish between these two sources through a formal test. However, very few papers even discussed the possibility that beliefs may be inaccurate (10.5%), and fewer still examined whether beliefs were accurate or inaccurate (6.7%).[6] Despite the lack of discussion and explicit tests, we would argue that inaccurate statistical discrimination is a reasonable alternative to the interpretation chosen by the authors in nearly all of these cases.

We classified papers as "discuss taste-based versus statistical source" if preference versus belief-based motives for the documented discrimination were discussed in the text, and as "test for taste-based versus statistical source" if the paper either explicitly tested between different models of preference versus belief-based discrimination or implicitly tested the predictions of a belief-based model while taking the taste-based model as the null hypothesis. If a paper mentioned inaccurate or biased beliefs as a potential source of discrimination, it was classified as "discuss accurate versus inaccurate beliefs." Papers that tested whether inaccurate beliefs could be driving discrimination, either by directly eliciting beliefs or through other tests, were classified as "test for inaccurate beliefs." Finally, papers that elicited beliefs were classified as "measure beliefs." Three of the seven papers in this category did not test whether these elicited beliefs were accurate.

Method. In this section, we outline the method that we used to determine which papers to include in the survey and the data that we collected for each paper.

Inclusion Criteria. We focused on empirical papers published between 1990 and 2018

[6]The papers that tested for inaccurate beliefs include List (2004); Hedegaard and Tyran (2018); Mobius and Rosenblat (2006); Fershtman and Gneezy (2001); Arnold, Dobbie, and Yang (2018a); Agan and Starr (2017). Beaman, Chattopadhyay, Duflo, Pande, and Topalova (2009) use the Implicit Association Test (IAT) to elicit a 'taste' against female politicians, though the authors also interpret IAT scores as a measure of implicit beliefs. We include this paper on the list as well.

in the following journals: American Economic Journal: Applied, American Economic Journal: Policy, American Economic Review (excluding the Papers & Proceedings issue), Econometrica, Journal of the European Economic Association, Journal of Labor Economics, Journal of Political Economy, the Quarterly Journal of Economics, Review of Economic Studies, and Review of Economics and Statistics. We acknowledge that the economics literature on discrimination includes important contributions from other journals. We restricted attention to these ten journals as a representative sample in order for the scope of the survey to include a manageable number of papers.

We proceeded in two steps to determine whether to include a paper published in the relevant time frame and journals. First, in each journal, we searched for all empirical papers that had at least one of the search terms *{discrimination, prejudice, bias, biases, biased, disparity, disparities, stereotype, stereotypes, premium}* in the title, or at least one of the search terms *{discrimination, prejudice}* in the abstract, or at least one of the search terms from *{racial, race, gender, sex, ethnic, religious, beauty}* and *{bias, biased, disparity, stereotype, stereotypes, premium}* in the abstract. Second, we restricted attention to papers that attempted to causally document differential treatment of individuals based on their group identity. This eliminated papers on unrelated topics, including the industrial organization literature on *price* discrimination, the financial literature on the *risk* premium, theoretical models, and the experimental literature that documents behavioral differences such as gender differences in risk preferences.[7]

Data Collection. For each paper that met our inclusion criteria, we recorded the following information: data source (laboratory experiment, field experiment, audit or correspondence study, observational data study, other), empirical method (reduced form analysis, structural analysis), group identity of interest (race, gender, ethnicity, religion, sexuality, class/income, other), domain of study (labor market, legal, education, financial, consumer purchases—non-financial, evaluations, other), measure of discrimination (i.e. difference in call back rates), whether the paper distinguishes between taste-based and statistical discrimination, whether the paper distinguishes between accurate and inaccurate statistical discrimination, whether discrimination was documented, whether the study identified the source of discrimination, and whether the study measured beliefs

[7] We also excluded some papers that met our objective criteria but which we viewed as not relevant to the spirit of the exercise. More specifically, we excluded papers that could not be classified as either a "Yes" or "No" for the criteria outlined in Table 1. For example, Gneezy, Niederle, and Rustichini (2003) examine behavioral differences between men and women but do not study discrimination per se. Similarly, Cameron and Heckman (2001) examine the extent to which the racial and ethnic gap in college attendance can be explained by long-run versus short-run factors but do not address discrimination.

Table 2. Publications by Journal and Decade

	Number of Papers			
	1990-99	**2000-09**	**2010-2018**	**Total**
AEJ: Applied	0	1	7	8
AEJ: Policy	0	0	2	2
AER	4	7	6	17
EMA	0	0	0	0
JEEA	0	1	1	2
JLE	2	8	12	22
JPE	2	6	1	9
ReStud	1	2	3	6
ReStat	5	6	11	22
QJE	4	4	9	17
Total	**18**	**35**	**52**	**105**

about an individual's predicted attribute by group identity.

Summary Statistics. We found 105 papers that met our inclusion criteria. Table 2 lists the number of papers broken down by journal and decade of publication. The full list of papers is included in the Supplemental Material. Out of the papers surveyed, 11 conducted audit or correspondence studies, 7 conducted another type of field experiment, 3 conducted a laboratory experiment and 84 analyzed observational data.

Discrimination was studied for a variety of group identities and in a variety of domains. The most frequent group identities were race (58 papers) and gender (37 papers), followed by physical traits / appearance (7 papers) and ethnicity (6 papers). The most frequent domain was labor markets (58 papers), followed by legal contexts (12 papers), education (9 papers), non-financial consumer markets (6 papers) and financial markets (5 papers). Table 3 summarizes the papers by group identity and domain. Some papers in the survey studied multiple group identities or domains; therefore, some papers are counted in multiple rows of the table.

Table 3. Type and Domain of Discrimination

	All Papers	Evidence of Discrimination	
	# Papers	*# Papers*	*% Total*
Group Identity			
Race	58	56	96.6%
Gender	37	35	94.6%
Ethnicity	6	6	100.0%
Religion	1	1	100.0%
Sexuality	1	1	100.0%
Class/Income	1	1	100.0%
Physical Traits / Appearance	7	7	100.0%
Other	5	5	100.0%
Domain of Discrimination			
Labor Market	58	57	98.3%
Legal	12	12	100.0%
Education	9	9	100.0%
Financial	5	4	80.0%
Consumer Markets (not financial)	6	6	100.0%
Other	17	16	94.1%

3 A Model of Discrimination with Inaccurate Beliefs

In this section, we develop a model of discrimination with inaccurate beliefs. An evaluator learns about a worker's productivity from a signal, then decides whether to hire the worker. Inaccurate beliefs refer to the possibility that the evaluator misperceives key population statistics, such as how the distribution of productivity 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 only reject accurate statistical discrimination—the method cannot separate whether discrimination stems from preferences or inaccurate beliefs, and therefore, it cannot identify the source. Further, erroneously assuming accurate beliefs and using the outcomes-based method leads to a misclassification of discrimination driven by inaccurate beliefs as arising from preferences. We conclude by outlining two methods—eliciting beliefs and manipulating information—that can be used to determine whether discrimination stems from preferences or beliefs.

3.1 Set-up

3.1.1 Model

Worker. 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 $\mu_g \in \mathbb{R}$ and concentration $\tau_g > 0$. 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 precision $\eta_g > 0$. Without loss of generality, we focus on discrimination against workers from group F.

Evaluator. 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 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 corresponds to the case in which these subjective distributions differ from the true distributions.[8]

Given these subjective distributions, the evaluator uses Bayes rule to update her belief about the worker's productivity. She 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.[9] 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 signal s, where $\hat{E}_\theta$ denotes the expectation taken with respect to θ's subjective beliefs.

[8] An additional form of inaccurate beliefs that we do not discuss 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 $b > 0$ i.e. $s = a + b + \epsilon$, and therefore, the evaluator discounts a signal to $s - b$ for group F.

[9] 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.

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 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 even 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*.*

Adopting terminology from Bartoš, Bauer, Chytilová, and Matějka (2016), we define how the evaluator's type informs her perception of the market. An evaluator believes the market is *lemon-dropping* for a group if the hiring threshold is low relative to perceived average productivity, $u_g < \hat{\mu}_g$, while she believes it is *cherry-picking* if the hiring threshold is high relative to average productivity, $\hat{\mu}_g < u_g$. In the former case, the evaluator wants to hire a worker from group g in the absence of a signal while in the latter case, she does not. A market is *mixed* if the evaluator believes it is cherry-picking for one group and lemon-dropping for the other group.

Discrimination. Discrimination captures the causal effect of group identity on hiring. Let $D(s, \theta) \equiv v(s, M, \theta) - v(s, F, \theta)$ denote the difference between the hiring decision for a worker from group M versus F by an evaluator of type θ who observes signal s. *Discrimination* occurs at signal s when two workers are evaluated differently based on their group identity, $D(s, \theta) \neq 0$; it occurs against group F if $D(s, \theta) > 0$ and against group M if $D(s, \theta) < 0$. Discrimination *exists* if there is an s such that $D(s, \theta) \neq 0$, and there is *no 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, while 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 set-up allows us to illustrate how inaccurate beliefs impact discrimination in a tractable and succinct way. Our main insights are robust to alternative forms of evaluations—i.e. selecting a wage offer or a rating from an interval—and to other productivity and signal distributions.

3.1.2 The Optimal Hiring Rule

In this section we 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 cut-off with respect to the signal.

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

$$\overline{s}(\theta, g) \equiv \left(\frac{\hat{\tau}_g + \hat{\eta}_g}{\hat{\eta}_g}\right) u_g - \frac{\hat{\tau}_g}{\hat{\eta}_g}\hat{\mu}_g. \tag{1}$$

The signal required to hire a worker is increasing in the evaluator's hiring preference u_g and decreasing in the prior belief about average productivity $\hat{\mu}_g$. In a cherry-picking market, it is increasing in the concentration of productivity $\hat{\tau}_g$ and decreasing in the signal precision $\hat{\eta}_g$. Intuitively, when the evaluator seeks workers in the top tail of the distribution, a higher signal realization is required to offset higher concentration. In contrast, when the evaluator would not hire a worker in the absence of a signal, the evaluator is willing to hire at lower signal realizations when the signal is more precise. These comparative statics reverse in a lemon-dropping market: when seeking to avoid workers in the bottom tail, the evaluator is willing to hire at lower signal realizations when productivity is more concentrated, while when the evaluator would hire a worker in the absence of a signal, the evaluator is willing to hire at lower signal realizations when the signal is less precise.

From Lemma 1, an evaluator of type θ discriminates against group F if she sets a higher signal threshold for group F, $\overline{s}(\theta, F) > \overline{s}(\theta, M)$. When this is the case, discrimination occurs on the interval of signals that lie between the two hiring thresholds, $s \in [\overline{s}(\theta, M), \overline{s}(\theta, F))$.

3.1.3 Equivalent Discrimination and Isodiscrimination Curves

We next derive the sets of beliefs and preferences that give rise to equivalent discrimination. From Lemma 1, types exhibit equivalent discrimination when they have preferences and beliefs that lead to the same signal thresholds for each group. Combining this observation with Eq. (1) yields the following characterization.

Proposition 1 (Equivalent Discrimination)**.** *For any constants $(s_M, s_F) \in \mathbb{R}^2$ with $s_F > s_F$ $(s_M > s_F)$, the set of types*

$$\left\{ (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g \in \{M,F\}} \;\middle|\; \begin{matrix} \dfrac{\hat{\tau}_M + \hat{\eta}_M}{\hat{\eta}_M} u_M - \dfrac{\hat{\tau}_M}{\hat{\eta}_M} \hat{\mu}_M = s_M \\ \dfrac{\hat{\tau}_F + \hat{\eta}_F}{\hat{\eta}_F} u_F - \dfrac{\hat{\tau}_F}{\hat{\eta}_F} \hat{\mu}_F = s_F \end{matrix} \right\} \tag{2}$$

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

Similar to indifference curves in utility theory, we can represent the sets of types that exhibit equivalent discrimination as *isodiscrimination curves*, where each curve corresponds to a different level of discrimination and is parameterized by a pair of signal thresholds $(s_M, s_F) \in \mathbb{R}^2$ with $s_M \neq s_F$.[10]

[10]The isodiscrimination curve corresponding to no discrimination is parameterized by $(s_M, s_F) \in \mathbb{R}^2$ such that $s_M = s_F$.

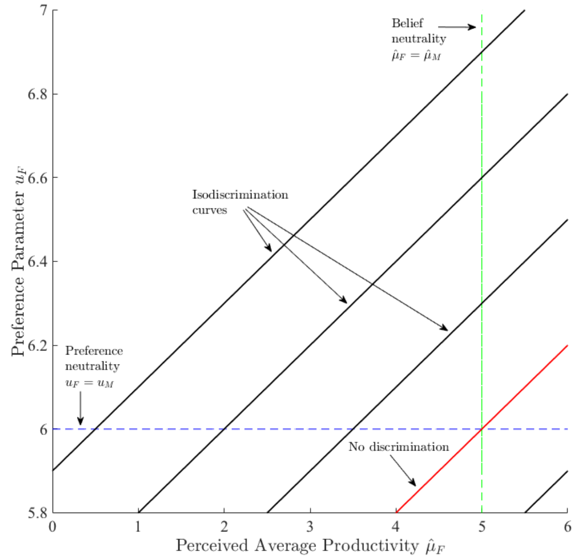


Figure 1. Isodiscrimination Curves
$(u_M, \hat{\mu}_M, \hat{\tau}_M, \hat{\eta}_M) = (6, 5, .5, 2)$, $(\hat{\tau}_F, \hat{\eta}_F) = (.5, 2)$

Fig. 1 illustrates several isodiscrimination curves in two-dimensions. Fixing the other parameters, each curve plots the continuum of preferences and perceived average productivities for group F that lead to a given level of discrimination. For example, an evaluator with mild preference partiality and extreme belief partiality exhibits equivalent discrimination to an evaluator with more extreme preference partiality and mild belief partiality. The red line traces out the isodiscrimination curve corresponding to no discrimination. Isodiscrimination curves above this line correspond to discrimination against group F, while isodiscrimination curves below this line correspond to discrimination against group M. Moving northwest from the line of no discrimination, i.e. increasing the preference against group F and the decreasing perceived average productivity of group F, leads to discrimination on a larger set of signals. The blue dotted line traces out preference neutrality (i.e. $u_F = u_M$) and the green dotted line traces out belief neutrality (i.e. $\hat{\mu}_F = \hat{\mu}_M$). The quadrant above and to the left of these lines corresponds to the region in which there is preference and belief partiality against group F, the quadrant below and to the left corresponds to the region in which there is belief partiality against group F and preference partiality in favor of group F, while the the quadrant above and to the right corresponds to the region in which belief partiality favors group F and preference partiality favors group M. As can be seen in the figure, a given level 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. Finally, the quadrant below and to the right of the lines of neutrality corresponds to the region in which there is preference and belief partiality against group M. All isodiscrimination curves in this region correspond to discrimination against group M.

3.2 Identifying the Source of Discrimination.

Researchers are often interested in identifying the *source* of discrimination i.e. the form of partiality that generates the observed discriminatory behavior. Informally, identifiability means that the property of interest can be backed out from available data.[11] In this section we explore how the possibility of inaccurate beliefs impacts such identification.

We first illustrate several difficulties that arise when allowing for inaccurate beliefs. Specifically, is is not possible to reject either preference based partiality or any of the forms of belief based partiality from observing solely the level of discrimination. Further, the traditional outcomes-based method used to classify source under the assumption of accurate beliefs can only be used to reject accurate statistical discrimination when beliefs are inaccurate—erroneously assuming accurate beliefs leads to a misclassification of source. We then present two methods that can separate preferences from beliefs.

To proceed, we assume that the researcher observes the group identity g, signal s and hiring decision v for each worker, and that the dataset includes a sufficiently rich set of workers such that the signal thresholds, and therefore, the isodiscrimination curve, are identified.[12]

3.2.1 An Identification Challenge

It is well known that identifying the level of discrimination—i.e. the isodicrimination curve—cannot distinguish between preference-based partiality and accurate belief-based partiality about the distributions of productivity (see for example Bertrand and Mullainathan 2004). The same insight extends to inaccurate beliefs and belief-based par-

[11] More formally, a property is *identified* if there exists an injective relationship between the observed data and the property (Haavelmo, 1944).

[12] 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 (for example, the classic resume study of Bertrand and 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.

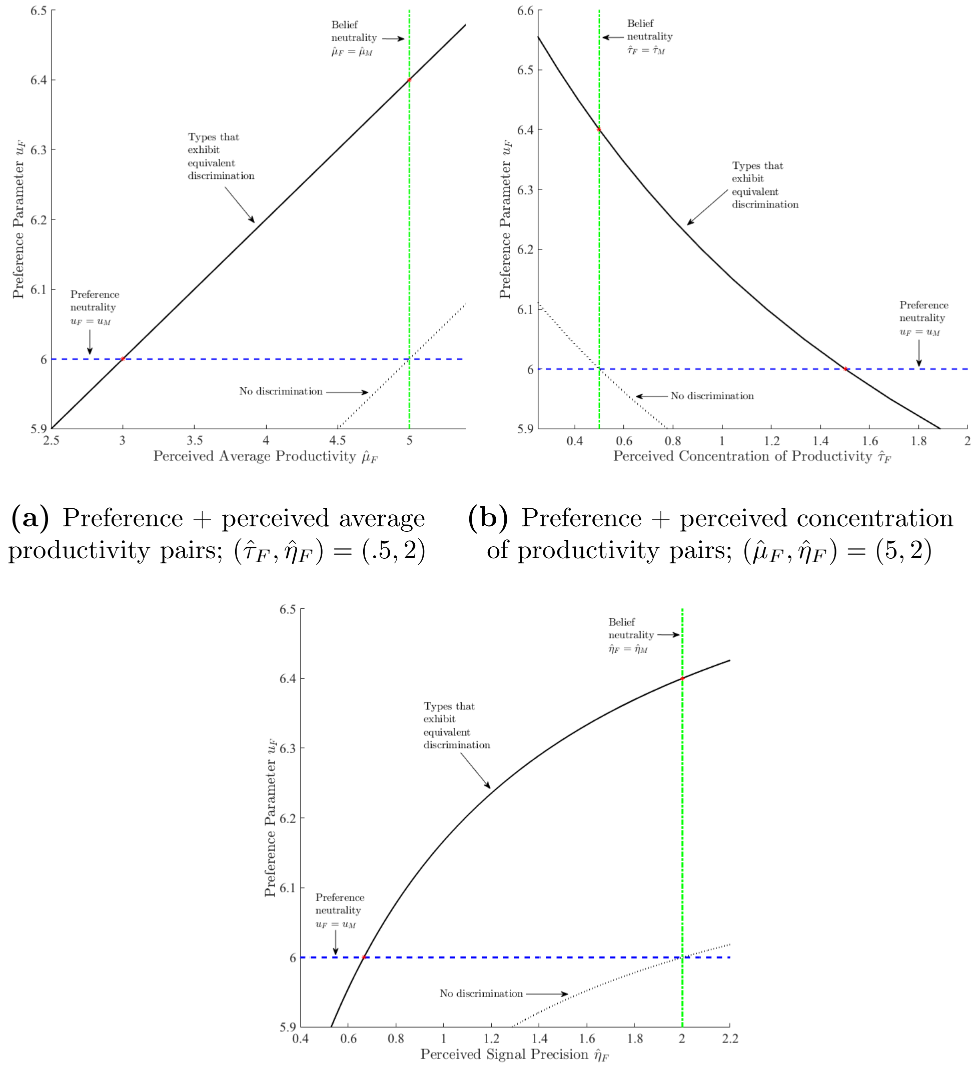
tiality about the signal distributions. To formalize this insight, we show that for each isodiscrimination curve, each form of partiality in isolation can generate the given level of discrimination. Therefore, identifying the isodiscrimination curve does not identify the source of discrimination or even rule out any of the potential sources.

Proposition 2 (Equivalent Sources)**.** *For any level of discrimination against group F, there are a continuum of types that exhibit equivalent discrimination, including:*

1. *A type with* preference partiality *against group F 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 that believes the market is cherry-picking (lemon-dropping) and has preference neutrality and belief partiality due to* higher (lower) concentration of productivity*, $\hat{\tau}_F > \hat{\tau}_M$ ($\hat{\tau}_F < \hat{\tau}_M$) and $(u_F, \hat{\mu}_F, \hat{\eta}_F) = (u_M, \hat{\mu}_M, \hat{\eta}_M)$;*
4. *A type that believes the market is cherry-picking (lemon-dropping) and has preference neutrality and belief partiality due to* lower (higher) signal precision*, $\hat{\eta}_F < \hat{\eta}_M$ ($\hat{\eta}_F > \hat{\eta}_M$) and $(u_F, \hat{\mu}_F, \hat{\tau}_F) = (u_M, \hat{\mu}_M, \hat{\tau}_M)$.*

From Proposition 2, when all other parameters are equal, a higher preference parameter or a lower perceived average productivity for group F relative to M generates discrimination against group F. For the other parameters, the direction of partiality that leads to discrimination against group F depends on the type of market: all else equal, a higher perceived concentration of productivity or a lower signal precision for group F generate discrimination against F in a cherry-picking market, while the opposite holds in a lemon-dropping market. This stems from the comparative static in Eq. (1) for the parameter of interest: as discussed following Lemma 1, how the parameter impacts the signal thresholds, and therefore, the level of discrimination, depends on the type of market.

Fig. 2 illustrates the evaluator types constructed in Proposition 2. Fixing an isodiscrimination curve and holding the other parameters equal across groups, the first panel illustrates the continuum of preference parameters and perceived average productivity for group F that exhibit this level of discrimination. This includes the type with belief neutrality described in part (i), denoted by the asterisk where the isodiscrimination curve intersects the line of belief neutrality, and the type with preference neutrality described in part (ii), denoted by the asterisk where the isodiscrimination curve intersects the line of preference neutrality. Panel (b) repeats this exercise for the continuum of preference



(a) Preference + perceived average productivity pairs; $(\hat{\tau}_F, \hat{\eta}_F) = (.5, 2)$

(b) Preference + perceived concentration of productivity pairs; $(\hat{\mu}_F, \hat{\eta}_F) = (5, 2)$

(c) Preference + perceived signal precision pairs; $(\hat{\mu}_F, \hat{\tau}_F) = (5, .5)$

Figure 2. Equivalent Sources

Isodiscrimination curve $(s_M, s_F) = (6.25, 6.75)$; $(u_M, \hat{\mu}_M, \hat{\tau}_M, \hat{\eta}_M) = (6, 5, .5, 2)$; red asterisks denote types with single form of partiality from Proposition 2.

parameters and perceived concentrations of productivity, illustrating the types described in parts (i) and (iii), while panel (c) does so for the continuum of preference parameters and perceived signal precisions, illustrating the types described in parts (i) and (iv).

Given Proposition 2, additional data is necessary in order to identify the source of discrimination. We next explore several methods

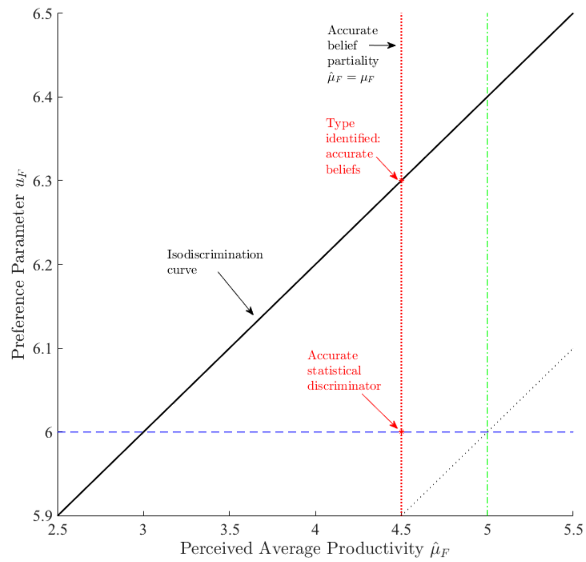


Figure 3. Outcomes-based Test
Isodiscrimination curve $(s_M, s_F) = (6.25, 6.75)$; true distributions $(\mu_M, \tau_M, \eta_M) = (5, .5, 2)$, $(\mu_F, \tau_F, \eta_F) = (4.5, .5, 2)$; $(u_M, \hat{\mu}_M, \hat{\tau}_M, \hat{\eta}_M) = (6, 5, .5, 2)$; blue dotted line: preference neutrality $(u_F = u_M)$; green dotted line: belief neutrality $(\hat{\mu}_F = \hat{\mu}_M)$.

3.2.2 Outcomes-based Test.

A common method to identify the source of discrimination is to compare evaluations to the outcome distribution for each group. In the current framework, this corresponds to comparing hiring decisions to the true productivity and signal distributions.

When implementing an outcomes-based test, the literature typically assumes accurate beliefs. Under this assumption, observing the isodiscrimination curve and the parameters (μ_g, τ_g, η_g) for $g \in \{M, F\}$ for the true distributions identifies the evaluator's type, and therefore, the source(s) of discrimination.[13] Differential treatment and identical true distributions implies preference partiality, whereas different true distributions implies belief partiality (potentially coupled with preference partiality). This is illustrated in Fig. 3: the unique preference parameters that are consistent with the observed isodiscrimination curve and true distributions are $u_F = 6.3$ and $u_M = 6$. Since $\mu_F = 4.5$ and $\mu_M = 5$, this evaluator has both preference partiality and accurate belief partiality.

We first show that the validity of the accurate beliefs assumption is crucial: erroneously imposing this assumption misidentifies the source of discrimination. To formal-

[13]See Lemma 2 in Appendix A for a formal statement of this insight.

ize this insight, we first define how inaccurate beliefs impact the level of discrimination. We say a type's inaccurate beliefs increase (decrease) discrimination if the type discriminates against group F on a strictly larger (smaller) set of signals than the type with accurate beliefs and the same preferences.

Definition 4. *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 decrease is analogous with the inequalities reversed.*

The following result establishes that the researcher mistakenly attributes discrimination stemming from inaccurate belief-based partiality to a 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 3 (Misidentified Source)**.** *Suppose a researcher incorrectly assumes an evaluator has accurate beliefs and uses the outcomes-based method to identify the evaluator's type. 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.*

Fig. 3 illustrates this result. 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 $u_M = 6$. Therefore, the researcher attributes discrimination stemming from the inaccurate belief to preference partiality when in actuality, the evaluator has preference neutrality, $u_F = u_M = 6$. In contrast, if the evaluator believes that the average productivity is the same for both groups, $\hat{\mu}_F = \hat{\mu}_M = 5$ when in fact $\mu_F = 4.5$, then assuming accurate beliefs will lead the researcher to underestimate the evaluator's preference partiality against group F, concluding $u_F = 6.3$ when in fact it is equal to 6.4.

We next show that when one allows for the possibility of inaccurate beliefs, the source is no longer identified: a researcher cannot separate preference partiality from inaccurate belief partiality. When the true distributions are the same, differential treatment may be due to preferences, inaccurate beliefs, or a combination of the two. Similarly, when the true distributions differ, differential treatment may stem from accurate beliefs about these differences or inaccurate beliefs coupled with preference partiality. One source of discrimination that the outcomes-based method *can* potentially rule out is *accurate statistical discrimination*—that is, discrimination stemming from accurate belief partiality and preference neutrality. This source 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).[14] The following result shows that the outcomes-based test can rule out accurate statistical discrimination by showing that the observed pattern of discrimination is inconsistent with accurate beliefs and preference neutrality. Ruling out this source establishes that the observed discrimination either stems from animus towards a group or inaccurate beliefs about them.

Proposition 4 (Rejecting Accurate Statistical Discrimination)**.** *Suppose a researcher identifies the isodiscrimination curve with thresholds* (s_M, s_F) *and observes* (μ_g, τ_g, η_g) *for* $g \in \{M, F\}$*. This does not identify the evaluator's type, but if*

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

then the evaluator is not an accurate statistical discriminator.

In Fig. 3, there are a continuum of evaluator types that exhibit discrimination consistent with the observed isodiscrimination curve, regardless of the true average productivity. However, the preferences and beliefs of an accurate statistical discriminator do not lie on this isodiscrimination curve, and therefore, are not consistent with the observed level of discrimination.

Of course, when the observed pattern of discrimination is consistent with accurate statistical discrimination, this does not identify the evaluator as an accurate statistical discriminator: there are many other types that could also generate the observed behavior. Importantly, even if a combination of preference partiality and inaccurate belief partiality yields equivalent discrimination to an accurate statistical discriminator for the current hiring decision, inaccurate beliefs are not innocuous. These inaccurate beliefs may negatively affect the worker in future performance evaluations and promotions. For example, consider an evaluator with beliefs that overestimate the differences in pro-

[14] 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.

ductivity between groups and preferences that somewhat favor the disadvantaged group for entry-level positions. Suppose this yields equivalent discrimination to an accurate statistical discriminator. Then if the evaluator only feels compelled to favor the disadvantaged group for entry-level hiring, these inaccurate beliefs will lead to persistently lower rates of promotion and advancement for the disadvantaged group.

3.2.3 Eliciting Beliefs.

Given the difficulty of using the outcomes-based method to identify the source of discrimination, we next explore other possible methods. 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.

Proposition 5 (Identifying Preferences from Beliefs)**.** *Suppose a researcher can identify the isodiscrimination curve, i.e. from observing v, g, and s. Observing $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$ identifies u_g, and therefore, the evaluator's type.*

Importantly, observing subjective beliefs does not identify whether they are accurate—additional data, such as outcomes, is necessary to determine this.[15]

In practice, this method will be difficult in certain 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 (for example, 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.

3.2.4 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 set-up, we model this as varying the number of signal draws that the

[15] 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, Exley, and Niederle, 2021). Since the performance distributions are the same regardless of the label, 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.

evaluator observes for a worker. If an evaluator believes that a single draw of the signal has precision $\hat{\eta}_g$, then she believes that observing $x > 1$ draws of this signal has precision $x\hat{\eta}_g$. The characterization of the optimal hiring rule and the isodiscrimination curves for x draws are identical to the case of a single draw, substituting $x\hat{\eta}_g$ for $\hat{\eta}_g$.

This method can be used to partially identify the source of discrimination. The following result establishes that there is a unique set of preference parameters that yield equivalent discrimination across multiple informational treatments. Therefore, manipulating the number of signal draws can identify the level of preference partiality. However, it is not possible to identify the form of belief partiality: given an evaluator of type θ, the result show that there are a continuum of distinct belief profiles that exhibit equivalent discrimination to θ across all informational treatments.

Proposition 6 (Identifying Preferences from Manipulating Information)**.** *Suppose an evaluator of type $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g\in\{M,F\}}$ observes either $x_1 \geq 1$ or $x_2 \neq x_1$ signal draws for each worker. Then the set of types*

$$\left\{ \theta' = (u_g, \hat{\mu}'_g, \hat{\tau}'_g, \hat{\eta}'_g)_{g\in\{M,F\}} \middle| \hat{\mu}'_g = u_g - \frac{\hat{\tau}_g/\hat{\eta}_g}{\hat{\tau}'_g/\hat{\eta}'_g}(u_g - \hat{\mu}_g) \right\} \tag{3}$$

exhibit equivalent discrimination to θ across both informational treatments. This set of types also exhibit equivalent discrimination for any number $x \geq 1$ of signal draws. Therefore, identifying the isodiscrimination curves for at least two informational treatments identifies the evaluator's preferences u_g, but does not identify beliefs $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$.

Fig. 4 illustrates how it is possible to identify a unique preference parameter by varying the number of signal draws. Only a type with preference parameter $u_F = 6.2$ exhibits the observed discrimination for both informational treatments. The other types on the isodiscrimination curve for one draw exhibit equivalent discrimination to the type with $u_F = 6.2$ when there is a single draw. But these types do *not* exhibit equivalent discrimination to the type with $u_F = 6.2$ when there are two signal draws, as they do not lie on the same isodiscrimination curve for two draws. While it may look like the perceived average productivity $\hat{\mu}_F = 4$ is also identified, this is just the belief for a type θ with perceived concentration $\hat{\tau}_F = .5$ and signal precision $\hat{\eta}_F = 1$. There are other types with different values of $(\hat{\tau}_F, \hat{\eta}_F)$ that exhibit equivalent discrimination across all informational treatments, and these types will have different values of $\hat{\mu}_F$ (but the same preference $u_F = 6.2$). For example, from Eq. (3), a type θ' with $u_F = 6.2$, $\tau'_F = .4$,

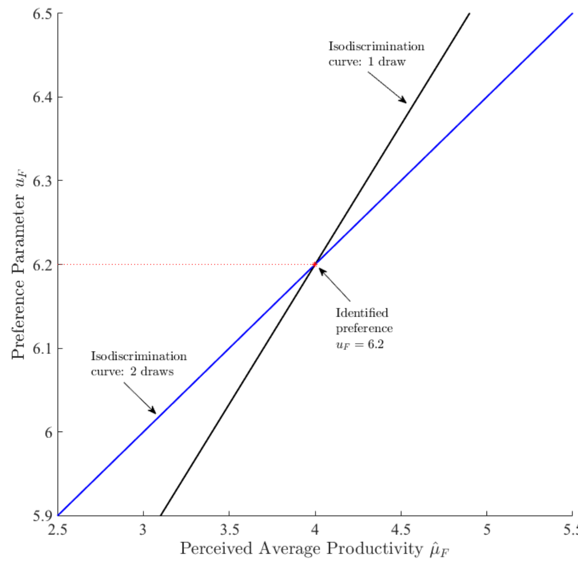


Figure 4. Information Manipulation
Isodiscrimination curves for $(u_F, \hat{\mu}_F, \hat{\tau}_F, \hat{\eta}_F) = (6.2, 4, .5, 1)$

$\eta'_F = 1$ and $\hat{\mu}'_F = 6.2 - \frac{.5/1}{.4/1}(6.2 - 4) = 3.45$ will exhibit equivalent discrimination to θ for any number of signal draws.

A crucial feature of this information manipulation is that the multiple signals are drawn from the same distribution. In other words, for each group, the evaluator has the same model of the precision for each draw of the signal. This contrasts with an information manipulation in which additional signals from other domains are included – for example, comparing discrimination in a treatment in which education is revealed to a treatment in which education and SAT score are revealed. In this case, the evaluator may have a different group-specific model of the signal distribution for each type of signal—for example, the evaluator believes that education is more informative for group M, while SAT score is more informative for group F. It is not possible to identify the evaluator's preference partiality—or any other aspect of her type—from varying the set of observed signals across different domains.

Taken together, either the belief elicitation or information manipulation method can separate preference partiality from 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 can be inaccurate.

4 Identifying the Source

In this section, we employ a stylized experimental setting to demonstrate the pitfalls of the identification problem posed by inaccurate beliefs—in particular, how assuming accurate beliefs can lead to erroneous conclusions—and propose a potential methodology to solve it. The experiment allows us to perform the accounting exercise employed in outcomes-based tests and to also elicit beliefs about relevant characteristics. We show that average beliefs are incorrect, violating the accurate beliefs assumption typically made when considering statistical discrimination, and that ignoring these inaccurate beliefs leads to a false identification of the source of discrimination. We also demonstrate how informational interventions can be used both to separate inaccurate beliefs from underlying animus and to correct these inaccurate beliefs. Specifically, we inform individuals of the true group-specific average productivities. 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.

4.1 Experimental Design

In this section, we provide a short summary of the pre-registered experimental design; we outline the design in detail in Appendix B.[16] We recruited two samples of subjects on Amazon Mechanical Turk (participants) to complete either a work task (Survey 1) or a hiring task (Survey 2). In the first survey, we recruited 589 participants to create a population of "workers" (392 from the US and 197 from India). These participants answered a set of demographic questions, followed by 50 multiple-choice math questions. They were told that their performance would not affect their payment. This design provided for relatively continuous and precise measures of productivity, and allowed us to study discrimination without using deception.

In the second survey, we recruited 577 different participants to create a population of "employers" (392 from the US, 185 from India). These participants were told about the task assigned to "workers" (they were shown five examples of the questions), that the average score was 36.95 out of 50, and that they would serve as an "employer" who could potentially hire one of the workers. The hiring task involved multiple stages. In the first stage, each employer was shown 20 profiles of potential workers, randomly

[16] The experiment was pre-registered on AsPredicted (#8678)

selected from the bank, and made a wage offer for each worker. The profiles included demographic variables which could be informative for forming beliefs about productivity (age, gender, and nationality), or over which employers may hold taste-based motives, as well as potentially irrelevant information (favorite high school subject, sport, color, movie, and coffee/tea preference); see Fig. B1 for an example. One of these workers was selected and their wage offer was accepted or rejected according to the Becker-DeGroot-Marschak mechanism to incentivize truthful reporting. Specifically, if the wage offer was larger than a randomly generated number, then it was 'accepted'—the employer paid the random number as a bonus to the selected worker and the employer received 1 cent for each question answered correctly, minus this random number (in addition to a $2 participation payment); if the wage offer was smaller than the random number, it was 'rejected'—the employer paid nothing and the worker did not receive a bonus.[17]

In the second stage, we elicited employers' beliefs about the average productivity of different groups (men/women, residents from the US/India, people above/below the median age of 33), randomizing whether or not this was incentivized.[18] The third stage involved the informational intervention, which provided employers with average productivity data for each of the six groups. As discussed in Section 1, providing accurate information about group-level statistics is one potential method for differentiating between inaccurate beliefs and "animus-driven" beliefs. While the former should shift in the direction of the information, the latter are unlikely to be moved because the errors are due to non-informational factors. Finally, the fourth stage of the experiment involved a second set of hiring decisions for 10 worker profiles, which was conducted in the same way as the first stage.

4.2 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

[17]Employers saw examples of the mechanism and passed comprehension checks before making wage offers.

[18]We only elicited beliefs about the first moment of the performance distribution. While participants may also have inaccurate beliefs about other statistics, demonstrating a difference in perceived versus actual means is sufficient to falsify the assumption that beliefs are correct, which was the primary goal of the illustrative experiment. Eliciting other moments of the distribution, e.g. variance, is more complex for participants relative to eliciting the mean. Given the multiple stages in the study, we sought to keep the belief elicitation task as simple as possible in order to curtail potential confusion and minimize noise.

Table 4. Wages and "Productivities", by Employee Characteristics (Hiring Task 1)

	(1) Group 1	(2) Group 2	(3) Diff.	(4) p-val	(5) #Obs. G1	(6) #Obs. G2
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 = US , 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 = US , 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)				

Notes: 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.

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.[19]

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 4 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, while 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, while profiles from the U.S. earned 30.71 cents, a significant 7.0% difference ($p < 0.01$). Finally, there was 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 C1 demonstrates that these results are robust to demographic controls and employer fixed effects.

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

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

two indicators to measure in-group bias (see Table C2). 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 C1 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 (since they are taking a "standard model of statistical discrimination" as the benchmark and note that "these beliefs must be correct in equilibrium"). In Table C3, we test 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 a significant difference in beliefs, namely, Indians hold beliefs that more strongly favor the out-group (Americans).

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).[20] In our context, this requires mapping disparities in the employers' stated wages to disparities in group-specific productivity differences, i.e. 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 4 shows the average number of correct answers by each sub-group (see Fig. C1 for probability density functions). As shown in Panel A of Table 4, the gap in average wages for men and women was lower than the gap in average performance (1.05 points versus 3.32 points).[21] Therefore, if we used the standard outcome method

[20]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 Antonovics and Knight (2009); Dharmapala and Ross (2004); Anwar and Fang (2006).

[21]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 (since each of the 589 employers saw independent random samples of 20 of the 577 workers). Due to 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

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 prejudice 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 prejudice against Americans.[22]

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, U.S.: 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 C4, 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 5, we present employer beliefs about the group-specific average performance, which can be compared directly to the actual group-specific performance reported in Table 4, Panel B. 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

the averages in Table B1 are the basis for the informational intervention.

[22]While 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. While the literature often equates taste-based discrimination with animus or prejudice, this link is 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 5. Beliefs about Productivity by Employee Characteristics

	Group 1 (1)	**Group 2** (2)	**Diff.** (3)	**p-val** (4)
Gender (1=Male, 2=Female)	34.04	32.14	1.89	0.00
	(8.26)	(8.41)		
Country (1=US, 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)		

Notes: 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). # Observations = 577.

larger than gaps in wage payments. Using employers' actual beliefs to identify the source of discrimination leads to substantially 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 2.17 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.[23]

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 6 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

[23] In Table C5, we show that the differences in beliefs are quite similar after trimming the top and bottom five percent of the distributions of belief differences by each group. Consistent with Fig. C2, 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.

Table 6. Effect of Information: Difference-in-Differences by Hiring Task

	(1)	(2)	(3)	(4)	(5)
	b/se	b/se	b/se	b/se	b/se
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 X Post-Info	-0.64*			-0.89**	-1.01***
	(0.38)			(0.38)	(0.29)
Indian		2.14***		2.01***	2.02***
		(0.41)		(0.43)	(0.38)
Indian X Post-Info		-1.07**		-1.20***	-1.65***
		(0.43)		(0.44)	(0.33)
Over 33			-0.54	0.06	0.29
			(0.39)	(0.39)	(0.35)
Over 33 X Post-Info			0.41	0.12	-0.21
			(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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: 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 30 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 (US), 36.53 (India), 36.84 (Under 33), 36.77 (Over 33), 36.81 (Prefer Coffee), 36.79 (Prefer Tea).

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.[24]

[24]There are several caveats to note 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, Haushofer, and Roth, 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 10 profiles and the

5 Conclusion

The study of discrimination and its motives has a rich history in economics. Separating out statistical and taste-based drivers of discrimination is a useful exercise, but as our survey of the literature illustrates, the empirical literature has thus far relied heavily on the assumption of accurate beliefs. There are many reasons to suspect that beliefs may not always be accurate. This paper formally outlines the identification problem inherent in distinguishing between belief-based and preference-based motives. A stylized experiment is used to highlight the pitfalls of not accounting for inaccurate beliefs when attempting to identify the source of discrimination, and illustrates a potential methodology for improved identification.

The results of the information intervention suggest that identifying inaccurate beliefs may have immediate policy implications for reducing discrimination. However, there are some important caveats to keep in mind when considering how this type of intervention would be implemented outside of the stylized exercise. 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 the underlying outcomes may differ by group; for example, police officers have been shown to be more likely to discount the recorded speed of a white driver than a minority driver (Goncalves and Mello, 2019). To the latter point, there are contexts in which discrimination at (often unobserved) intermediate stages renders final productivity measures unreliable due to behavioral responses. For example, minority pitchers correctly anticipate discrimination by umpires and modify their behavior, resulting in a downward bias for performance measures (Parsons, Sulaeman, Yates, and Hamermesh, 2011). Studies have also documented that bias at intermediate stages can skew final productivity measures among grocery store workers (Glover, Pallais, and Pariente, 2017) and academic economists (Hengel, 2019). It is important to also take into account the underlying psychology of how people will respond to the information. Selection decisions such as hiring are rarely one-dimensional. Drawing attention to a (smaller than expected) productivity gap could correct beliefs, while nonetheless increasing discrimination if it increases the salience of the gap as an input into the hir-

second 10 profiles during the initial task. We do not find evidence for an experience effect (36.86 vs. 36.72; p=.39). While 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).

ing decision. These concerns highlight the need for future tests that operationalize and examine similar 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 that of the gap in pay. This leads to the conclusion that there is preference-based partiality against the group that received higher wages—male workers. While 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 actions (e.g. equalizing wage rates). For example, people may treat groups similarly regardless of actual or believed productivity differences due to fairness concerns. Additionally, there is often 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 group-level attributes, there is evidence that fairness norms may contribute to observed wage compression (e.g. Breza, Kaur, and Shamdasani (2018))

Just as decomposing the nature of belief-based discrimination has implications for policy, the same may be true for preference-based partiality. For example, if the basis for preference-based partiality is animus or prejudice, then a policy that increases contact between groups may reduce disparities (Dobbie and Fryer, 2015; Paluck, Green, and Green, 2018; Rao, 2019). By contrast, if the behavior is instead sanction- or value-oriented, then such interventions will likely have little impact. While 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.

Lastly, our findings speak to the need for continued work such as Bordalo et al. (2016) that may help to identify situations when inaccurate beliefs are especially likely to be prevalent. 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 reduce discrimination as a result.

References

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

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

Albrecht, K., E. Von Essen, J. Parys, and N. Szech (2013): "Updating, self-confidence, and discrimination," *European Economic Review*, 60, 144–169.

Antonovics, K. and B. G. Knight (2009): "A New Look at Racial Profiling: Evidence from the Boston Police Department," *Review of Economics and Statistics*, 91, 163–177.

Anwar, S. and H. Fang (2006): "An alternative test of racial prejudice in motor vehicle searches: Theory and evidence," *American Economic Review*, 96, 127–151.

Arnold, D., W. Dobbie, and C. Yang (2018a): "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 1885–1932.

Arnold, D., W. Dobbie, and C. S. Yang (2018b): "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 133, 1885–1932.

Arrow, K. J. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter and A. Rees, Princeton, NJ: Princeton University Press.

——— (1998): "What has economics to say about racial discrimination?" *Journal of economic perspectives*, 12, 91–100.

Ayres, I. (2002): "Outcome Tests of Racial Disparities in Police Practices," *Justice Research and Policy*, 4, 131–142.

Bar, R. and A. Zussman (2019): "Identity and Bias: Insights from Driving Tests," *Working Paper*, 1–45.

Bartoš, V., M. Bauer, J. Chytilová, and F. Matějka (2016): "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition," *American Economic Review*, 106, 1437–75.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova (2009): "Powerful Women: Does Exposure Reduce Bias?" *The Quarterly Journal of Economics*, 124, 1497–1540.

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

Bertrand, M. and E. Duflo (2017): "Field Experiments on Discrimination," *Handbook of Field Experiments*, 1, 110.

Bertrand, M. and S. Mullainathan (2004): "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 94, 991–1013.

Bohren, J. A., A. Imas, and M. Rosenberg (2019): "The Dynamics of Discrimination: Theory and Evidence," *Working Paper*.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Stereotypes," *Quarterly Journal of Economics*, 1753–1794.

Brehm, J. W. (1966): *A Theory of Psychological Reactance*, New York: Academic Press.

Breza, E., S. Kaur, and Y. Shamdasani (2018): "The morale effects of pay inequality," *Quarterly Journal of Economics*, 133, 611–663.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2018): "Misperceived social norms: Female labor force participation in Saudi Arabia," Tech. rep., National Bureau of Economic Research.

Cameron, S. V. and J. J. Heckman (2001): "The Dynamics of Educational Attainment for Black, Hispanic, and White Males," *Journal of Political Economy*, 109, 455–499.

Charles, K. K. and J. Guryan (2011): "Studying Discrimination: Fundamental Challenges and Recent Progress," *Annual Review of Economics*, 3, 479–511.

Coffman, K. B., C. L. Exley, and M. Niederle (2021): "The role of beliefs in driving gender discrimination," *Management Science*.

De Quidt, J., J. Haushofer, and C. Roth (2018): "Measuring and bounding experimenter demand," *American Economic Review*, 108, 3266–3302.

Dharmapala, D. and S. L. Ross (2004): "Racial bias in motor vehicle searches: Additional theory and evidence," *Contributions to Economic Analysis and Policy*, 3, 89–111.

Dobbie, W. and R. G. Fryer (2015): "The Impact of Youth Service on Future Outcomes: Evidence from Teach for America," *The B.E. Journal of Economic Analysis and Policy*, 15, 1031–1066.

Durlauf, S. N. (2005): "Racial Profiling as a Public Policy Question: Efficiency, Equity, and Ambiguity," *American Economic Review*, 95, 132–136.

Fang, H. and A. Moro (2011): "Theories of statistical discrimination and affirmative action: A survey," in *Handbook of social economics*, Elsevier, vol. 1, 133–200.

Fershtman, C. and U. Gneezy (2001): "Discrimination in a Segmented Society: An Experimental Approach," *Quarterly Journal of Economics*, February, 351–377.

Fiske, S. T. (1998): "Stereotyping, prejudice, and discrimination," *The handbook of social psychology*, 2, 357–411.

Gagnon-Bartsch, T., M. Rabin, and J. Schwartzstein (2018): *Channeled attention and stable errors*, Harvard Business School.

Glover, D., A. Pallais, and W. Pariente (2017): "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores," *Quarterly Journal of Economics*, 1219–1260.

Gneezy, U., M. Niederle, and A. Rustichini (2003): "Performance in competitive Environments: Gender differences," *Quarterly Journal of Economics*, 1049–1074.

Goncalves, F. and S. Mello (2019): "A Few Bad Apples? Racial Bias in Policing," *Working Paper*, 1–79.

Haavelmo, T. (1944): "The Probability Approach in Econometrics," *Supplement to Econometrica*, 12, 1–115.

Hedegaard, M. S. and J.-R. Tyran (2018): "The Price of Prejudice," *American Economic Journal: Applied Economics*, 10, 40–63.

Hengel, E. (2019): "Publishing while female," *Working Paper*, 1–67.

Hilton, J. L. and W. V. Hippel (1996): "Stereotypes," *Annual Review of Psychology*, 47, 237–271.

Hübert, R. and A. T. Little (2020): "A Behavioral Theory of Discrimination in Policing," *Working Paper*.

Jensen, R. (2010): "The (perceived) returns to education and the demand for schooling," *Quarterly Journal of Economics*, 515–548.

Judd, C. M. and B. Park (1993): "Definition and assessment of accuracy in social stereotypes," *Psychological Review*, 100, 109–128.

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

Knox, D. E., W. Lowe, and J. Mummolo (2020): "Administrative Records Mask Racially Biased Policing," *American Political Science Review*, 1–19.

Kravitz, D. A. and J. Platania (1993): "Attitudes and Beliefs About Affirmative Action: Effects of Target and of Respondent Sex and Ethnicity," *Journal of Applied Psychology*, 78, 928–938.

Lepage, L.-P. (2020): "Endogenous Learning and the Persistence of Employer Biases in the Labor Market," Tech. rep., mimeo.

List, J. A. (2004): "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *Quarterly Journal of Economics*, 49–89.

Mobius, M. and T. Rosenblat (2006): "Why Beauty Matters," *American Economic Review*, 96, 222–235.

Norton, M. I. and D. Ariely (2011): "Building a better America-one wealth quintile at a time," *Perspectives on Psychological Science*, 6, 9–12.

Paluck, E. L., S. Green, and D. P. Green (2018): "The Contact Hypothesis Reevaluated," *Behvioural Public Policy*, 1–30.

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

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

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

Rao, G. (2019): "Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools," *American Economic Review*, 109, 774–809.

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

Simoiu, C., S. Corbett-Davies, and S. Goel (2017): "The problem of inframarginality in outcome tests for discrimination," *Annals of Applied Statistics*, 11, 1193–1216.

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

Zimmermann, F. (2019): "The Dynamics of Motivated Beliefs," *Working Paper*.

Appendix A. Proofs from Section 3

Proof of Lemma 1. The normal distribution is the conjugate prior to a normal likelihood function. Therefore, the evaluator's posterior belief about productivity is normally distributed with mean $(\hat{\tau}_g\hat{\mu}_g + \hat{\eta}_g s)/(\hat{\tau}_g + \hat{\eta}_g)$ and variance $1/(\hat{\tau}_g + \hat{\eta}_g)$ and the optimal decision rule is $v(s, g, \theta) = 1$ iff $(\hat{\tau}_g\hat{\mu}_g + \hat{\eta}_g s)/(\hat{\tau}_g + \hat{\eta}_g) \geq u_g$. Rearranging terms yields Eq. (1). □

Proof of Proposition 1. The characterization of the set of types that exhibit equivalent discrimination follows from Eq. (1) and the discussion in the text. If two types do not discriminate, which corresponds to setting the same thresholds for each group, then trivially they exhibit equivalent discrimination, even when they set different signal thresholds from each other. □

Proof of Proposition 2. Consider an evaluator of type $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g\in\{M,F\}}$ who discriminates against group F. This evaluator generates discrimination that lies on isodiscrimination curve $(s_F, s_M) = (\overline{s}(\theta, F), \overline{s}(\theta, M))$. Given that θ discriminates against group F, $s_F > s_M$. It is immediately apparent from Eq. (2) that there are a continuum of other types that exhibit equivalent discrimination. We next construct types with a single form of partiality.

Part (1): Consider a type θ' with belief neutrality, $(\hat{\mu}'_F, \hat{\tau}'_F, \hat{\eta}'_F) = (\hat{\mu}'_M, \hat{\tau}'_M, \hat{\eta}'_M)$. Let $(\hat{\mu}', \hat{\tau}', \hat{\eta}')$ denote the type's subjective beliefs for a worker from either group. Given preference parameters (u'_F, u'_M), this type hires members of group g with signals above $\overline{s}(\theta', g) = \left(\frac{\hat{\tau}'+\hat{\eta}'}{\hat{\eta}'}\right) u'_g - \frac{\hat{\tau}'}{\hat{\eta}'}\hat{\mu}'$. This type exhibits equivalent discrimination to θ if $\overline{s}(\theta', g) = s_g$ for each $g \in \{M, F\}$. Rearranging terms, this corresponds to preference parameter

$$u'_g = \left(\frac{\hat{\eta}'}{\hat{\tau}' + \hat{\eta}'}\right) s_g + \frac{\hat{\tau}'}{\hat{\tau}' + \hat{\eta}'}\hat{\mu}'$$

for group g. Note $u'_F > u'_M$ since $s_F > s_M$, so there is preference partiality against group F.

Part (2): Consider a type θ' with preference neutrality, $u'_F = u'_M$ and belief neutrality with respect to concentration and signal precision, $(\hat{\tau}'_F, \hat{\eta}'_F) = (\hat{\tau}'_M, \hat{\eta}'_M)$. Let $(u', \hat{\tau}', \hat{\eta}')$ denote these common parameters. Given perceived means $(\hat{\mu}'_M, \hat{\mu}'_F)$, this type hires members of group g with signals above $\overline{s}(\theta', g) = \left(\frac{\hat{\tau}'+\hat{\eta}'}{\hat{\eta}'}\right) u' - \frac{\hat{\tau}'}{\hat{\eta}'}\hat{\mu}'_g$. This type exhibits equivalent discrimination to θ if $\overline{s}(\theta', g) = s_g$ for each $g \in \{M, F\}$. Rearranging terms, this corresponds to perceived mean

$$\hat{\mu}'_g = \left(\frac{\hat{\tau}' + \hat{\eta}'}{\hat{\tau}'}\right) u' - \frac{\hat{\eta}'}{\hat{\tau}'} s_g$$

for group g. Note $\hat{\mu}'_F < \hat{\mu}'_M$ since $s_F > s_M$, so there is belief partiality in the form of lower expected productivity against group F.

Part (3): Consider a type θ' with preference neutrality, $u'_F = u'_M$ and belief neutrality with respect to average productivity and signal precision, $(\hat{\mu}'_F, \hat{\eta}'_F) = (\hat{\mu}'_M, \hat{\eta}'_M)$. Let $(u', \hat{\mu}', \hat{\eta}')$ denote these common parameters. Given perceived concentration of productivity $(\hat{\tau}'_M, \hat{\tau}'_F)$, this type hires members of group g with signals above $\overline{s}(\theta', g) = \left(\frac{\hat{\tau}'_g + \hat{\eta}'}{\hat{\eta}'}\right) u' - \frac{\hat{\tau}'_g}{\hat{\eta}'}\hat{\mu}'$. This type exhibits equivalent discrimination to θ if $\overline{s}(\theta', g) = s_g$ for each $g \in \{M, F\}$. Rearranging terms, this corresponds to perceived concentration

$$\hat{\tau}'_g = \hat{\eta}' \left(\frac{s_g - u'}{u' - \hat{\mu}'}\right)$$

for group g. Given $s_F > s_M$, $\hat{\eta}'(s_F - u') > \hat{\eta}'(s_M - u')$. Therefore, whether $\hat{\tau}'_F$ is greater than or less than $\hat{\tau}'_M$ depends on the sign of $u' - \hat{\mu}'$.

If θ' believes the market is lemon-dropping, i.e. $u' - \hat{\mu}' < 0$, then $\hat{\tau}'_F < \hat{\tau}'_M$ and a less concentrated perceived productivity distribution generates the discrimination against group F. The fatter low productivity tail for F relative to M means that a larger share of workers from group F fall below the threshold ex-ante. We also need to check that $\hat{\tau}'_F > 0$ for these to both be a valid precisions. This will be the case for $u' > s_F$, so that the numerator is also negative. In summary, any type with $\hat{\mu}' > s_F$, $u' \in (s_F, \hat{\mu}')$ and $\hat{\tau}'_g = \hat{\eta}' \left(\frac{s_g - u'}{u' - \hat{\mu}'}\right)$ has belief partiality in the form of lower perceived concentration for group F and exhibits equivalent discrimination to θ.

If θ' believes the market is cherry-picking, i.e. $u' - \hat{\mu}' > 0$, then $\hat{\tau}'_F > \hat{\tau}'_M$ and a more concentrated perceived productivity distribution generates the discrimination against group F. The thinner high productivity tail for F relative to M means that a smaller share of workers from group F lie above the threshold ex-ante. We also need to check that $\hat{\tau}'_M > 0$ for these to both be valid precisions. This will be the case for $s_M > u'$, so that the numerator is also positive. In summary, any type with $\hat{\mu}' < s_M$, $u' \in (\hat{\mu}', s_M)$ and $\hat{\tau}'_g = \hat{\eta}' \left(\frac{s_g - u'}{u' - \hat{\mu}'}\right)$ has belief partiality in the form of higher perceived concentration for group F and exhibits equivalent discrimination to θ.

Part (4): Consider a type θ' with preference neutrality, $u'_F = u'_M$ and belief neutrality with respect to average productivity and concentration, $(\hat{\mu}'_F, \hat{\tau}'_F) = (\hat{\mu}'_M, \hat{\tau}'_M)$. Let $(u', \hat{\mu}', \hat{\tau}')$ denote these common parameters. Given perceived signal precision $(\hat{\eta}'_M, \hat{\eta}'_F)$, this type hires members of group g with signals above $\overline{s}(\theta', g) = \left(\frac{\hat{\tau}'+\hat{\eta}'_g}{\hat{\eta}'_g}\right) u' - \frac{\hat{\tau}'}{\hat{\eta}'_g}\hat{\mu}'$. This type exhibits equivalent discrimination to θ if $\overline{s}(\theta', g) = s_g$ for each $g \in \{M, F\}$. Rearranging terms, this corresponds to perceived signal precision

$$\hat{\eta}'_g = \hat{\tau}' \left(\frac{u' - \hat{\mu}'}{s_g - u'}\right)$$

for group g. Given $s_F > s_M$, $s_F - u' > s_M - u'$. We need $\hat{\eta}'_g > 0$ for each g in order for these to be valid precisions. This is the case when (i) $u' - \hat{\mu}' < 0$ and $s_F - u' < 0$, which also implies $s_M - u' < 0$, or (ii) $u' - \hat{\mu}' > 0$ and $s_M - u' > 0$, which also implies $s_F - u' > 0$.

First consider case (i). In this case, θ' believes the market is lemon-dropping since $u' < \hat{\mu}'$. Further, $0 > s_F - u' > s_M - u' \Rightarrow 1/(s_M - u') > 1/(s_F - u') \Rightarrow (u' - \hat{\mu}')/(s_M - u') < (u' - \hat{\mu}')/(s_F - u')$. Therefore, $\hat{\eta}'_M < \hat{\eta}'_F$ and a higher perceived signal precision generates the discrimination against group F. In summary, any type with $\hat{\mu}' > s_F$, $u' \in (s_F, \hat{\mu}')$ and $\hat{\eta}'_g = \hat{\tau}' \left(\frac{u' - \hat{\mu}'}{s_g - u'}\right)$ has belief partiality in the form of higher perceived signal precision for group F and exhibits equivalent discrimination to θ.

Next consider case (ii). In this case, θ' believes the market is cherry-picking since $u' > \hat{\mu}'$. Further, $s_F - u' > s_M - u' > 0 \Rightarrow 1/(s_M - u') > 1/(s_F - u') \Rightarrow (u' - \hat{\mu}')/(s_M - u') > (u' - \hat{\mu}')/(s_F - u')$. Therefore, $\hat{\eta}'_M > \hat{\eta}'_F$ and a lower perceived signal precision generates the discrimination against group F. In summary, any type with $\hat{\mu}' < s_F$, $u' \in (\hat{\mu}', s_M)$ and $\hat{\eta}'_g = \hat{\tau}' \left(\frac{u' - \hat{\mu}'}{s_g - u'}\right)$ has belief partiality in the form of lower perceived signal precision for group F and exhibits equivalent discrimination to θ. □

Lemma 2. *Assume an evaluator has accurate beliefs and suppose a researcher can identify the isodiscrimination curve. Observing (μ_g, τ_g, η_g) for $g \in \{M, F\}$ identifies the evaluator's type, and therefore, the source(s) of discrimination.*

Proof of Lemma 2. Suppose the evaluator has type $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g \in \{M,F\}}$. This evaluator exhibits discrimination that lies on isodiscrimination curve $(s_F, s_M) = (\overline{s}(\theta, F), \overline{s}(\theta, M))$. Suppose the researcher identifies the isodiscrimination curve (s_F, s_M) and the true productivity and signal distributions (μ_g, τ_g, η_g) for each group g. Under the assumption of accurate beliefs, i.e. $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g) = (\mu_g, \tau_g, \eta_g)$, solving Eq. (2) for u_g uniquely identifies the preference parameters as

$$u_g = \left(\frac{\eta_g}{\tau_g + \eta_g}\right) s_g + \left(\frac{\tau_g}{\tau_g + \eta_g}\right) \mu_g. \tag{4}$$

Therefore, the evaluator's type is identified. □

Proof of Proposition 3. Given true productivity and signal distributions $(\mu_g, \tau_g, \eta_g)_{g\in\{M,F\}}$, suppose the evaluator has type $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g\in\{M,F\}}$ with inaccurate beliefs, $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g) \neq (\mu_g, \tau_g, \eta_g)$. This evaluator exhibits discrimination that lies on isodiscrimination curve $(s_F, s_M) = (\overline{s}(\theta, F), \overline{s}(\theta, M))$. Suppose a researcher identifies the isodiscrimination curve (s_F, s_M) and the true productivity and signal distributions (μ_g, τ_g, η_g) for each group g. When the researcher assumes belief are accurate, i.e. the evaluator is a type θ' with beliefs $(\hat{\mu}'_g, \hat{\tau}'_g, \hat{\eta}'_g) = (\mu_g, \tau_g, \eta_g)$, then from Lemma 2, the researcher concludes that the evaluator has preference parameter

$$u'_g = \left(\frac{\eta_g}{\tau_g + \eta_g}\right) s_g + \left(\frac{\tau_g}{\tau_g + \eta_g}\right) \mu_g. \tag{5}$$

In contrast, the the true preference parameter satisfies

$$u_g = \left(\frac{\hat{\eta}_g}{\hat{\tau}_g + \hat{\eta}_g}\right) s_g + \left(\frac{\hat{\tau}_g}{\hat{\tau}_g + \hat{\eta}_g}\right) \hat{\mu}_g. \tag{6}$$

When beliefs are inaccurate, this identified preference parameter is equal to the true parameter, $u'_g = u_g$, if and only if

$$\mu_g = \left(\frac{\tau_g + \eta_g}{\tau_g}\right) \left[\left(\frac{\hat{\eta}_g}{\hat{\tau}_g + \hat{\eta}_g}\right) s_g - \left(\frac{\eta_g}{\tau_g + \eta_g}\right) s_g + \left(\frac{\hat{\tau}_g}{\hat{\tau}_g + \hat{\eta}_g}\right) \hat{\mu}_g\right]. \tag{7}$$

Therefore, the preference parameter is misidentified for a generic set of true beliefs (μ_g, τ_g, η_g) and evaluator types $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g\in\{M,F\}}$, $u'_g \neq u_g$.

Let $\theta^* = (u_g, \mu_g, \tau_g, \eta_g)_{g\in\{M,F\}}$ denote the type with accurate beliefs and the same preferences as θ. Suppose type θ's inaccurate beliefs increase discrimination against group F, i.e. $\overline{s}(\theta, F) \geq \overline{s}(\theta^*, F)$ and $\overline{s}(\theta^*, M) \geq \overline{s}(\theta, M)$ with at least one strict inequality. Then given the observed isodiscrimination curve is consistent with type θ, i.e. $s_F = \overline{s}(\theta, F)$, $s_F \geq \overline{s}(\theta^*, F) = \frac{\tau_F + \eta_F}{\eta_F} u_F - \frac{\tau_F}{\eta_F} \mu_F$. Combining this inequality with Eq. (5) establishes that

$$u'_F = \left(\frac{\eta_F}{\tau_F + \eta_F}\right) s_F + \left(\frac{\tau_F}{\tau_F + \eta_F}\right) \mu_F \geq u_F. \tag{8}$$

Similarly, $u'_M \leq u_M$, with a strict inequality for at least one of the expressions. Therefore, the researcher overestimates the preference parameter for group F and/or underestimates the preference parameter for group M, leading her to overestimate the preference partiality against group F. The proof for the case of decreasing discrimination is analogous. □

Proof of Proposition 4. Suppose the researcher identifies the isodiscrimination curve (s_F, s_M) and the true productivity and signal distributions (μ_g, τ_g, η_g) for each group g. From Eq. (2), for any $u \in \mathbb{R}$, the corresponding accurate statistical discriminator with preferences $u_M = u_F = u$ lies on isodiscrimination curve (s'_F, s'_M) with $s'_g = \left(\frac{\tau_g+\eta_g}{\eta_g}\right) u - \frac{\tau_g}{\eta_g}\mu_g$. If $\frac{\tau_M \mu_M + \eta_M s_M}{\tau_M + \eta_M} \neq \frac{\tau_F \mu_F + \eta_F s_F}{\tau_F + \eta_F}$, then there is no u such that $(s'_F, s'_M) = (s_F, s_M)$, i.e. an accurate statistical discriminator exhibits discrimination that is consistent with the observed isodiscrimination curve.

Suppose the evaluator has type $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g \in \{M,F\}}$ such that $\overline{s}(\theta, F), \overline{s}(\theta, M)) = (s_F, s_M)$. From Eq. (2), it is clear that when beliefs may be inaccurate, this provides no additional information about $(u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$ for each group g—any type that satisfies Eq. (2) for the observed isodiscrimination curve can exhibit the observed behavior. □

Proof of Proposition 5. Suppose the evaluator has type $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g \in \{M,F\}}$. This evaluator exhibits discrimination that lies on isodiscrimination curve $(s_F, s_M) = (\overline{s}(\theta, F), \overline{s}(\theta, M))$. Suppose the researcher identifies the isodiscrimination curve (s_F, s_M) and the perceived productivity and signal distributions $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$ for each group g. Solving Eq. (2) for u_g uniquely identifies the preference parameters (u_F, u_M) as

$$u_g = \left(\frac{\hat{\eta}_g}{\hat{\tau}_g + \hat{\eta}_g}\right) s_g + \left(\frac{\hat{\tau}_g}{\hat{\tau}_g + \hat{\eta}_g}\right) \hat{\mu}_g. \tag{9}$$

Therefore, the evaluator's type is identified. □

Proof of Proposition 6. Given a signal with precision $\eta > 0$, observing $x \geq 1$ draws of the signal is equivalent to observing a single signal that is normally distributed with precision $x\eta$. Suppose the evaluator has type $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g \in \{M,F\}}$. From the signal thresholds in Eq. (1), if a type $\theta' = (u'_g, \hat{\mu}'_g, \hat{\tau}'_g, \hat{\eta}'_g)_{g \in \{M,F\}}$ exhibits equivalent discrimination to θ when observing $x \geq 1$ signal draws, then

$$\frac{\hat{\tau}_g u_g + x\hat{\eta}_g u_g - \hat{\tau}_g \hat{\mu}_g}{x\hat{\eta}_g} = \frac{\hat{\tau}'_g u'_g + x\hat{\eta}'_g u'_g - \hat{\tau}'_g \hat{\mu}'_g}{x\hat{\eta}'_g} \tag{10}$$

for $g \in \{M, F\}$. Rearranging terms, this is equivalent to

$$\frac{\hat{\tau}_g(u_g - \hat{\mu}_g)}{\hat{\eta}_g} - \frac{\hat{\tau}'_g(u'_g - \hat{\mu}'_g)}{\hat{\eta}'_g} = x(u'_g - u_g). \tag{11}$$

Suppose θ' exhibits equivalent discrimination to θ when observing $x_1 \geq 1$ and $x_2 > x_1$ signal draws. Then Eq. (11) must be simultaneously satisfied at $x = x_1$ and $x = x_2$. Since the left hand side of Eq. (11) is independent of x, this requires $x_1(u'_g - u_g) = x_2(u'_g - u_g)$. Given $x_1 \neq x_2$, it must be that $u'_g = u_g$. Therefore, all types that exhibit equivalent discrimination to θ in both informational treatments have the same preferences as θ, $u'_F = u_F$ and $u'_M = u_M$. The right hand side of Eq. (11) is equal to zero for these types. Therefore, all types that exhibit equivalent discrimination to θ in both informational treatments have beliefs and preferences that satisfy $\hat{\tau}_g(u_g - \hat{\mu}_g)/\hat{\eta}_g = \hat{\tau}'_g(u_g - \hat{\mu}'_g)/\hat{\eta}'_g$. Solving for $\hat{\mu}'_g$, the set of types that exhibit equivalent discrimination to θ corresponds to Eq. (3). If $\hat{\mu}'_g = \hat{\mu}_g$, this corresponds to the set of types that preserve the ratio of the precisions. If the perceived precisions are equal across groups, i.e. $\hat{\tau}'_g = \hat{\tau}_g$ and $\hat{\eta}'_g = \hat{\eta}_g$, then it must be that $\hat{\mu}'_g = \hat{\mu}_g$, which means $\theta' = \theta$. In this case, θ is identified from observing discrimination for two informational treatments.

Suppose θ' exhibits equivalent discrimination to θ for two informational treatments, denoted x_1 and x_2. Then $u'_F = u_F$ and $u'_M = u_M$, so Eq. (11) is satisfied for any other informational treatment $x \notin \{x_1, x_2\}$, i.e. θ' also exhibits equivalent discrimination for informational treatment x. Therefore, if type θ' exhibits equivalent discrimination to θ for two informational treatments, θ' also exhibits equivalent discrimination to θ for all possible informational treatments. □

Given the evaluator has type $\theta = (u_g, \hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)_{g\in\{M,F\}}$, let $\theta(x) = (u_g, \hat{\mu}_g, \hat{\tau}_g, x\hat{\eta}_g)_{g\in\{M,F\}}$ denote a type with the same preferences and beliefs about the productivity distribution as θ and perceived signal precision $x\hat{\eta}_g$. When type θ observes x draws of the signal, it behaves as if it is type $\theta(x)$. Therefore, it exhibits discrimination that lies on isodiscrimination curve $(\overline{s}(\theta(x), F), \overline{s}(\theta(x), M))$. Suppose the researcher identifies the isodiscrimination curves for θ when it observes x_1 and $x_2 \neq x_1$ signal draws, $(s_{F,1}, s_{M,1}) \equiv (\overline{s}(\theta(x_1), F), \overline{s}(\theta(x_1), M))$ and $(s_{F,2}, s_{M,2}) \equiv (\overline{s}(\theta(x_2), F), \overline{s}(\theta(x_2), M))$.

Then from Eq. (2), we know

$$\frac{\hat{\tau}_g + x_i\hat{\eta}_g}{x_i\hat{\eta}_g}u_g - \frac{\hat{\tau}_g}{x_i\hat{\eta}_g}\hat{\mu}_g = s_{g,i}. \tag{12}$$

for $i = 1, 2$. Rearranging terms,

$$\left(\frac{\hat{\tau}_g}{\hat{\eta}_g} + x_i\right)u_g = x_i s_{g,i} + \frac{\hat{\tau}_g}{\hat{\eta}_g}\hat{\mu}_g. \tag{13}$$

Subtracting Eq. (13) evaluated at x_2 from Eq. (13) evaluated at x_1 yields $(x_1 - x_2)u_g = x_1 s_{g,1} - x_2 s_{g,2}$. Solving for u_g uniquely identifies the evaluator's preferences,

$$u_g = \frac{x_1 s_{g,1} - x_2 s_{g,2}}{x_1 - x_2}. \tag{14}$$

However, as shown in Proposition 6, multiple sets of beliefs $(\hat{\mu}_g, \hat{\tau}_g, \hat{\tau}_g)$ can satisfy Eq. (13) for x_1 and x_2. Therefore, the evaluator's beliefs are not identified. □

Appendix B. Experimental Design

Our experimental design includes two separate, pre-registered surveys: (1) a work task (math quiz) performed by 589 Amazon Mechanical Turk subjects (MTurkers), who comprise the prospective "workers" for the second survey, (2) a hiring task in which each of 577 different MTurkers, who comprise the "employers," stated a wage (willingness to pay) for 20 prospective worker profiles.[25] The second survey also contains a belief elicitation and an information intervention followed by a second hiring task. The full surveys are in the Supplemental Material. We describe the experimental design and provide summary statistics below.

Survey 1 (Work Task): We recruited 589 subjects from MTurk on February 23, 2018 for the first survey.[26] The survey was posted with the title "Math Questions and Demographics" and the description "A 20-minute task of answering math questions." We

[25] We pre-registered the study on AsPredicted.org. There are two minor differences between the pre-registration plan and the actual study. First, we pre-registered that we would recruit 400 employers in the hiring task survey, but decided to recruit closer to 600. Second, we did not pre-register sample restrictions due to completing the task too quickly or slowly. We dropped 12 subjects in the work task survey and 5 in the hiring task survey due to these restrictions.

[26] We received 604 responses in total, but dropped 12 responses that corresponded to the top 1% (< 227 seconds) and bottom 1% (> 3274 seconds) in terms of survey duration. Of the remaining 592 responses, we dropped 3 whose Qualtrics survey responses could not be matched to their MTurk records, leaving 589 final respondents.

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.[27] After starting the survey, subjects were informed that they would first answer demographic questions and then answer 50 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 vs. 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 irrelevant information (e.g. tea preference). The irrelevant information ensures that the relevant information is not the only salient information provided to the employer (this mimics the irrelevant 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.

Survey 2: We recruited 577 different MTurk subjects on February 26, 2018. We used the same hiring criteria as the first survey (392 from U.S., 185 from India, >= 80%

[27] 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 one only eligible for U.S. workers.

approval rate).[28] 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 50 math questions, showed them five examples of the math questions, and told them that on average, participants answered 36.95 out of 50 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 20 profiles of potential workers and state the highest wage (between 0 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 20 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 Fig. B1) 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).[29] Finally, employers were presented

[28] We recruited 587 subjects in total, but dropped 7 whose surveys were completed in under 300 seconds and 3 whose stimuli (the profiles they evaluated) could not be matched to the first survey.

[29] Entering an incorrect an answer would generate a pop-up with "Wrong Answer" and restrict the

with a second wage (15), and answered the same questions. They were then presented with 20 profiles, each randomly selected with replacement from the bank of 589 profiles produced by the first survey.

Figure B1. Example Profile Used in First Hiring Task Description

Country:	United States
Gender:	Female
Age:	63
Favorite High School Subject:	English
Favorite Sport:	Gymnastics
Favorite Color:	Sea Green
Favorite Movie:	Overboard
Prefers Coffee/Tea:	Tea

Belief Elicitation Task: Next, subjects were randomly assigned to one of two different conditions: an incentivized or un-incentivized belief elicitation. Across both conditions, subjects were reminded that the full sample answered 36.95 out of 50 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." 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 & 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 U.S. (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). 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

individual from moving to the next page.

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 10 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 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 6 demographic groups used in the second survey. On average, the work task (survey 1) took subjects 19 minutes to complete, while the hiring task took 23 minutes. There is variation in this timing across groups. Subjects from the U.S. took an average of 19 minutes to complete the hiring task, while 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 American 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). While we primarily focus on simple comparisons between each demographic group, these observed differences motivate our use of multivariate regressions as well.

Table B1. Summary Statistics

	Total	**Male**	**Female**	**US**	**India**	**Under 33**	**Over 33**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Worker							
Trivia Score	36.95	38.32	35.28	37.14	36.58	37.10	36.79
	(8.73)	(8.52)	(8.70)	(8.93)	(8.31)	(8.55)	(8.94)
Survey Duration (Minutes)	18.82	19.03	18.56	16.19	24.04	20.25	17.18
	(10.39)	(10.52)	(10.25)	(8.12)	(12.31)	(11.82)	(8.20)
Prefer Tea (Yes=1)	0.39	0.38	0.41	0.37	0.44	0.42	0.36
	(0.49)	(0.49)	(0.49)	(0.48)	(0.50)	(0.49)	(0.48)
Age (Worker)	35.89	35.30	36.62	38.55	30.61	27.38	45.61
	(11.57)	(11.27)	(11.91)	(12.16)	(8.01)	(3.50)	(9.76)
Female (Yes=1)	0.45	0.00	1.00	0.52	0.32	0.43	0.48
	(0.50)	(0.00)	(0.00)	(0.50)	(0.47)	(0.50)	(0.50)
From India (Yes=1)	0.33	0.42	0.23	0.00	1.00	0.47	0.18
	(0.47)	(0.49)	(0.42)	(0.00)	(0.00)	(0.50)	(0.39)
# Observations	589	324	265	392	197	314	275
Panel B: Employer							
Survey Duration (Minutes)	23.09	23.59	22.37	19.08	31.60	22.53	23.87
	(17.23)	(15.57)	(19.43)	(11.70)	(23.04)	(19.00)	(14.44)
College Education or Above	0.67	0.70	0.62	0.56	0.90	0.67	0.67
	(0.47)	(0.46)	(0.49)	(0.50)	(0.30)	(0.47)	(0.47)
Age (Employer)	34.36	32.66	36.88	35.73	31.46	27.09	44.36
	(11.02)	(9.92)	(12.07)	(11.63)	(8.96)	(3.59)	(9.91)
Female (Yes=1)	0.40	0.00	1.00	0.49	0.23	0.34	0.49
	(0.49)	(0.00)	(0.00)	(0.50)	(0.42)	(0.47)	(0.50)
From India (Yes=1)	0.32	0.41	0.19	0.00	1.00	0.40	0.29
	(0.47)	(0.49)	(0.39)	(0.00)	(0.00)	(0.49)	(0.41)
# Observations	577	344	233	392	185	334	243

Notes: Standard deviations in parentheses. One observation per worker (survey 1) or employer (survey 2).

Appendix C. Additional Tables and Figures

Figure C1. Kernel Densities of Productivities (Trivia Scores) by Group

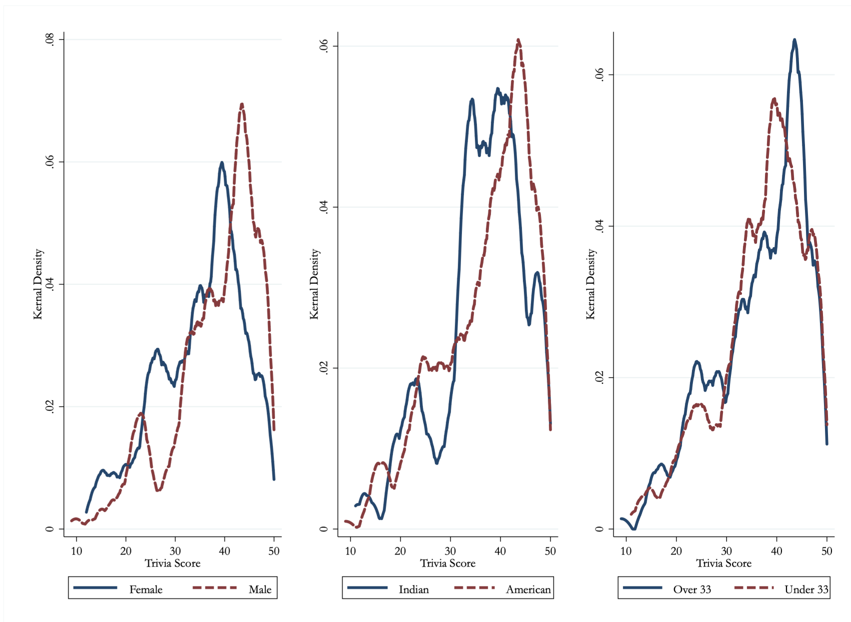


Figure C2. Kernel Densities of Beliefs about Differences by Group (Within-Employer)

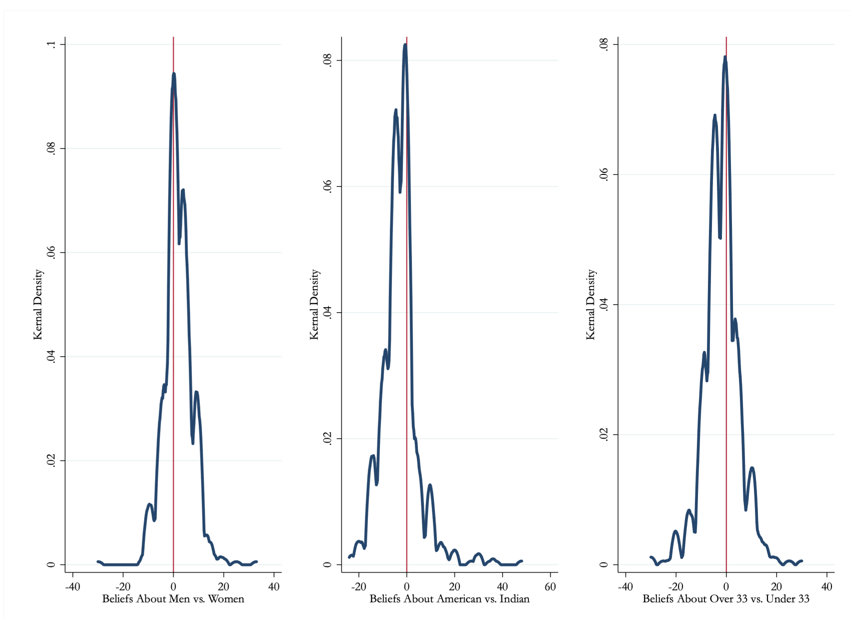


Table C1. Discrimination in Wages, by Employee Characteristics (Hiring Task 1)

	(1)	(2)	(3)	(4)	(5)
	b/se	b/se	b/se	b/se	b/se
Female	-1.05***			-0.66*	-0.78**
	(0.38)			(0.37)	(0.33)
Indian		2.14***		2.01***	2.03***
		(0.41)		(0.43)	(0.38)
Over 33			-0.54	0.06	0.31
			(0.39)	(0.39)	(0.35)
N	11,540	11,540	11,540	11,540	11,540
R^2	0.00	0.01	0.00	0.01	0.49
DepVarMean	31.90	30.71	31.67	30.71	30.71
Employer FE?	No	No	No	No	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: 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 for column (1)).

Table C2. In-Group Bias Test (Hiring Task 1)

	(1)	(2)	(3)	(4)
	b/se	b/se	b/se	b/se
Female Worker	-1.42***			-1.20***
	(0.37)			(0.37)
Female Employer	1.78**			1.91***
	(0.69)			(0.72)
Female Worker X Employer	0.26			0.41
	(0.44)			(0.44)
Indian Worker		2.04***		1.88***
		(0.44)		(0.45)
Indian Employer		0.99		1.70**
		(0.71)		(0.75)
Indian Worker X Employer		-0.79		-0.82
		(0.51)		(0.51)
Over 33 Worker			-0.86**	-0.39
			(0.37)	(0.37)
Over 33 Employer			0.31	0.22
			(0.69)	(0.71)
Over 33 Worker X Employer			1.10***	1.19***
			(0.40)	(0.40)
N	17,310	17,310	17,310	17,310
R^2	0.01	0.01	0.00	0.02
DepVarMean	31.90	30.71	31.67	31.67

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: 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 evaluated by Male Employers for column (1)).

Table C3. In-Group vs. Out-Group Beliefs about Productivity by Employee Characteristics

	Out Group	**In Group**	**Diff.**	**p-val**	**#Obs. Out**	**#Obs. In**
	(1)	(2)	(3)	(4)	(5)	(6)
Prediction for Female Workers	31.70	32.79	-1.09	0.13	344	233
	(8.78)	(7.81)				
Prediction for Male Workers	34.68	33.60	1.09	0.12	233	344
	(6.59)	(9.20)				
Prediction for Indian Workers	36.09	32.06	4.04	0.00	392	185
	(7.10)	(12.67)				
Prediction for US Workers	30.46	32.84	-2.38	0.00	185	392
	(12.04)	(6.15)				
Prediction for Over 33 Workers	30.92	32.47	-1.55	0.04	334	243
	(9.82)	(7.66)				
Prediction for Under 33 Workers	33.85	33.09	0.77	0.31	243	334
	(7.03)	(10.14)				

Notes: Standard deviations in parentheses. "In-Group" refers to a match in the characteristic between the employer and the group of workers over which they are making a prediction, e.g. column (1), row 1 is the average prediction made by female employers about the average productivity of female workers.

Table C5. Beliefs about Productivity by Employee Characteristics, Trimmed

	(1)	(2)	(3)	(4)
	Group (1 or 2)		Diff.	P-Val
	1	2	[(1)-(2)]	
Gender (1 = Male , 2 = Female)	34.26	32.30	1.96	0.00
	(8.23)	(8.20)		
Country (1 = US , 2 = India)	32.00	35.21	-3.22	0.00
	(8.49)	(8.85)		
Age (1 = Under 33 , 2 = Over 33)	33.42	31.78	1.64	0.00
	(8.84)	(8.83)		

Notes: This table repeats Table 5 after trimming the top and bottom 5 percent of observations by the within-employer difference in beliefs about the two groups (e.g. on the Male - Female difference for the first row). Standard deviations in parentheses. One observation per employer combination. Column (4) shows the p-value from regression of the outcome on a dummy variable for group membership, with standard errors two-way clustered by employer and worker. # Observations = 528 (Gender), 541 (Country), and 528 (Age).

Table C4. Effects of Large Incentives for Accurate Predictions

	Incentivized?		**Diff.**	**p-val**
	No	*Yes*		
	(1)	(2)	(3)	(4)
Prediction for Female Workers	32.36	31.93	0.44	0.53
	(7.71)	(9.08)		
Prediction for Male Workers	34.22	33.86	0.36	0.60
	(7.37)	(9.08)		
Prediction for Indian Workers	35.29	34.31	0.98	0.21
	(8.49)	(10.30)		
Prediction for US Workers	32.28	31.87	0.41	0.56
	(8.21)	(8.90)		
Prediction for Over 33 Workers	31.95	31.19	0.75	0.32
	(8.39)	(9.58)		
Prediction for Under 33 Workers	33.73	33.09	0.64	0.39
	(8.58)	(9.35)		
# Observations	290	287		

Notes: Standard deviations in parentheses. One observation per employer. The joint f-statistic from regression of an indicator for the "Incentivized" treatment on set of employer observable characteristics in Table B1, Panel B (duration, education, age, female, from India) is 1.25 (p=0.286).