Inaccurate Statistical Discrimination

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

JUNE 2019
INACCURATE STATISTICAL DISCRIMINATION

J. Aislinn Bohren
Kareem Haggag
Alex Imas
Devin G. Pope

June 2019

We thank Emir Kamenica, Alex Frankel, and seminar participants at Harvard Business School, Stanford University, UCLA, University of Chicago, University of Melbourne, University of Pennsylvania, University of Sydney, and University of Virginia for helpful comments and suggestions. Jihong Song provided excellent research assistance.

© 2019 by J. Aislinn Bohren, Kareem Haggag, Alex Imas, and Devin G. Pope. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.
ABSTRACT

Discrimination has been widely studied in economics and other disciplines. In addition to identifying evidence of discrimination, economists often categorize the source of discrimination as either taste-based or statistical. Categorizing discrimination in this way can be valuable for policy design and welfare analysis. We argue that a further categorization is important and needed. Specifically, in many situations economic agents may have inaccurate beliefs about the expected productivity or performance of a social group. This motivates our proposed distinction between accurate (based on correct beliefs) and inaccurate (based on incorrect beliefs) statistical discrimination. We do a thorough review of the discrimination literature and argue that this distinction is rarely discussed. Using an online experiment, we illustrate how to identify accurate versus inaccurate statistical discrimination. We show that ignoring this distinction – as is often the case in the discrimination literature – can lead to erroneous interpretations of the motives and implications of discriminatory behavior. In particular, when not explicitly accounted for, inaccurate statistical discrimination can be mistaken for taste-based discrimination, accurate statistical discrimination, or a combination of the two.
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. In addition to observational data analysis, a variety of empirical techniques have been used to document discrimination, including audit studies, own-group bias, and correspondence studies (for recent reviews of the discrimination literature, see Bertrand and Duflo 2017 and Charles and Guryan 2011).

In addition to establishing the existence of discrimination, economists typically categorize discrimination as one of two types. The first type, taste-based discrimination (Becker 1957), posits that an individual or firm has animus towards members of a particular group, and therefore may choose to discriminate against them because he receives disutility from providing services to or interacting with members of the group. The second type, statistical discrimination, suggests that discrimination may occur against members of a particular group because productivity is unobserved and the group is perceived to have a lower average productivity (Phelps 1972; Arrow 1973) or a different variance (Aigner and Cain 1977).

Distinguishing between these two types 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/efficiency analyses differ for each type of discrimination. For example, statistical discrimination is sometimes referred to as “efficient discrimination,” in that it is the optimal response to a signal-extraction problem. Importantly, this premise of efficiency relies on the assumption that statistical discrimination stems from rational expectations, or correct beliefs, about the group distributions of the relevant outcome. Finally, the extent to which competitive markets will eliminate discrimination depends on its source.

In this paper, we argue that in many situations, an individual’s beliefs about the productivity of different social groups may be inaccurate. We refer to discrimination that stems from inaccurate beliefs as inaccurate statistical discrimination. Just as it is important to distinguish between taste-based and statistical discrimination for policy design and welfare analysis, we show that it is equally important to separate inaccurate from accurate statistical discrimination. For example, if discrimination stems from inaccurate beliefs, an effective policy response could be to provide individuals with information about the correct distributions (as Jensen 2010 did in the case of inaccurate beliefs about the returns to education).

We discuss two broad sources for inaccurate beliefs. First, psychological biases and heuris-
tics may lead to inaccurate beliefs. Psychologists have long explored and tried to understand why inaccurate stereotypes exist (Schneider et al. 1979; Judd and Park 1993, Hilton and Hippel 1996). Bordalo et al. (2016) provide a model for inaccurate stereotype formation based on the representativeness heuristic. Biased beliefs can also form and persist in a dynamic social learning setting when individuals have incorrect models of how others evaluate workers (e.g. Bohren et al. 2019). Second, inaccurate beliefs may simply be due to a lack of information. A completely rational actor may lack the relevant information necessary to form correct beliefs. For example, an employer may have 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. Thus, an employer may not hire members of a particular group because he/she has an inaccurate belief due to limited information about the selection process. Over time, learning will mitigate inaccurate beliefs in some settings. But in other situations, there will be little or no feedback on the decisions being made, leading to learning traps in which inaccurate beliefs persist. Whether due to psychological bias or lack of information, 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).

By not allowing for the possibility of inaccurate beliefs, we demonstrate that the discrimination literature may mistakenly interpret evidence of inaccurate statistical discrimination as taste-based and/or accurate statistical discrimination. For example, a paper that measures productivity outcomes may find that discrimination against a particular group was not statistically warranted, as different groups had identical distributions of productivity, and therefore argue that it must be taste-based. However, an alternative explanation is that decision-makers had inaccurate beliefs and were inaccurately statistically discriminating, rather than discriminating due to animus.

To understand the prevalence of these mistaken interpretations, we review the empirical literature on discrimination. We find 105 papers published in the top economics journals between 1990 and 2018 that test for evidence of discrimination. Of these papers, 61.9% of them discuss statistical vs. taste-based discrimination and 46.7% of them test between the sources. Only 10.5% of them discuss the possibility of inaccurate beliefs, and 4.8% test for inaccurate beliefs. In nearly all of these cases, we argue that inaccurate statistical discrimination is a reasonable alternative to the interpretation chosen by the authors.\footnote{The line between inaccurate beliefs and taste-based discrimination may sometimes become a bit blurry.}
To demonstrate how inaccurate beliefs influence patterns of discrimination, we run an online hiring experiment. We first have 589 participants (the “workers”) take a 50-question math quiz. These workers differ in their country of origin (US vs. India), gender, and age. A different set of 577 participants (the “employers”) are shown the profiles of 20 participants who took the math quiz and are asked the maximum they would be willing to pay in order to hire each participant (using a Becker-DeGroot-Marschak mechanism). If an employer hires a worker, the worker earns the wage and the employer is paid proportional to how many math questions the worker answered correctly.

Our first finding is that employers discriminate. Americans and females receive systematically lower wage offers than Indians and males (we find no statistically significant difference in wages offered to workers of different ages). Given the conventional categories of discrimination, this finding has two potential explanations. Employers may have animus towards Americans/females relative to Indians/males, and offer them lower wages because they do not want them to receive money. Alternatively, employers may statistically discriminate and offer Americans/females a lower wage because they believe that these groups have lower expected productivity than Indians/males.

One method often used in the discrimination literature to distinguish between statistical and taste-based discrimination is to measure the productivity of outcomes and compare them across groups. This determines whether the discrimination was “warranted” from a statistical perspective. If the difference in distributions by group can explain the observed level of discrimination, then statistical discrimination is at play, and otherwise, discrimination is taste-based. In our experiment, we find that Americans and Indians perform equally well on the math quiz, while females significantly underperform. Thus, using the standard approach to study discrimination, one would argue that the discrimination we observe against Americans stems from a taste-based source, as it was not warranted from a statistical perspective. Further, because we observe a level of discrimination against women that is too small relative to the large differences in productivity, this approach would also conclude that there is taste-based discrimination against men.

An alternative explanation for the discrimination that we find is that individuals have no animus towards a particular group, but rather, have inaccurate beliefs about how well these

What if individuals develop inaccurate beliefs because they have animus towards members of a particular group? Perhaps at times decision makers justify their animus with beliefs that deep down inside they know are incorrect. We agree that this is likely to be the case in many situations. However, we also think these things are separately identifiable. For example, if decision makers are provided with credible information, those that have inaccurate beliefs should change their behavior. Those whose inaccurate beliefs are merely masking an underlying animus are unlikely to change their behavior.
different groups performed on the math test.\textsuperscript{2} To test for this alternative source, we elicit the beliefs of the employers. We find that employers mistakenly thought that Indians would perform much better on the math test than Americans and that females would only slightly underperform relative to males. Accounting for these inaccurate beliefs turns the prior conclusions on their head. What would have been diagnosed as taste-based discrimination \textit{in favor of} Indians appears to be a combination of inaccurate statistical discrimination and taste-based discrimination \textit{against} Indians. Similarly, a large portion of the gender gap is explained by inaccurate statistical discrimination.

As mentioned above, a potential policy intervention to mitigate inaccurate statistical discrimination is to provide individuals with information about how the distribution of an outcome varies by group. We provide the employers in our experiment with information about how the average performance on the math test varied by gender, nationality, and age. After receiving this information, participants were asked to make wage offers to 10 additional potential workers. We find that employers significantly changed their wage offers in the direction of correcting their inaccurate beliefs (less discrimination against Americans and more discrimination against females).

We view our experiment as a tool to illustrate the potential for erroneous conclusions when claiming to identify statistical versus taste-based discrimination without considering the possibility of inaccurate beliefs, rather than a domain for discrimination that is of inherent interest on its own. Our experiment also demonstrates the importance of understanding beliefs in the relevant market when choosing a policy intervention to mitigate discrimination, and the potential for policy interventions to effectively change beliefs.\textsuperscript{3} Building on the nascent literature that discusses the possibility of discrimination due to inaccurate beliefs (see section 2 for a review), our results flesh out the implications for this additional source of discrimination, and provide an illustrative example of how allowing for inaccurate beliefs impacts the interpretation of experimental and observational findings.

The paper proceeds as follows. Section 2 presents a survey of the economics literature on discrimination. Section 3 presents the experimental design, analysis, and interpretation. Section 4 concludes.

\textsuperscript{2}These inaccurate beliefs could be due to a psychological bias or simply a lack of information about how good/motivated people who are recruited to do this online task happen to be at math.

\textsuperscript{3}Though see the end of Section 3 for some strong caveats on porting the information intervention in this experiment to the field.
2 Literature Review

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. 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 of the papers (46.7%) attempted to distinguish between these two sources. However, very few papers discussed the possibility that beliefs may be inaccurate (10.5%), and fewer still examined whether beliefs were accurate or inaccurate (4.8%). Table 1 summarizes these findings.

<table>
<thead>
<tr>
<th>Table 1: Summary of Literature Survey on Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>All: 1990 - 2018</td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td><strong># Papers</strong></td>
</tr>
<tr>
<td>Total Papers</td>
</tr>
<tr>
<td>Evidence of Discrimination</td>
</tr>
<tr>
<td>Discuss taste-based versus statistical source</td>
</tr>
<tr>
<td>Test for taste-based versus statistical source</td>
</tr>
<tr>
<td>Discuss accurate versus inaccurate beliefs</td>
</tr>
<tr>
<td>Test for inaccurate beliefs</td>
</tr>
<tr>
<td>Measure beliefs</td>
</tr>
</tbody>
</table>

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 six papers in this category did not test
whether these elicited beliefs were accurate.

Of the five papers that tested for inaccurate beliefs, three papers did so by directly eliciting beliefs (List 2004, Hedegaard and Tyran 2018, Mobius and Rosenblat 2006) and two did so by either comparing behavior across different contexts (Fershtman and Gneezy 2001) or using theoretical arguments (Arnold et al. 2018). List 2004 finds discrimination against minorities in a marketplace for sports cards. He elicits perceptions of reservation values from a separate group of card dealers and shows that these perceptions qualitatively match the distributional differences observed in the experiment. Experienced dealers are most accurate and exhibit the most discrimination. This suggests that the discrimination observed in that study is statistical. Hedegaard and Tyran (2018) examine the “price of prejudice” using a field experiment where workers can forgo potential earnings in order to work with an in-group member instead of an out-group member. They find evidence for animus workers are willing to forgo 8 percent of earnings to work with a member of the same ethnic group. They elicit productivity beliefs from a separate group of workers and show that these beliefs are qualitatively correct. This suggests that incorrect beliefs are not driving the observed discrimination. Mobius and Rosenblat (2006) was the only paper in our survey that documented inaccurate beliefs. They examine the beauty premium in an experimental principal-agent market and show that more attractive people are offered higher wages but are not more productive. By eliciting beliefs, the authors demonstrate that this discrimination is driven by an incorrect perception of productivity differences.

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 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 following search terms in the title:
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.4

Data Collection. For each paper that met our inclusion criteria, we recorded the following information: data source (laboratory experiment, field experiment, audit 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 about an individuals 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.

---

4We 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 et al. (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

<table>
<thead>
<tr>
<th>Journal</th>
<th>1990-99</th>
<th>2000-09</th>
<th>2010-2018</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AEJ: Applied</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>AEJ: Policy</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>AER</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>EMA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>JEEA</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>JLE</td>
<td>2</td>
<td>8</td>
<td>12</td>
<td>22</td>
</tr>
<tr>
<td>JPE</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>ReStud</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>ReStat</td>
<td>5</td>
<td>6</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>QJE</td>
<td>4</td>
<td>4</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>18</strong></td>
<td><strong>35</strong></td>
<td><strong>52</strong></td>
<td><strong>105</strong></td>
</tr>
</tbody>
</table>

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.

Table 3: Publications by Method

<table>
<thead>
<tr>
<th>Method</th>
<th># Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audit/Correspondence Study</td>
<td>11</td>
</tr>
<tr>
<td>Other Field Experiment</td>
<td>7</td>
</tr>
<tr>
<td>Laboratory Experiment</td>
<td>3</td>
</tr>
<tr>
<td>Observational Data</td>
<td>84</td>
</tr>
</tbody>
</table>

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 4 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 4: Type and Domain of Discrimination

<table>
<thead>
<tr>
<th>Group Identity</th>
<th>Total Papers</th>
<th>Evidence of Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td># Percent</td>
</tr>
<tr>
<td>Race</td>
<td>58</td>
<td>56 96.6%</td>
</tr>
<tr>
<td>Gender</td>
<td>37</td>
<td>35 94.6%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>6</td>
<td>6 100.0%</td>
</tr>
<tr>
<td>Religion</td>
<td>1</td>
<td>1 100.0%</td>
</tr>
<tr>
<td>Sexuality</td>
<td>1</td>
<td>1 100.0%</td>
</tr>
<tr>
<td>Class/Income</td>
<td>1</td>
<td>1 100.0%</td>
</tr>
<tr>
<td>Physical Traits / Appearance</td>
<td>7</td>
<td>7 100.0%</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>5 100.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain of Discrimination</th>
<th>Total Papers</th>
<th>Evidence of Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td># Percent</td>
</tr>
<tr>
<td>Labor Market</td>
<td>58</td>
<td>57 98.3%</td>
</tr>
<tr>
<td>Legal</td>
<td>12</td>
<td>12 100.0%</td>
</tr>
<tr>
<td>Education</td>
<td>9</td>
<td>9 100.0%</td>
</tr>
<tr>
<td>Financial</td>
<td>5</td>
<td>4 80.0%</td>
</tr>
<tr>
<td>Consumer Markets (not financial)</td>
<td>6</td>
<td>6 100.0%</td>
</tr>
<tr>
<td>Other</td>
<td>17</td>
<td>16 94.1%</td>
</tr>
</tbody>
</table>

3 Hiring Experiment

We have argued that it is necessary to elicit beliefs in order to identify the relative importance of (accurate) statistical versus taste-based motives for discrimination. We illustrate this point in a stylized hiring experiment that allows us to perform an accounting exercise typical of the outcomes-based tests common in the literature and to elicit relevant beliefs. We show that actual beliefs are inaccurate, violating the rational expectations assumption typically made in the statistical discrimination literature and altering the conclusions drawn from the exercise, compared to those that would be drawn under rational expectations. We also test a simple intervention to correct inaccurate beliefs: we provide individuals with information on the distributions of productivities by groups and conduct an additional hiring task. This measures the residual importance of taste-based discrimination and demonstrates the effectiveness of this type of informational intervention.

3.1 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.\(^5\) The second survey also contains a belief elicitation and an information intervention followed by a second hiring task. 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.\(^6\) The survey was posted with the title “Math Questions and Demographics” and the description “A 20-minute task of answering math questions.” We paid $2 (i.e. a projected $6/hour wage) and recruited a subject pool of 392 from the United States and 197 from India, all of whom had completed at least 500 prior tasks and had an 80% or higher approval rate for these tasks.\(^7\) 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 \times 6 \times 7 =?”), algebra (e.g. “If \((y + 9) \times \(y^2 - 121\) = 0, then which of the following cannot be \(y\)?”), and more conceptual questions (e.g. “Which of the following is not a prime number?”).\(^8\) 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 corre-

---

\(^5\)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.

\(^6\)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.

\(^7\)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.

\(^8\)The full survey is available in the Online Appendix.
responded 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 serves as a placebo test and 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% approval rate). 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 workers 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.

---

9We 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.
To ensure comprehension, we showed subjects an example profile (see Figure 1) 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). Finally, employers were presented 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 1:** Example Profile Used in First Hiring Task Description

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

---

10Entering an incorrect answer would generate a pop-up with Wrong Answer and restrict the individual from moving to the next page.
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 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 5 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 5: Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
<th>US</th>
<th>India</th>
<th>Under 33</th>
<th>Over 33</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Worker</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trivia Score</td>
<td>36.95</td>
<td>38.32</td>
<td>35.28</td>
<td>37.14</td>
<td>36.58</td>
<td>37.10</td>
<td>36.79</td>
</tr>
<tr>
<td>Survey Duration</td>
<td>18.82</td>
<td>19.03</td>
<td>18.56</td>
<td>16.19</td>
<td>24.04</td>
<td>20.25</td>
<td>17.18</td>
</tr>
<tr>
<td>(Minutes)</td>
<td>(10.39)</td>
<td>(10.52)</td>
<td>(10.25)</td>
<td>(8.118)</td>
<td>(12.31)</td>
<td>(11.82)</td>
<td>(8.199)</td>
</tr>
<tr>
<td>Prefer Tea (Yes = 1)</td>
<td>0.394</td>
<td>0.380</td>
<td>0.411</td>
<td>0.370</td>
<td>0.442</td>
<td>0.424</td>
<td>0.360</td>
</tr>
<tr>
<td></td>
<td>(0.489)</td>
<td>(0.486)</td>
<td>(0.493)</td>
<td>(0.483)</td>
<td>(0.498)</td>
<td>(0.495)</td>
<td>(0.481)</td>
</tr>
<tr>
<td>Age (Worker)</td>
<td>35.89</td>
<td>35.30</td>
<td>36.62</td>
<td>38.55</td>
<td>30.61</td>
<td>27.38</td>
<td>45.61</td>
</tr>
<tr>
<td></td>
<td>(11.57)</td>
<td>(11.27)</td>
<td>(11.91)</td>
<td>(12.16)</td>
<td>(8.014)</td>
<td>(3.503)</td>
<td>(9.762)</td>
</tr>
<tr>
<td>Female (Yes = 1)</td>
<td>0.450</td>
<td>0</td>
<td>1</td>
<td>0.518</td>
<td>0.315</td>
<td>0.427</td>
<td>0.476</td>
</tr>
<tr>
<td></td>
<td>(0.498)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.500)</td>
<td>(0.466)</td>
<td>(0.495)</td>
<td>(0.500)</td>
</tr>
<tr>
<td>From India (Yes = 1)</td>
<td>0.334</td>
<td>0.417</td>
<td>0.234</td>
<td>0</td>
<td>1</td>
<td>0.468</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td>(0.472)</td>
<td>(0.494)</td>
<td>(0.424)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.500)</td>
<td>(0.386)</td>
</tr>
<tr>
<td>N</td>
<td>589</td>
<td>324</td>
<td>265</td>
<td>392</td>
<td>197</td>
<td>314</td>
<td>275</td>
</tr>
<tr>
<td><strong>Panel B: Employer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survey Duration</td>
<td>23.09</td>
<td>23.59</td>
<td>22.37</td>
<td>19.08</td>
<td>31.60</td>
<td>22.53</td>
<td>23.87</td>
</tr>
<tr>
<td>(Minutes)</td>
<td>(17.23)</td>
<td>(15.57)</td>
<td>(19.43)</td>
<td>(11.70)</td>
<td>(23.04)</td>
<td>(19.00)</td>
<td>(14.44)</td>
</tr>
<tr>
<td>College Education</td>
<td>0.669</td>
<td>0.701</td>
<td>0.622</td>
<td>0.559</td>
<td>0.903</td>
<td>0.671</td>
<td>0.667</td>
</tr>
<tr>
<td>or Above</td>
<td>(0.471)</td>
<td>(0.459)</td>
<td>(0.486)</td>
<td>(0.497)</td>
<td>(0.297)</td>
<td>(0.471)</td>
<td>(0.472)</td>
</tr>
<tr>
<td>Age (Employer)</td>
<td>34.36</td>
<td>32.66</td>
<td>36.88</td>
<td>35.73</td>
<td>31.46</td>
<td>27.09</td>
<td>44.36</td>
</tr>
<tr>
<td></td>
<td>(11.02)</td>
<td>(9.917)</td>
<td>(12.07)</td>
<td>(11.63)</td>
<td>(8.962)</td>
<td>(3.587)</td>
<td>(9.906)</td>
</tr>
<tr>
<td>Female (Yes = 1)</td>
<td>0.404</td>
<td>0</td>
<td>1</td>
<td>0.485</td>
<td>0.232</td>
<td>0.341</td>
<td>0.490</td>
</tr>
<tr>
<td></td>
<td>(0.491)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.500)</td>
<td>(0.424)</td>
<td>(0.475)</td>
<td>(0.501)</td>
</tr>
<tr>
<td>From India (Yes = 1)</td>
<td>0.321</td>
<td>0.413</td>
<td>0.185</td>
<td>0</td>
<td>1</td>
<td>0.395</td>
<td>0.218</td>
</tr>
<tr>
<td></td>
<td>(0.467)</td>
<td>(0.493)</td>
<td>(0.389)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0.490)</td>
<td>(0.414)</td>
</tr>
<tr>
<td>N</td>
<td>577</td>
<td>344</td>
<td>233</td>
<td>392</td>
<td>185</td>
<td>334</td>
<td>243</td>
</tr>
</tbody>
</table>

*Notes:* Standard deviations in parentheses. One observation per worker (survey 1) or employer (survey 2).
3.2 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 being observed as part of a research study, the stakes were low, there was no direct interaction or expectation of future interaction between the employers and the workers, 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.\footnote{For example, Bar and Zussman (2019) argue that a lack of interaction may attenuate the extent of taste-based discrimination in driving test examinations.}

Despite these attenuating factors, we did find evidence of discrimination with respect to two out of three group identities: gender and nationality, but not age. Panel A of Table 6 presents the differences in average wages paid by employers to worker profiles from each group. Related to gender, male profiles are paid on average 31.90 cents, while female profiles are paid 30.85 cents, a 3.4% difference that is significant at the 1% level (p<0.01). With respect to nationality, profiles from India are favored, earning an average of 32.85 cents, while profiles from the U.S. earn 30.71 cents, a 7.0% difference also significant at the 1% level (p<0.01). Finally, we see modest evidence of age discrimination: subjects at or below age 33 are paid an average of 31.67 cents and those above age 33 are paid 31.14 cents, a 1.7% difference that is significant at the 5% level (p=0.02). To put these differences into context, we show a “placebo” comparison for a profile characteristic that was unlikely to be either a source of animus or a proxy for math ability: the workers preference for tea versus coffee. We find a similar level of discrimination to the level that we found for age, with tea drinkers earning a 1.7% higher average wage (31.74 vs. 31.22 cents), significant at the 5% level (p=0.03).

In Table 7, we present regression results that simultaneously control for all demographic variables, as the pairwise mean comparisons do not account for any correlations with other demographics. Columns 1-4 repeat the mean comparisons in regression format with standard errors clustered at the employer level. In column 5, we present the multivariate regression with all demographic controls and in column 6, we add employer fixed effects. In both cases, the levels of discrimination for gender and age are slightly diminished in magnitude, but remain statistically significant. To check for robustness, we run similar regressions controlling for the employer belonging to the group of interest (e.g. female) and the interaction of the two indicators to measure in-group bias (see Appendix Table 1). We find that the interaction is
Table 6: Wages and “Productivities”, by Employee Characteristics (Hiring Task 1)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>Diff.</th>
<th>P-Val</th>
<th>N(1)</th>
<th>N(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group (1 or 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>[(1)-(2)]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender (1 = Male, 2 = Female)</td>
<td>31.90</td>
<td>30.85</td>
<td>1.05</td>
<td>0.00</td>
<td>6,306</td>
<td>5,234</td>
</tr>
<tr>
<td></td>
<td>(12.07)</td>
<td>(12.23)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Country (1 = US, 2 = India)</td>
<td>30.71</td>
<td>32.85</td>
<td>-2.14</td>
<td>0.00</td>
<td>7,700</td>
<td>3,840</td>
</tr>
<tr>
<td></td>
<td>(12.20)</td>
<td>(11.95)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (1 = Under 33, 2 = Over 33)</td>
<td>31.67</td>
<td>31.14</td>
<td>0.54</td>
<td>0.02</td>
<td>6,139</td>
<td>5,401</td>
</tr>
<tr>
<td></td>
<td>(12.00)</td>
<td>(12.33)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Placebo (1 = Prefer Coffee, 2 = Prefer Tea)</td>
<td>31.22</td>
<td>31.74</td>
<td>-0.52</td>
<td>0.03</td>
<td>7,075</td>
<td>4,465</td>
</tr>
<tr>
<td></td>
<td>(12.32)</td>
<td>(11.89)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel B: Employee Productivity, by Employee Characteristics

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>Diff.</th>
<th>P-Val</th>
<th>N(1)</th>
<th>N(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group (1 or 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>[(1)-(2)]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender (1 = Male, 2 = Female)</td>
<td>38.30</td>
<td>34.98</td>
<td>3.32</td>
<td>0.00</td>
<td>6,306</td>
<td>5,234</td>
</tr>
<tr>
<td></td>
<td>(8.55)</td>
<td>(8.73)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Country (1 = US, 2 = India)</td>
<td>37.01</td>
<td>36.36</td>
<td>0.65</td>
<td>0.00</td>
<td>7,700</td>
<td>3,840</td>
</tr>
<tr>
<td></td>
<td>(8.93)</td>
<td>(8.49)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (1 = Under 33, 2 = Over 33)</td>
<td>36.96</td>
<td>36.60</td>
<td>0.37</td>
<td>0.03</td>
<td>6,139</td>
<td>5,401</td>
</tr>
<tr>
<td></td>
<td>(8.62)</td>
<td>(8.98)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Placebo (1 = Prefer Coffee, 2 = Prefer Tea)</td>
<td>36.64</td>
<td>37.03</td>
<td>-0.40</td>
<td>0.02</td>
<td>7,075</td>
<td>4,465</td>
</tr>
<tr>
<td></td>
<td>(8.77)</td>
<td>(8.82)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard deviations in parentheses. One observation per worker-employer combination. Column 4 shows the p-value from two-sample t-tests for the equality of columns 1 and 2.

Having demonstrated moderate levels of discrimination in hiring, we turn to measured productivity differences between groups in the work task (the math quiz). The typical outcomes-based test of statistical discrimination requires mapping disparities between groups in the evaluators’ relevant decision (e.g. police searches of motor vehicles) to disparities in an outcome in the evaluators’ objective function (e.g. the likelihood of finding contraband).\(^{12}\)

\(^{12}\)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 ones 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 Appendix Table A3, 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).

\(^{13}\)Translating the two measures may require strong modeling assumptions (e.g. whether there is hetero-
Table 7: Discrimination in Wages, by Employee Characteristics (Hiring Task 1)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
</tr>
<tr>
<td>Female</td>
<td>-1.05***</td>
<td>-0.67***</td>
<td>-0.80***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(0.24)</td>
<td>(0.19)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian</td>
<td>2.14***</td>
<td>1.98***</td>
<td>2.00***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.29)</td>
<td>(0.29)</td>
<td>(0.25)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 33</td>
<td></td>
<td>-0.54**</td>
<td>0.07</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.26)</td>
<td>(0.26)</td>
<td>(0.22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Placebo: Prefers Tea</td>
<td></td>
<td>0.52**</td>
<td>0.39*</td>
<td>0.37**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.24)</td>
<td>(0.24)</td>
<td>(0.18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>11,540</td>
<td>11,540</td>
<td>11,540</td>
<td>11,540</td>
<td>11,540</td>
<td>11,540</td>
</tr>
<tr>
<td>R^2</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.49</td>
</tr>
<tr>
<td>DepVarMean</td>
<td>31.90</td>
<td>30.71</td>
<td>31.67</td>
<td>31.22</td>
<td>30.18</td>
<td>30.18</td>
</tr>
<tr>
<td>Employer FE?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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

Notes: Standard errors clustered by employer. “DepVarMean” is the mean of the dependent variable (wage WTP) in the omitted group (e.g. Male Workers for column 1).

To simplify and make the link between the employers’ objective function and decision transparent, we required employers to state a wage that is directly comparable to the outcome they are maximizing (predicting the number of correct questions by the worker) – that is, each cent for the wage is equivalent to one correct question. Therefore, we can directly compare disparities in wages to disparities in performance to measure the relative importance of (accurate) statistical versus taste-based sources. For this comparison, we will refer to both disparities as measured in “points.”

Figure 2 plots probability density functions for the number of correct answers by each sub-group. Men and subjects from India outperformed women and subjects from the U.S., respectively, whereas there was no significant difference by age. As shown in Table 6, the gap in average wages for men and women was lower than the gap in average performance (1.05 points versus 3.32 points). Therefore, if we used this standard approach to separate

covariates 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.

14 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 5. For example, the male-female performance gap is 3.04 points in Table 5 and 3.32 points in this weighted sample. Note that the averages in Table 5 are the
statistical and taste-based discrimination, we would conclude that the entire 1.05 point disparity in wages is due to (accurate) statistical discrimination. Further, we would conclude that the remaining 2.17 point difference in performance suggests taste-based discrimination against men. Turning to nationality-based discrimination, there was a wage gap of -2.14 points in favor of Indians, compared to a performance gap of 0.65 points in favor of Americans. Under the standard approach, we would conclude that both the -2.14 point disparity in wages and the 0.65 point difference in performance suggest taste-based discrimination against Americans.

Figure 2: Kernal Densities of Productivities (Trivia Scores) by Group

As we argued above, this standard approach naively assumes that employers have accurate beliefs about group-level performance differences. To investigate this assumption, we elicited beliefs about performance. As a check that employers’ decisions were guided by the beliefs we elicited, we correlate the wages they assigned in the prior hiring task with the predictions they made for average productivity of each group. We find modest 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 reflect assimilation driven by cognitive dissonance (e.g. an individual first chooses to discriminate against women in setting wages and then chooses beliefs to justify this decision) or by an audience effect (e.g. an individual falsely reports beliefs that justify the discriminatory decision to the experimenter). To test for these types of cheap talk effects, we provided large incentives for accuracy for half of the employers. In Appendix basis for the informational intervention.
Table 2, we show that beliefs are nearly identical across both incentive conditions, with none of the six comparisons being statistically different. Together these findings suggest that the employers’ group-level performance predictions provide meaningful information about their beliefs.

In Table 8, we present employer beliefs about the group-level average performance (math quiz scores) that can be compared to the actual performance reported in Table 6, 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 a negative payment, taken from their $0.50 bonus, if they overestimate performance). More to our point, gaps in beliefs about performance are larger than gaps in wage payments. Repeating the accounting exercise above using reported beliefs about performance, rather than actual performance (which implicitly assumes correct beliefs) leads to markedly different conclusions. By nationality, the wage gap is -2.14 points and the performance gap is 0.65 points, whereas the belief gap is -2.72 points. Thus, the entire wage gap is explained by inaccurate beliefs. The remaining 0.58 point (2.72 minus 2.14) difference between the belief and wage gaps suggests taste-based discrimination against Indians. This contrasts with the comparison using performance as beliefs, where we concluded that there was taste-based discrimination in favor of Indians. Similarly, we reach different conclusions about the sources of gender discrimination. The wage gap is 1.05 points, the performance gap is 3.32 points, and the belief gap is 1.89 points. The wage gap is still fully accounted for by beliefs, but the gap between performance and wages is partly explained by inaccurate beliefs; the difference explained by a taste-based source shrinks from 2.17 to 0.84 points. Finally, despite the minimal gap in wages and performance for age, employers believed that young workers outperform older ones. This suggests taste-based discrimination against younger workers.

The role of inaccurate beliefs in explaining part of the wage gap suggests a policy intervention: inform employers of the correct distributions and provide them with the opportunity to set wages for a different set of potential workers to see if this induces a behavioral change. To examine the effects of such a policy, we had employers assign wages to 20 profiles (hiring task 1), provided them with information about group-level math quiz score averages, and then had them evaluate 10 more profiles (hiring task 2).

Table 9 demonstrates the effects of this informational intervention. We compare the differences between the two hiring rounds (Post-Info), the differences between wages assigned to profiles of each worker demographic group (e.g. “Female”), and the resulting difference-indifferences (e.g. “Female X Post-Info”). The coefficients on “Post-Info” suggests substantial
### Table 8: Beliefs about Productivity by Employee Characteristics

<table>
<thead>
<tr>
<th>Group (1 or 2)</th>
<th>Diff. [(1)-(2)]</th>
<th>P-Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Gender (1 = Male , 2 = Female)</td>
<td>34.04</td>
<td>32.14</td>
</tr>
<tr>
<td>Country (1 = US , 2 = India)</td>
<td>32.08</td>
<td>34.80</td>
</tr>
<tr>
<td>Age (1 = Under 33 , 2 = Over 33)</td>
<td>33.41</td>
<td>31.57</td>
</tr>
</tbody>
</table>

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. N = 577.

Belief updating across all demographic groups, partially correcting the large level differences in hiring task 1 between wages and actual productivity (a gap of roughly 5 points on average; see Table 6). For example, Column 1 shows that the wage rate for Men increased by 1.53 points. The coefficients on the interaction between “Post-Info” and each demographic group show that the intervention had a significant effect on wage gaps by gender and nationality, but not on the more modest gap by age. The gender wage gap expands by 0.64 points, increasing to 1.71 points in hiring task 2 (Column 1). This gap remains smaller than the 3.01 performance gap between the male and female profiles. Thus the wage gap still reflects a mix of accurate statistical discrimination against women and either (a smaller amount of) taste-based discrimination against men (1.30 points instead of 2.71 points) or significant limits to belief updating in response to our intervention.\(^{15}\) For nationality, the intervention

\(^{15}\)Throughout the paper we find discrimination in wage rates by gender (i.e. men paid more than women). As such, we carry out the accounting exercise that would typically follow in the literature. This reveals that the gap in performance in fact exceeds that of the gap in pay. Thus, we end up concluding that there is taste-based discrimination against the group that received higher wages, e.g. men. While the literature often treats taste-based discrimination as analogous to animus or prejudice against a group, when discrimination works toward equalizing actions (e.g. wage rates), this analogy seems misplaced. For example, in many real world domains, it is illegal to discriminate on the basis of gender, religion, or race, even if those attributes correlate with the outcomes of interest. Thus, employers may fail to discriminate (or not discriminate enough, given their beliefs) against the less productive group, not because they favor them, but because they fear legal or social sanctions. Relatedly, there is often an equity-efficiency trade-off to discriminating, 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 et al. 2018).

Just as decomposing the nature of belief-based discrimination has implications for policy, the same may be true for taste-based discrimination. For example, if the basis for taste-based discrimination is animus or prejudice, then a policy that increases contact between groups may reduce disparities (Dobbie and Fryer 2015; Paluck et al. 2018). By contrast, if the behavior is instead sanction- or value-oriented, then such
cut the wage gap in half, from 2.14 points down to 1.07 (Column 2). This wage gap still exceeds the performance gap of 0.42 points, again suggesting that there remains a modest amount of taste-based discrimination against Americans or limits to belief updating.

**Table 9: Effect of Information: Difference-in-Differences by Hiring Task**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
</tr>
<tr>
<td>Post-Info</td>
<td>1.53***</td>
<td>1.60***</td>
<td>1.06***</td>
<td>1.28***</td>
<td>1.94***</td>
<td>2.39***</td>
</tr>
<tr>
<td></td>
<td>(0.29)</td>
<td>(0.27)</td>
<td>(0.31)</td>
<td>(0.29)</td>
<td>(0.44)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>Female</td>
<td>-1.05***</td>
<td>-0.67***</td>
<td>-0.81***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(0.24)</td>
<td>(0.19)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female X Post-Info</td>
<td>-0.64*</td>
<td>-0.90**</td>
<td>-1.00***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.37)</td>
<td>(0.37)</td>
<td>(0.29)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian</td>
<td>2.14***</td>
<td>1.98***</td>
<td>1.99***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.29)</td>
<td>(0.29)</td>
<td>(0.25)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian X Post-Info</td>
<td>-1.07***</td>
<td>-1.20***</td>
<td>-1.63***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.40)</td>
<td>(0.42)</td>
<td>(0.34)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 33</td>
<td>-0.54**</td>
<td>0.07</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.26)</td>
<td>(0.26)</td>
<td>(0.22)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 33 X Post-Info</td>
<td>0.41</td>
<td>0.14</td>
<td>-0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.40)</td>
<td>(0.42)</td>
<td>(0.30)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prefers Tea</td>
<td>0.52**</td>
<td>0.39*</td>
<td>0.35**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.24)</td>
<td>(0.18)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prefers Tea X Post-Info</td>
<td>-0.08</td>
<td>0.06</td>
<td>-0.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.38)</td>
<td>(0.38)</td>
<td>(0.27)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>17,310</td>
<td>17,310</td>
<td>17,310</td>
<td>17,310</td>
<td>17,310</td>
<td>17,310</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.48</td>
</tr>
<tr>
<td>DepVarMean</td>
<td>31.90</td>
<td>30.71</td>
<td>31.67</td>
<td>31.22</td>
<td>30.18</td>
<td>30.18</td>
</tr>
<tr>
<td>Employer FE?</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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

**Notes:** Standard errors clustered by employer. “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).

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 the model of discrimination used in economics and the tools used to identify discrimination and design policy.
Our interpretation of the second hiring task results is that employers updated their beliefs toward the signal provided by the information intervention. However, a few challenges remain in determining whether this is the causal effect of the intervention through the channel of updated beliefs. First, we did not measure beliefs again after the information intervention, and therefore, we cannot verify that employers’ beliefs changed. However, as noted earlier, the fact that we find a large level effect on wages across all demographic groups is suggestive of belief updating. A second concern is that part of the change in wages could reflect an experience effect between rating the first 20 profiles and the subsequent 10 after the intervention. To investigate this channel, we perform a t-test comparing the average wages assigned in the first 10 profiles and the second 10 profiles in the first hiring task. There is no significant difference between these two sets of profiles (36.86 vs. 36.72; p=.39). A third concern is that the belief elicitation, rather than the information intervention, is what changed behavior (e.g. one possible mechanism is that stating beliefs about productivity over the given profile characteristics changes the salience of these characteristics, and therefore the weight placed on them in hiring). Since we did not have a group that had a belief elicitation but no information intervention, we cannot rule this out. Finally, the information intervention could have produced an experimenter demand effect. Subjects may have inferred that we wanted them to move their wage offers toward the performance levels, and thus done so to please the experimenter without actually changing their beliefs. While we cannot fully rule out all of these possible confounds, the information intervention serves as a proof of concept of the type of policy that could change behavior in the face of inaccurate statistical discrimination.

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 toy exercise in this paper. First, such an intervention may only be appropriate in contexts where the underlying target outcome (e.g. productivity) is reliably measured and reflects the appropriate counterfactual outcomes for all groups were they to be selected by the agents whose beliefs are being corrected (e.g. employers). To the first point, in some cases the underlying measures may be inaccurately recorded; for example, biased police officers are less likely to record a minority driver at a discounted speed than a white 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 downwardly biased performance measures.
(Parsons et al. 2011). Studies have also documented that bias at intermediate stages can skew final productivity measures among grocery store workers (Glover et al. 2017) and academic economists (Hengel 2019).

A second set of caveats relates to the underlying psychology of how people would respond to the information. First, selection decisions 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 and thus the weight placed on that attribute in the selection decision. Second, as noted in footnote 1, the line between inaccurate beliefs and animus is sometimes blurry. For example, individuals may use motivated reasoning to arrive at and preserve closely held inaccurate beliefs. As noted in the footnote, we can separately identify animus-driven inaccurate beliefs through an information experiment if individuals asymmetrically update (i.e. they discard the good news about the out-group); however, such an exercise would be difficult if individuals correctly update in the short run and only exhibit motivated memory (i.e. they discard the good news with a delay, as in Zimmermann 2019). If individuals inaccurate beliefs are in fact driven by animus, then the information intervention should merely be ineffective if it is discarded. However, it is possible that individuals with such animus-driven beliefs would exhibit psychological reactance (Brehm 1966) to a perceived attempt at manipulation and thus exhibit even more biased selections as a way of expressing their agency. Third, there may be spillover effects of a social planner selectively revealing information about gaps. Consider a planner that wishes to minimize discrimination. If such a planner only reveals information about gaps when they are smaller than stereotypes would suggest, then the absence of an information intervention itself reveals information. Its not obvious how individuals would update their beliefs in response to not getting an information summary on other attributes of the individuals under consideration or how it would spillover into other selection tasks. Such updating would likely depend on agents beliefs about the threshold being used by the social planner and the motives of the planner (e.g. whether the planner is willing to allow statistical discrimination or aims to minimize all discrimination). These various concerns limit the external validity of our information intervention, but we look forward to future tests that operationalize it appropriately for field contexts.
4 Conclusion

The study of discrimination and its motives has a rich history in economics. Separating out statistical and taste-based intentions for discrimination is a useful exercise, but as our literature review illustrates, is one that has typically relied heavily on the assumption of accurate beliefs. However, there are many reasons to suspect that beliefs may not always be accurate (both psychological explanations of bias and also due to limited information of rational agents). In this paper, we document and provide an example for the misinterpretation of data that can result from researchers assuming accurate beliefs.

Going forward, we argue that researchers studying motives of discrimination should in all situations consider and discuss the possibility of inaccurate beliefs and elicit beliefs when possible. By doing so, researchers are less likely to mistakenly assign motives of animus or accurate statistical discrimination when those may actually not be the actual intentions. Further, this research speaks to the need for continued work such as Bordalo et al. 2016 that might help 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 improve beliefs and thereby reduce discrimination.

References


## Appendix: Additional Tables

### Table A1: Effects of Large Incentives for Accurate Predictions

<table>
<thead>
<tr>
<th></th>
<th>(1) Incentivized?</th>
<th>(2)</th>
<th>(3) Diff. [(1)-(2)]</th>
<th>(4) P-Val</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Female Workers</td>
<td>32.36</td>
<td>31.93</td>
<td>0.44</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>(7.71)</td>
<td>(9.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Male Workers</td>
<td>34.22</td>
<td>33.86</td>
<td>0.36</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>(7.37)</td>
<td>(9.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Indian Workers</td>
<td>35.29</td>
<td>34.31</td>
<td>0.98</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>(8.49)</td>
<td>(10.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for US Workers</td>
<td>32.28</td>
<td>31.87</td>
<td>0.41</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>(8.21)</td>
<td>(8.90)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Over 33 Workers</td>
<td>31.95</td>
<td>31.19</td>
<td>0.75</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>(8.39)</td>
<td>(9.58)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Under 33 Workers</td>
<td>33.73</td>
<td>33.09</td>
<td>0.64</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>(8.58)</td>
<td>(9.35)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>290</td>
<td>287</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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 observables characteristics in Table 1, Panel B (duration, college education or above, age, female, from India) is 1.25 (p=0.2863).
Table A2: In-Group Bias Test (Hiring Task 1)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
<td>b/se</td>
</tr>
<tr>
<td>Female Worker</td>
<td>-1.42</td>
<td>-1.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(0.27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female Employer</td>
<td>1.78</td>
<td>1.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.69)</td>
<td>(0.72)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female Worker X Employer</td>
<td>0.26</td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.43)</td>
<td>(0.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian Worker</td>
<td>2.04</td>
<td>1.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.31)</td>
<td>(0.31)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian Employer</td>
<td>0.99</td>
<td>1.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.70)</td>
<td>(0.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian Worker X Employer</td>
<td>-0.79</td>
<td>-0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.48)</td>
<td>(0.48)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 33 Worker</td>
<td>-0.86</td>
<td>-0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.29)</td>
<td>(0.28)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 33 Employer</td>
<td>0.31</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.69)</td>
<td>(0.71)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Over 33 Worker X Employer</td>
<td>1.10</td>
<td>1.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.42)</td>
<td>(0.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>17,310</td>
<td>17,310</td>
<td>17,310</td>
<td>17,310</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>DepVarMean</td>
<td>31.90</td>
<td>30.71</td>
<td>31.67</td>
<td>31.67</td>
</tr>
</tbody>
</table>

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

Notes: Standard errors clustered by employer. “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 A3: In-Group vs. Out-Group Beliefs about Productivity by Employee Characteristics

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out vs. In Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diff.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-Val</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Female Workers</td>
<td>31.70</td>
<td>32.79</td>
<td>-1.09</td>
<td>0.13</td>
<td>344</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>(8.78)</td>
<td>(7.81)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Male Workers</td>
<td>34.68</td>
<td>33.60</td>
<td>1.09</td>
<td>0.12</td>
<td>233</td>
<td>344</td>
</tr>
<tr>
<td></td>
<td>(6.59)</td>
<td>(9.20)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Indian Workers</td>
<td>36.09</td>
<td>32.06</td>
<td>4.04</td>
<td>0.00</td>
<td>392</td>
<td>185</td>
</tr>
<tr>
<td></td>
<td>(7.10)</td>
<td>(12.67)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for US Workers</td>
<td>30.46</td>
<td>32.84</td>
<td>-2.38</td>
<td>0.00</td>
<td>185</td>
<td>392</td>
</tr>
<tr>
<td></td>
<td>(12.04)</td>
<td>(6.15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Over 33 Workers</td>
<td>30.92</td>
<td>32.47</td>
<td>-1.55</td>
<td>0.04</td>
<td>334</td>
<td>243</td>
</tr>
<tr>
<td></td>
<td>(9.82)</td>
<td>(7.66)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prediction for Under 33 Workers</td>
<td>33.85</td>
<td>33.09</td>
<td>0.77</td>
<td>0.31</td>
<td>243</td>
<td>334</td>
</tr>
<tr>
<td></td>
<td>(7.03)</td>
<td>(10.14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. column1, row 1 is the average prediction made by female employers about the average productivity of female workers.

B Appendix: Papers Included in Literature Survey

In this section, we list the citation for each paper included in the literature survey.


C Appendix: Qualtrics Surveys

In this section, we include the Qualtrics survey used in the MTurk worker math trivia task, followed by the survey used in the MTurk employer hiring task. Survey block titles (not shown to participants) are in bold and underlined.
Thank you for participating in this survey!

The survey has two parts. In the first part you will answer some very basic demographics questions. In the second part you will answer 50 multiple-choice math questions.

We are interested in determining how many of these math questions you can get right without any help. So please do not use a calculator or look up the answers online, but rather just do your best. The number of questions you answer correctly will not affect your payment in any way.

Demographics

Please answer the personal profile questions below:

What is your favorite color?

What is your favorite movie?

Do you prefer coffee or tea?

- Tea
What is your age?

What is your gender?

- Female
- Male

What is your favorite subject in high school?

What is your favorite sport?

Math

Q1. What is the square root of 289?

- 17
- 19
- 15
- 21

Q2. 4-8*9/2 = ?

- -6
- -32
- -18
- -12

Q3. 3^5 = ?

- 243
- 405
Q4. $5 \times 6 \times 7 = ?$
- 729
- 81

Q5. What is the reduced form of the fraction $70/42$?
- $7/5$
- $14/10$
- $5/3$
- $10/6$

Q6. What is the cubic root of 64?
- 4
- 6
- 5
- 3

Q7. $(4+5)/5 = ?$
- 6.25
- 1
- 1.8
- 5

Q8. $x+2 < 18/3$. Which of the following is necessarily false?
- $x > 4$
- $x < 3$
- $x > 3$
- $x < 4$
Q9. $x^5 \times x^8 = ?$
- $x^{11}$
- $x^{14}$
- $x^{13}$
- $x^{12}$

Q10. Which of the following is approximately equal to 0.833?
- $\frac{5}{6}$
- $\frac{4}{5}$
- $\frac{6}{7}$
- $\frac{3}{4}$

Q11. $x=5$, $y=6$, $z=7$, then what is $\frac{xy}{z-4}$?
- 8
- 10
- 6
- 4

Q12. Which of the following is the closest integer to $\frac{45}{7}$?
- 6
- 5
- 7
- 8

Q13. Which of the following is an integer multiple of 9?
- 3618
- 3619
- 3617
- 3620
Q14. $\frac{10}{5} + 34 - 4 = ?$

- 32
- 34
- 30
- 36

Q15. $(x-1)(x^2-4)=0$, then which of the following cannot be $x$?

- 2
- -2
- -1
- 1

Q16. What is the square root of 196?

- 12
- 13
- 15
- 14

Q17. $5 - \frac{6}{18/9} = ?$

- 2
- -2
- -0.5
- 0.5

Q18. $(y+9)(y^2-121)=0$, then which of the following cannot be $y$?

- 11
- 9
- -9
- -11

Q19. Which of the following is an integer multiple of 11?
Q20. 5+6+7+8+9+10 = ?
- 45
- 51
- 42
- 48

Q21. What is the binary form of 7?
- 101
- 100
- 111
- 110

Q22. 35/7+1 = ?
- 6
- 4
- 7
- 5

Q23. 24/4/3 = ?
- 4
- 3
- 1
- 2

Q24. Which of the following in an integer multiple of 4?
- 66
- 62
Q25. Which of the following is not a prime number?

- 4
- 2
- 3
- 5

Q26. $2 \times 3 \times 4 \times 5 = ?$

- 720
- 24
- 240
- 120

Q27. $6^3 = ?$

- 216
- 432
- 36
- 128

Q28. $(4 \times 2 + 7 \times 8) / 4 = ?$

- 20
- 24
- 16
- 12

Q29. Which of the following is a prime number?

- 23
- 27
- 21
Q30. 16 < x+8 < 26. Which of the following could x be?

- 23
- 18
- 13
- 8

Q31. 45+3-1 = ?

- 48
- 46
- 47
- 49

Q32. x^6 + x^6 = ?

- x^12
- x^36
- (2x)^6
- 2x^6

Q33. Which of the following fractions cannot be further reduced?

- 7/35
- 46/2
- 3/5
- 3/6

Q34. Which of the following numbers has an integer square root?

- 40
- 48
- 32
- 36
Q35. $5 \cdot (7 + 3) + 5 - 4 = ?$

- 51
- 55
- 39
- 32

Q36. Which of the following is not a factor of 30?

- 3
- 5
- 2
- 4

Q37. $x^6 / x^4 = ?$

- $x^{24}$
- $x^{10}$
- $x^2$
- $x^2$
- $x^{(2/3)}$

Q38. $56 / 8 = ?$

- 6
- 5
- 7
- 8

Q39. $2^4 - 3^3 = ?$

- 11
- 9
- -11
- -9

Q40. $(18 + 19 + 20) / 3 = ?$
Q41. Twenty cannot be divided by which of the following?

☐ 5
☐ 3
☐ 2
☐ 4

Q42. 4+8+12+16 = ?

☐ 40
☐ 20
☐ 25
☐ 45

Q43. (x^5)^3 = ?

☐ 5x^3
☐ 3x^5
☐ x^15
☐ x^8

Q44. Which of the following is the correct factorization of 36?

☐ 4 * 9
☐ 2^2 * 3^2
☐ 4 * 3^2
☐ 2^2 * 9

Q45. 3^2*2 = ?

☐ 18
☐ 42
Q46. \(-2\times(-3-8) = ?\)
- -14
- 14
- 22
- -22

Q47. Which of the following is an integer multiple of 5?
- 44
- 46
- 43
- 45

Q48. \(x^4 = 81\). What is \(x\)?
- 9
- 20.5
- 3
- 6

Q49. \(76/4 = ?\)
- 18
- 19
- 17
- 20

Q50. Which of the following is negative?
- \(2^2\)
- \((-2)^2\)
- \((-2)^2\)
- \((-2)^3\)
Thank you for your participation. In addition to your base payment, we may put a small bonus into your account sometime in the next few weeks. Who receives the bonus payment is determined by a different experiment that we are doing and is unrelated to how well you did in the task. Please just think of it as an additional appreciation for your efforts.
Introduction

Thank you for participating in this survey.

The survey has four parts. You will first answer some simple demographic questions. Then you will answer three sets of questions related to people’s performance in math questions.

The survey will take approximately 20 minutes.

Please enter your M-Turk ID:

What is your gender?

- [ ] Male
- [ ] Female

What is your age?

Please indicate the highest level of education you have completed.

- [ ] Less than High School
- [ ] High School or equivalent
- [ ] Vocational/Technical School (2 year)
Setup

We recently paid many people to answer 50 math questions each. Here are some examples of the types of math questions we asked:

**Question 1:** What is the square root of 289?
- Choices: 15, 17, 19, 21

**Question 2:** 4-8*9/2 = ?
- Choices: -6, -12, -18, -32

**Question 3:** What is the reduced form of the fraction 70/42?
- Choices: 5/3, 10/6, 7/5, 14/10

**Question 4:** x^5 * x^8 = ?
- Choices: x^11, x^12, x^13, x^14

**Question 5:** What is the binary form of 7?
- Choices: 100, 101, 110, 111

On average, participants answered **36.95 out of 50 questions correctly.**

Today, you are going to be an employer. You will hire one of the people who answered our math questions. The person you hire will be given a bonus (the wage that you choose to pay them) and in return you will receive money based on how many of the math questions they answered correctly.

Specifically, we are going to provide you with the profiles of 20 people (potential
Specifically, we are going to provide you with the profiles of 20 people (potential employees) who answered our math questions. For each of the 20 people that we present, you will indicate what is the highest wage (between 0 and 50 cents) you would be willing to pay that person. In return, you will be paid 1 cent for every question that the person you end up hiring answered correctly.

After you indicate the highest wage you would be willing to give to each employee, we will randomly draw a number between 0 and 50. If the wage you chose for the employee is equal to or higher than the randomly-drawn number, then that employee will receive the random number as a bonus, and you will receive a profit equal to the number of correct answers given by the individual minus the random number that was drawn. If the highest wage you were willing to pay the individual is lower than the random number, you will not hire the employee and neither you nor the employee will receive a bonus.

Let's walk through an example of how this works. Below is an example of a potential employee profile that you might see:

<table>
<thead>
<tr>
<th>Country:</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender:</td>
<td>Female</td>
</tr>
<tr>
<td>Age:</td>
<td>63</td>
</tr>
<tr>
<td>Favorite High School Subject:</td>
<td>English</td>
</tr>
<tr>
<td>Favorite Sport:</td>
<td>Gymnastics</td>
</tr>
<tr>
<td>Favorite Color:</td>
<td>Sea Green</td>
</tr>
<tr>
<td>Favorite Movie:</td>
<td>Overboard</td>
</tr>
<tr>
<td>Prefers Coffee/Tea:</td>
<td>Tea</td>
</tr>
</tbody>
</table>

We will ask you the highest amount you would be willing to pay this employee. Let's imagine that you say you would be willing to pay this employee 40 cents.
We will then select a random number between 0 and 50. Let’s say the randomly-selected number is 20. Because the highest wage you are willing to pay that person is more than 20, you will "hire" this person and they will receive 20 cents. You will then be paid based on the number of correct answers this person gave. If the person answered 30 questions correctly, you will be paid 10 cents (30-20). If the person answered 10 questions correctly, you will be paid -10 cents (10-20).

Imagine instead that the randomly-drawn number is 45. Then you will not "hire" the person and neither you nor the person will receive a bonus.

In today’s task, you will actually only hire 1 person. After you decide the most you would be willing to pay to each of the 20 people we present, we will randomly select one profile to use as the actual hiring decision. We will then draw the random number between 0 and 50 and pay you the profit you’ve earned for that profile and pay the wage to the person whose profile you pick. We are going to automatically give you a $0.50 bonus in addition to what money you make with your hiring decision (so that there is no way you end up owing us any money after doing this task).

Just to make sure you understand, imagine you saw a profile and entered 43 as the highest amount you would be willing to pay. Now imagine the random number generated was 18 and the individual answered 10 questions correctly.

---

How many cents would you have to pay the individual?

How many cents would you be paid based on the individual’s performance (before subtracting the wage you have to pay the individual)?

Suppose instead that you had reported 15 as the highest wage you would pay, and everything else stayed the same:

How many cents would you have to pay the individual?
How many cents would you be paid based on the individual’s performance (before subtracting the wage you have to pay the individual)?

Enter the highest wage you would be willing to pay this individual (between 0 and 50 cents):

Prediction

Thank you for completing part 2 of 4 of this survey. As promised, we will randomly select one profile and pay you your $0.50 bonus plus whatever money you make
based on the hiring of the randomly-selected profile.

For the third part of this survey, please answer the six questions below. Please remember that people answered 36.95 questions correctly on average.

On average, how many math questions out of 50 do you think **women** answered correctly?  

On average, how many math questions out of 50 do you think **men** answered correctly?  

On average, how many math questions out of 50 do you think **people from the United States** answered correctly?  

On average, how many math questions out of 50 do you think **people from India** answered correctly?  

On average, how many math questions out of 50 do you think **people below or at the age of 33** answered correctly?  

On average, how many math questions out of 50 do you think **people above the age of 33** answered correctly?  

Thank you for completing part 2 of 4 of this survey. As promised, we will randomly select one profile and pay you your $0.50 bonus plus whatever money you make based on the hiring of the randomly-selected profile.

For the third part of this survey, please answer the six questions below. Please remember that people answered **36.95** questions correctly on average.

You have the chance to earn a significant bonus if you answer these questions correctly. We will randomly pick one question and pay you $5 minus your deviation from the correct answer. For example, if your answer for the randomly picked question is 40 and the truth is 37, then you will get a $2 bonus. You cannot receive a negative bonus. So, please answer the questions as carefully as possible so that you can potentially win a large bonus.
Here are the correct answers for the 6 questions you have answered above. On average:

- Women got **35.28** questions right.
- Men got **38.32** questions right.
- People from the U.S. got **37.14** questions right.
- People from India got **36.58** questions right.
- People below or at the age of 33 got **37.10** questions right.
- People above the age of 33 got **36.79** questions right.

Now that you have learned those facts, we would like you to work on 10 more profiles.

As before, after you finish working on those 10 additional profiles, we will randomly select one profile and randomly select a number between 0 and 50. If your highest wage is more than the randomly-selected number, we will pay you the profit you've earned for that profile as a bonus and pay the wage to the person who answered the math questions.
You have completed ${lm://Field/1} of 10 additional profiles.

Please indicate the highest wage you would be willing to pay this employee in the text box below.

Enter the highest wage you would be willing to pay this individual (between 0 and 50 cents):

---

Final

Thank you for your participation. We will calculate your bonus based on the rules specified in each part above, and pay the bonus to your account within a week.

If you have any additional comments about this survey, please provide them below.
(Optional)

---