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

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–26
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Base Rate Neglect as a Source of Inaccurate Statistical Discrimination

David Hagmann,[a,*] Gwendolin B. Sajons,[b] Catherine H. Tinsley[c]

[a] Department of Management, The Hong Kong University of Science and Technology, Hong Kong; [b] ESCP Business School, 14059 Berlin, Germany; [c] McDonough School of Business, Georgetown University, Washington, District of Columbia 20057

*Corresponding author

Contact: hagmann@ust.hk, https://orcid.org/0000-0002-2080-997X (DH); gsajons@escp.eu, https://orcid.org/0000-0002-0469-3153 (GBS); tinsleyc@georgetown.edu, https://orcid.org/0000-0002-8586-0776 (CHT)

Received: February 24, 2023
Revised: July 3, 2024; November 14, 2024
Accepted: December 7, 2024
Published Online in Articles in Advance: September 24, 2025

https://doi.org/10.1287/mnsc.2023.00603

Abstract. Statistical discrimination relies on people inferring unobservable characteristics of group members based on their beliefs about the group. Across four preregistered experiments (N = 9,002), we show that accurate information about the composition of top performers can induce incorrect beliefs about performance differences across groups when the groups are of unequal size. Because people fail to account for base rates, they underestimate the performance of individuals from smaller groups. As a result, when participants in our experiments receive true information about the gender composition of top performers in a male-dominated candidate pool, they are less likely to hire women, even when there are no gender differences in performance (Study 1). Similarly, they are less likely to hire better-performing non-White candidates when the racial demographics of the candidate pool reflect the U.S. population (Study 4). We show that these choices reflect an error in statistical reasoning, rather than being motivated by a desire to discriminate against any particular group (Study 2). Despite leading to less accurate beliefs, participants disproportionately seek out information about top performers when given the choice, and discrimination thus persists when information selection is endogenous (Study 3).

History: Accepted by Dorothea Kübler, behavioral economics and decision analysis.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2023.00603.

Keywords: beliefs • discrimination • base rate neglect • information

Introduction

Workplace discrimination based on demographic characteristics such as gender and race has been widely documented (Bertrand and Mullainathan 2004, Bertrand and Duflo 2017) and can result from *animus* toward a particular group (Becker 1957, Hedegaard and Tyran 2018) as well as *beliefs* held about a group. The latter type of discrimination, which is the focus of this paper, relates to how people's beliefs about social groups shape their inferences about individual group members (Arrow 1971, Phelps 1972). Such beliefs need not be accurate to affect the behavior of decision makers and impose costs on those facing discrimination (see Bohren et al. 2019, 2023). Yet how could incorrect beliefs emerge and persist? We propose base rate neglect as a cognitive driver of false beliefs about the performance of numerically smaller groups. Across four preregistered experiments (N = 9,002), we show that people underestimate the performance of smaller groups after observing the group characteristics of top performers and subsequently engage in statistical discrimination. Our findings demonstrate that accurate information may not only fail to correct false beliefs but can create them in the first place.

Our experiments simulate a hiring context. Participants receive information about the demographic composition of a pool of candidates, which is comprised of either equally sized groups (e.g., 50 women and 50 men) or unbalanced with groups of different sizes (e.g., 20 women and 80 men). We further manipulate the availability of demographic information about top performers (e.g., how many of the top performers are women). Our findings consistently show that information about top performers leads to beliefs and decisions that favor the majority group, even when that group performs worse. Notably, when given the option to view data from across the performance distribution, participants (including a sample with hiring experience) focus on top performers over middle and bottom performers. Across all studies, group imbalance is salient, with clearly explained uneven base rates and a comprehension check regarding this imbalance as a prerequisite for entering the study.[1]

Our results contribute to the literature on base rate neglect by demonstrating how it can lead to statistical

discrimination in settings where decision makers rely on unbalanced samples to infer performance differences between demographic groups. Whereas existing studies have documented the prevalence of base rate neglect (Kahneman and Frederick 2002, Pennycook et al. 2014, Benjamin et al. 2019, Stengård et al. 2022), this paper extends the understanding of its consequences in contexts relevant to organizational behavior and hiring practices.

Whereas our experimental setting resembles a standard base rate neglect problem, the information people receive and the task they need to solve differ. In the traditional paradigm, people receive information about the base rates for two groups as well as a conditional probability indicating how informative a certain characteristic is for inferring group membership. People are then asked to judge the likelihood that someone with the characteristic belongs to one of the groups. A widely replicated finding is that people fail to account for the groups' base rates. For example, people may be provided with the proportion of a population that suffers from a rare disease and the likelihood that a positive medical test result is indicative of having the disease. They then need to estimate the likelihood of someone testing positive to have indeed contracted it. Because they fail to give sufficient weight to the base rates, they overestimate this likelihood (e.g., Stengård et al. 2022).

In our experiments, participants similarly learn the base rates of two (or more) groups, such as the share of a sample being female. The additional information they receive, however, is a conditional proportion: the share of women among the top performers. Participants' task then is to evaluate how predictive (if at all) gender (i.e., group membership) is of performance. This resembles the challenge inherent in a hiring context: how diagnostic are different characteristics for inferring performance? Hiring managers have many potential signals of performance (e.g., major, GPA, ranking of university, internship experience, and recommendation letters), and the challenge is to determine how informative those are in a particular occupation when choosing whom to hire.

Hiring decisions are particularly challenging because they require a prediction about someone's performance based on limited information. One source of information available to hiring managers is the performance of existing employees, allowing them to identify characteristics that may be predictive of success. Information about the composition of top performers may be particularly salient within organizations because outstanding employees often receive explicit recognition (e.g., teaching/research awards at universities, "employee of the month" awards at companies) and rewards (e.g., larger offices, promotions). Outside of any one organization, top 10 lists are similarly ubiquitous, covering everything from the top-paid CEOs to the most successful entrepreneurs or the fastest-growing companies. Prior work has shown that these lists draw considerable attention and can have economically relevant consequences in financial markets (Isaac and Schindler 2014). In addition to being easily available, information about top performers is also something people actively seek. Academic (Starbuck 2006) and nonacademic books (Collins 2001, Gladwell 2008) illustrate the appeal of attempting to learn about characteristics of success from nonrepresentative top performers. Research in psychology suggests that we tend to focus on outliers, attending to distinctive (Hilton and von Hippel 1996) or extreme (Fiske 1980) attributes about others rather than those that are more average and overweight information about outliers when drawing inferences about the group to which the outliers belong (Dannals and Miller 2017). Some scholarship even explicitly advocates more focus on the rare cases at the right tail of a distribution to learn from these extremes (McKelvey and Andriani 2005, Baum and McKelvey 2006, Forgues 2012). In the context of our experiments, we empirically document this demand for information about top performers for both novices and participants with hiring experience (Study 3). Overall, the subtle differences from the structure of a standard base rate problem may explain how base rate neglect could have been "hiding in plain sight" as a potential driver of discrimination against minority groups.

Making correct inferences based on information about the group membership of top performers, however, requires adjusting for the groups' base rates in the overall population. In labor markets, occupations with sizable demographic imbalance are common.[2] The reasons for such imbalance are myriad, including discrimination (Goldin and Rouse 2000) and self-selection into job or study domains (Buser et al. 2014, Buser et al. 2017, Samek 2019). Some demographic characteristics, such as race, are represented unevenly in the population. Consider a firm where the proportion of White, Black, and Asian workers corresponds to their proportions within the U.S. population and in which there are no performance differences across groups. If a hiring manager examined the characteristics of the top performers, they would see mostly White workers. If they failed to adjust for the demographic proportions of the population, they would incorrectly infer that White workers are more productive at the task because they make up the majority of the top performers. Such an error would systematically penalize those who belong to (numerical) minority groups.[3]

One question is whether information about the group membership of top performers is at all informative of performance differences across the groups as a whole. We demonstrate in Appendix A, using simulations of hypothetical performance data, that information about the group membership of top performers in conjunction with base rate information can be informative for

inferring performance differences between groups—and more so when groups are imbalanced.

Specifically, we repeatedly simulate the performance of two groups (X and Y) by randomly sampling from a standard normal distribution. For each simulation, we determine how many of the five largest values are associated with group X and, conditional on this count, calculate the probability that a randomly selected value associated with X is greater than a randomly selected value associated with Y (the probability of superiority). When the two groups each consist of 50 draws (resembling balanced groups), the information that four of the five highest values are associated with X occurs 15% of the time, and the probability of superiority is 53%—slightly higher than the 50% in the absence of information about the group association of the highest values. When there are only 20 draws associated with X and 80 associated with Y (resembling unbalanced groups), then the same information (four of the five highest values are associated with X) is rare, occurring only 0.5% of the time. When it does, however, the corresponding probability of superiority is higher (59%, see Table A.1 for all values).[4]

Drawing proper inferences from unbalanced samples requires adjusting for base rates, which participants in our experiments fail to do. Our aim here, however, is not simply to document this error but also to highlight the organizational and societal consequences thereof. We hypothesize and empirically demonstrate that base rate neglect systematically leads people to underestimate the performance of smaller groups.

In male-dominated industries, which we model in Studies 1 and 3, information about top performers leads people to hire men at a higher rate than women (and, of course, we would expect the reverse for female-dominated industries). For other characteristics that are inherently unequally represented in the population, such as ethnicity in the United States, we find that non-White workers are systematically less likely to be hired when participants have information about the ethnicity of top performers (Study 4). Whereas accurate information should reduce bias and lead to better selection in the absence of inference errors, we show that accurate information is misinterpreted because of base rate neglect, leading to systematic bias.

The next section provides an overview of our four experiments and a description of the performance data that underlie all hiring tasks. We then present the experimental studies as well as their results in more detail and conclude with a general discussion.

Experiments

Our main experiments take place in the context of a stylized hiring task. Participants are presented with real workers (recruited via Amazon Mechanical Turk) who had completed a string reversal task (Huck et al. 2015) in exchange for a piece-rate wage. We will refer to these workers as "candidates," and the number of strings they reverse reflects their performance on the task.

We then present participants with pairs of candidates and ask them to make a series of binary hiring choices. Each pair is drawn from an individual pool of candidates, which we create for each participant by sampling from all available workers. Prior to presenting participants with candidate pairs, we construct four treatments by manipulating the demographic makeup of their pool (balanced or unbalanced), crossed with whether we provide information about the composition of top performers (Studies 1, 2, and 4) or information about the composition of any set of performers they wish to see (Study 3).

In our first study, participants choose between a male and a female candidate, and the pool is either gender balanced or unbalanced (20% female, 80% male). We find that participants who receive information about the composition of top performers in a gender-unbalanced pool incorrectly infer that men perform better on the task and are substantially less likely to hire the female candidate compared with participants in the other three treatments. Our second study replicates this finding with neutral labels—candidates are randomly assigned to "Team A" and "Team B"—suggesting the statistical discrimination observed in our studies does not reflect a motivated error (as in Exley and Kessler 2024). In Study 3, we render performance information endogenous in that people in two information treatments can choose whether to acquire information on low, middle, and/or high performers. We find that 87% of participants seek out information about the top five, whereas just over half of the participants inquire for information about the middle five and bottom five participants. The availability of information likewise leads participants to hire women at a significantly lower rate when they can get information about an unbalanced pool. Finally, in our fourth study, we turn to a setting where not accounting for base rate information comes at a direct cost to the decision maker. Participants choose between a White and a non-White candidate, and the pool is either race balanced (an equal number of White, Black, and Asian candidates) or unbalanced (in proportions representative of the United States population). We find that participants who receive information about the composition of top performers in a balanced pool are more likely to hire the non-White candidate than those who do not receive this information. This finding is consistent with an observed (and unexpected) performance difference in our data, as non-White candidates performed better than White candidates. However, participants who receive information about top performers in the representative pool

end up less likely to hire the non-White candidate than those who do not receive any information.

Open Science Statement

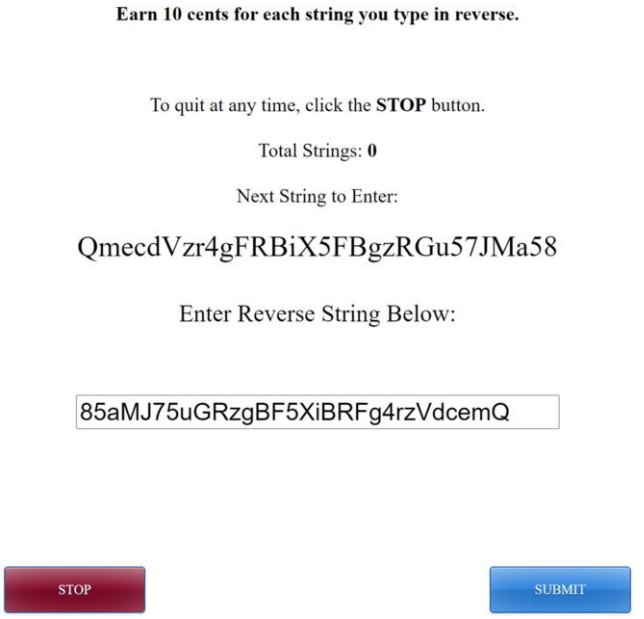
We report all sample sizes, data exclusions, manipulations, and measures. Screen captures of the experimental materials are available in the supplemental information. The complete data and code to reproduce all statistical analyses and figures in the manuscript are available via OSF. Our studies were preregistered on AsPredicted.[5]

Performance Data

We recruit 400 workers via Amazon Mechanical Turk and begin by asking them standard demographic questions, including gender and race. To ensure a gender-balanced sample, we limit advancement to the main module of the study to 200 men and 200 women.[6]

Participants are then introduced to a real effort task in which they have to enter alphanumeric strings of length 30 in reverse (Huck et al. 2015).[7] For each string that they reverse correctly, participants earn 10 cents in addition to a fixed payment of 25 cents. Incorrect entries lead to an error message, and participants can revise their entry with no penalty. They can stop completing strings at any time, including before submitting the first string, by clicking a "stop" button. At that point, the study is completed. They face no time constraints in order to increase variation of performance, but after completing 50 strings, no new strings are generated. We do not inform them of this maximum in advance to avoid creating a "target" level of performance. Figure 1 shows a screenshot of the task.

Figure 1. (Color online) Screenshot of the String Reversal Task



Note. Workers receive a piece-rate incentive for each string they reverse correctly and can decide to stop at any time.

Results. On average, participants successfully reverse 16.1 strings. Twenty percent of participants complete all 50 strings, and 15% of participants complete none. We show the full distribution in Appendix C (Figure C.1).

The number of strings completed do not differ significantly between women and men (16.5 and 15.6 strings, respectively; $t[398] = 0.47$, $p = 0.638$). The chance that a randomly selected female worker completed more strings than a randomly selected male worker also does not significantly differ from 50% (52%; 95% confidence interval (CI): [47, 58]). In Study 2, we ex post randomly assign participants to be members of either Team A or Team B. As expected, we again do not see a significant difference in the number of strings the two teams complete (15.8 and 16.4, respectively; $t[398] = -0.31$, $p = 0.754$).

For Study 4, we look at performance by race. Among our workers, 322 identify as White, 43 as Asian, and 20 as Black or African American (and 15 in other categories). Here, we find differences in performance such that White workers complete the fewest strings, with an average of 14.7. Black workers complete an average of 19.0 (not significantly different from the number of strings completed by White workers; $t[340] = 1.01$, $p = 0.315$). Asian workers complete the most strings, with an average of 22.2, which is significantly more than White workers ($t[363] = 2.48$, $p = 0.014$), but not than Black workers ($t[61] = 0.54$, $p = 0.589$). As we show participants pairs of White and non-White (Black or Asian) candidates, we combine Black and Asian workers and find that they complete significantly more strings on average than do White workers ($t[383] = 2.50$, $p = 0.013$).

Study 1

In our first experiment, we test in an incentivized hiring context whether information on the gender composition of top performers from a pool that is gender imbalanced can induce incorrect beliefs about each gender's relative performance and lead to statistical discrimination. We manipulate, in a 2×2 experimental design, (1) whether participants receive a gender-balanced pool or one that contains more male participants, and (2) whether participants receive information about the gender composition of the top performers in their sample. Our key prediction is that despite the absence of real gender differences in performance, participants who receive information about an unbalanced sample are less likely to hire women than those who receive information about the balanced sample and that this

will not be the case for the treatments where participants receive no information about top performers.

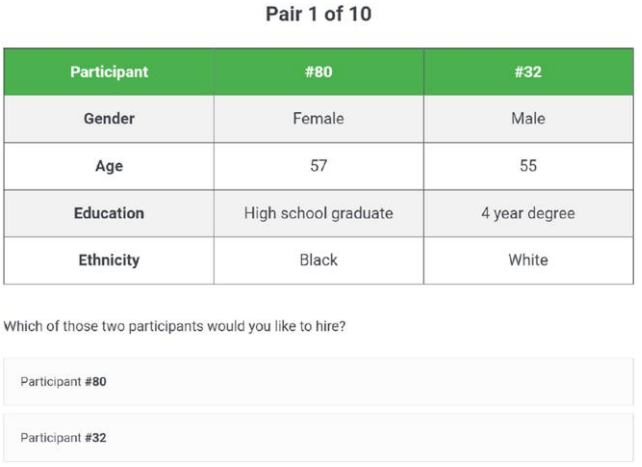
Experimental Design. We open recruitment to 3,000 participants via Prolific Academic, resulting in a sample of 3,002 completing the experiment. All participants learn that their task is to make choices between pairs of candidates and that each candidate is a worker whom we recruited previously from Amazon Mechanical Turk. They see a screenshot of the string reversal task and are informed that the workers could revise their entries in case of error, that they were incentivized via a piece-rate wage, and that they could stop at any time.

For each participant, the candidate pairs are drawn from a randomly generated pool of 100 (out of the 400) previous workers. We vary between subjects whether their pool contains 50 women and 50 men (*Balanced*) or 20 women and 80 men (*Unbalanced*). Participants have to pass a comprehension check consisting of three questions, including one about the gender composition of their pool, on the first attempt to advance to the experiment. Next, we remind participants of the gender composition of their pool and inform them that they will receive information about each candidate's gender, age, level of education, and ethnicity. We include additional demographic information beyond gender to make the task more realistic and provide plausible decision strategies that do not rely on gender. Moreover, the additional information allows us to ask participants to make repeated choices.

We then vary whether participants receive information about the gender composition of the five top-performing candidates in their pool (*Information*) or not (*NoInformation*). Each participant receives individualized information (e.g., three women and two men) based on the randomly drawn pool of candidates. On average, participants in the *Balanced* treatment learn that there are 2.57 women among the top five performers, whereas those in the *Unbalanced* treatment learn that there is 1.00 woman. Note that these are the approximate values expected in the absence of a gender difference in performance. We show the distribution of the information participants receive in Appendix C (Figure C.2).

Participants are then presented with 10 pairs of candidates, each consisting of one woman and one man. We show one example of such a pair in Figure 2. We incentivize participants by selecting one of their pairs at random and paying them five cents for each string that their chosen candidate has successfully completed. After participants make all 10 hiring decisions, we ask them to estimate the number of times out of 100 random comparisons between a man and a woman that the woman completed more tasks, with ties repeated and where each participant could be selected more than once. To incentivize their answer, we pay an additional bonus of 20 cents if their guess is within five of the correct answer. We divide this response by 100 to transform it into the probability of superiority (Ruscio 2008) as a direct and separately incentivized measure of beliefs regarding performance differences between the two groups. Importantly, the probability of superiority measure is invariant to differences in base rates. The true probability of superiority for women is 52% (95% CI: [47, 58]), consistent with no gender difference. The study concludes with basic demographic questions.

Figure 2. (Color online) The Hiring Screen Shown to Participants in Our Four Main Experiments



Notes. In Study 1, participants select from a female and male candidate. In five pairs, the female candidate is on the left side, and in five pairs, the male candidate is on the left. The hiring screens look identical in Studies 3 and 4. In Study 2, the participant number is replaced with team A and team B, and the position is fixed with team A on the left side.

We use the "Quota" option in Qualtrics to limit advancement past the comprehension check to 750 participants in each of the four treatments. We anticipated differential dropout as participants who merely guessed on the gender composition comprehension check might be more likely to select the balanced response. Indeed, whereas 71 participants fail the comprehension check in the *Balanced* treatments, 205 participants do so in the *Unbalanced* treatments ($\chi^2(1, n = 3,280) = 58.56$, $p < 0.001$). To the extent that this differential dropout may bias our results, it likely leads to a conservative test of our hypothesis: here, we include only participants who pay attention to the different base rates, remember them, and may even infer that inclusion in a comprehension check implies that this information is important to their task. They should therefore be less likely than the population overall to succumb to neglecting the base rates.

Results. A test of balance shows successful randomization (see Table S1 in the supplemental information).

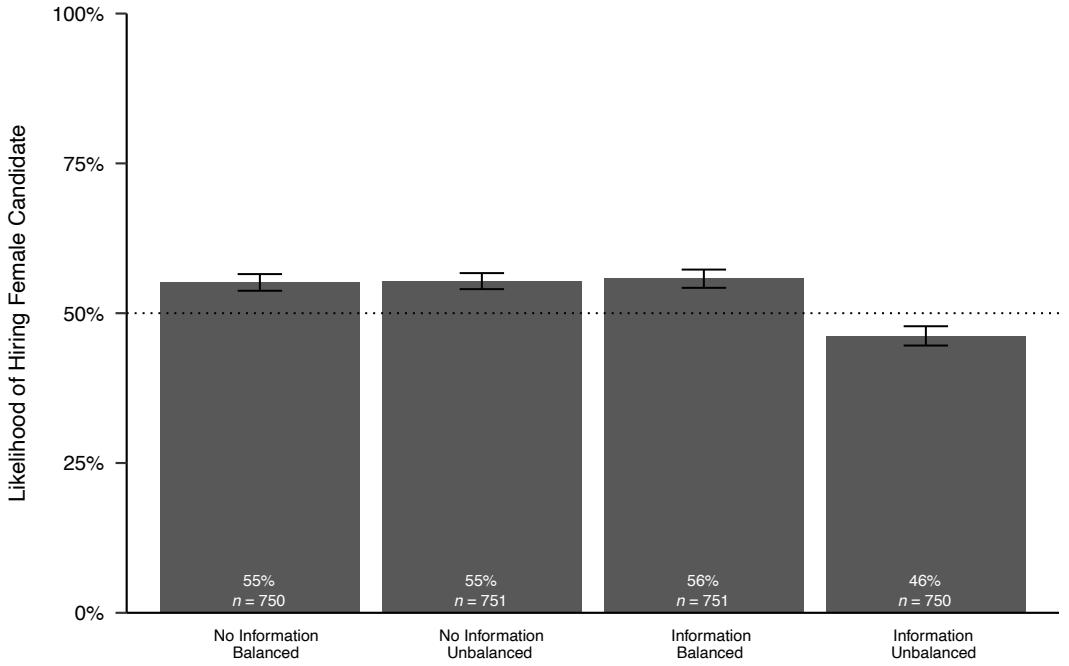
Figure 3 shows the proportion of times the female candidate is selected in each of the treatments. We begin our analyses with a linear probability model in which we use our experimental treatments and their interaction as predictors of the likelihood of hiring the female candidate in each pair (column (1) of Table 1). We present these and all following analyses without demographic controls for the participants or the candidates from whom they choose. Results are robust to including such controls across all studies, as shown in Tables C.2–C.5 in Appendix C. Because each participant makes 10 such binary decisions, we cluster standard errors at the participant level. We subtract 0.50 from the constant such that a positive constant term reflects a preference for hiring the woman in the *Balanced-NoInformation* baseline treatment. As we predicted, the interaction term is negative: participants who receive information about the composition of top performers for an unbalanced pool are about 10 percentage points less likely to hire the female candidate in any given pair than when receiving information about a balanced pool ($p < 0.001$)—a difference that we do not observe in the *NoInformation* treatments.

In the other three experimental treatments, participants exhibit a preference for the female candidate ($p < 0.001$ for proportion tests for a difference from 50%). This could reflect their beliefs about the performance of women or may relate to impression concerns such that they do not want to appear to be discriminating against women. Receiving information for a balanced candidate pool or merely learning that the pool is unbalanced does not significantly affect hiring choice relative to the *Balanced-NoInformation* baseline. In the treatment presenting information about the unbalanced candidate pool, however, participants are significantly less likely than chance to hire the female candidate ($t(749) = -4.61, p < 0.001$).

Next, we look at the probability of superiority (column (2), $P(F > M)$). Similar to our previous regression, we subtract 0.50 such that a significant constant term implies that people anticipate a gender difference in performance. Consistent with the hiring choice, we find that participants in the *Balanced-NoInformation* treatment believe women complete more tasks than men ($p < 0.01$). Participants who receive information about a balanced sample do not differ in their estimate. Unlike in the hiring task, participants adjust their probability of superiority estimate when they are presented with an unbalanced candidate pool. Here, it is possible that participants are unaware that the probability of superiority measure is invariant to base rates and may incorrectly think the larger standard errors of smaller groups affect this measure. Importantly for our hypothesis, however, we find that information about the top performers further decreases the effect of an unbalanced pool on the probability of superiority estimate, as we had predicted ($p < 0.001$).

To look directly at the influence of top performer information on participants' choices, we next compare the two *Information* treatments.[8] Because the candidate pool is generated independently for each participant,

Figure 3. Hiring Decisions in Study 1



Notes. Participants who receive information about the gender composition of the top performers in a pool that is unbalanced toward men are less likely to hire female candidates. Error bars show 95% confidence intervals.

Table 1. OLS Regressions for the Hiring Choice and Probability of Superiority Estimate in Study 1

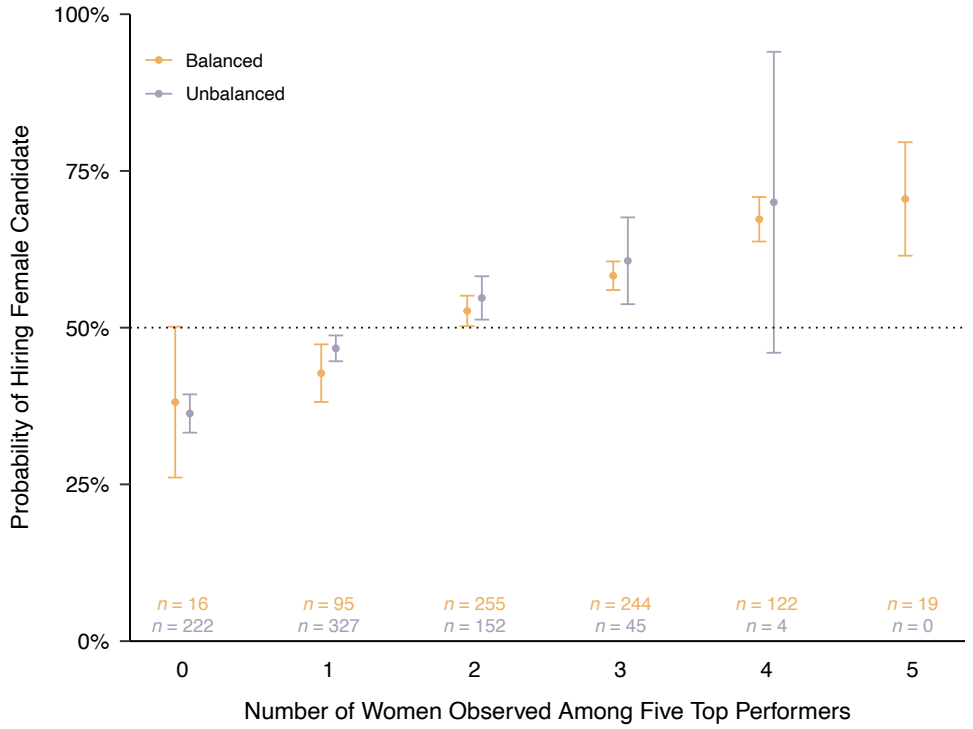
	P(Hire Woman)	*P(F > M)*	*P(Hire Woman)*	*P(F > M)*	*Performance*
Information	0.006	−0.012			0.237
	(0.011)	(0.010)			(0.309)
Unbalanced	0.002	−0.037***	0.002	−0.020	−0.435
	(0.010)	(0.011)	(0.024)	(0.023)	(0.304)
Information × Unbalanced	−0.098***	−0.057***			0.002
	(0.015)	(0.016)			(0.432)
Women In Top 5			0.073***	0.049***	
			(0.007)	(0.007)	
Women In Top 5 × Unbalanced			0.013	0.000	
			(0.012)	(0.012)	
Constant	0.051***	0.023**	−0.130***	−0.115***	16.767***
	(0.007)	(0.007)	(0.021)	(0.019)	(0.217)
N	30,020	3,002	15,010	1,501	30,020
Clustered SE	Participant	No	Participant	No	Participant

Notes. Receiving information about the number of women among the top performers in the candidate pool decreases the likelihood of hiring women if that pool is unbalanced toward male candidates (column (1)) and lowers the estimated probability of superiority for women (column (2)). Seeing more women among the five top performers increases the probability of hiring a woman (column (3)) and the probability of a superiority estimate (column (4)). We subtract 0.50 from these constant terms such that a positive coefficient implies a preference for hiring the female candidate or a probability of superiority estimate greater than equivalence. Statistical discrimination on the basis of false beliefs does not lead to hiring less productive workers (column (5)). Parentheses show standard errors that are clustered at the participant level for hiring choices and robust standard errors for the probability of superiority estimate. SE, standard error.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.1$.

there is variation in the information they receive. In Figure 4, we show how likely participants are to hire the female candidate in the pair, conditional on the information they receive. As one might expect, participants who learn that there are more women among the top performers are also more likely to hire the female candidates. However, participants fail to account for the relative proportion of women in their candidate pool: their choices

Figure 4. (Color online) Likelihood of Hiring the Female Candidate in the Pair as a Function of Information Received



Notes. Participants who learn that there are more women among the top performers in their pool are more likely to hire the female candidate. However, they fail to infer that an equivalent number of female top performers in a sample consisting of 20 women and 80 men as one consisting of 50 women and 50 men implies better performance among female candidates in the former. Error bars show 95% confidence intervals and differ in size as a result of how likely such information is, given the absence of a gender difference in the underlying data. Note that no participant in the unbalanced sample treatment observes five women among the top five performers.

are invariant to whether the number of women among the top five come from a gender-balanced pool or one in which there are four times as many men as there are women. As such, we see a significant effect of the number of women in the top five, but no significant effect of the sample composition nor a significant interaction—another demonstration of base rate neglect (column (3) of Table 1). In column (4), we show that the same finding holds for participants' estimate of the probability of superiority.[9]

We may wonder if statistical discrimination on the basis of incorrect beliefs is costly for the decision maker. Because performance in our task does not differ by gender, we would not anticipate the hiring participants to incur a cost as a result of discrimination. Indeed, that is what we find. Column (5) of Table 1 shows an ordinary least squares (OLS) regression on the number of tasks completed by the selected candidates (recall that participants are paid a bonus based on how many tasks one of their selected candidates had completed). The absence of significant main and interaction effects suggests that information about the number of women among the top performers, as well as the resulting discrimination in one of the treatments, does not affect the productivity of the workers who are ultimately selected. That is, the participants who form incorrect beliefs do not suffer any losses in the hiring task.

Study 2

In this study, we give participants choices between members of two neutrally labeled groups. This design allows us to test whether the statistical discrimination observed in the previous experiment is the result of a motivated error (Exley and Kessler 2024) or whether it can arise in the absence of existing stereotypes and biases. Specifically, some participants may wish to discriminate against women and neglect the group base rates in their candidate pool as cover to enact their discriminatory hiring preferences.

Experimental Design. We create two groups of candidates with neutral labels by randomly assigning the workers from the candidate pool to either Team A or Team B. We show in the supplemental information (Table S2) that the two teams are balanced across observable demographics and performance. We again randomly assign participants to one of four treatments in a 2×2 experimental design and vary whether they receive a balanced pool, with an equal number of Team A and Team B members, or an unbalanced pool, consisting of 20 members of Team A and 80 members of Team B. We further vary whether they receive information about how many members of Team A and Team B are among the five top performers of their pool.

We recruit 2,000 new participants via Prolific Academic. The experiment closely follows the design of Study 1. That is, participants have to pass a comprehension check, including a question asking about the proportion of workers in their candidate pool from Team A and Team B. They then make 10 hiring choices between pairs of candidates (one member of Team A and one member of Team B) and estimate the probability that a member of Team A completed more tasks than a member of Team B (i.e., the probability of superiority). The study concludes with basic demographic questions.

Results. A test of balance shows successful randomization (Table S3 in the supplemental information). A linear probability model replicates our key result from Study 1: participants who receive information on the composition of top performers for the unbalanced pool are significantly less likely to hire minority (Team A) members—an effect that we again do not find in the balanced pool (column (1) of Table 2). We show these results graphically in Figure 5.

Next, we look at the estimate of the probability of superiority. As we predicted, there is a negative interaction of the information treatment and having an unbalanced pool: participants estimate that members of the smaller group complete fewer tasks when they receive information about the top performers in the unbalanced pool, whereas there is no such difference in the balanced pool. As in the previous study, we see a main effect of the unbalanced sample that is not reflected in the hiring choice.

For this second study, we preregistered our previously exploratory analysis, examining participants in the two *Information* treatments. In Figure 6, we show how likely participants are to hire the Team A candidate in the pair, conditional on the information they receive. As we predicted, participants who learn that there are more members of Team A among the top performers are also more likely to hire the Team A candidates. We show the corresponding OLS regression in column (3) of Table 2. Contrary to our prediction, the interaction with the unbalanced sample treatment is significant. That means participants take into account whether the information comes from a balanced or unbalanced sample—yet they adjust inadequately to the imbalance. In column (4), we show that the main effect of the unbalanced sample persists for the probability of superiority estimate, but that the interaction is (as predicted) no longer significant.

Study 3

So far, we exogenously varied whether participants receive information about the composition of top performers and showed that this information can lead to discrimination against the smaller group. However, neglecting the composition of the pool would also have

Table 2. OLS Regressions for the Hiring Choice and Probability of Superiority Estimate in Study 2

	P(Hire A)	*Prob. (A > B)*	*P(Hire A)*	*Prob. (A > B)*
Information	0.006	−0.008		
	(0.011)	(0.012)		
Unbalanced	−0.011	−0.066***	−0.033	−0.044
	(0.009)	(0.012)	(0.026)	(0.026)
Information × Unbalanced	−0.080***	−0.038*		
	(0.015)	(0.017)		
Team A In Top 5			0.056***	0.033***
			(0.008)	(0.008)
Team A In Top 5 × Unbalanced			0.029*	−0.009
			(0.012)	(0.014)
Constant	0.015*	−0.030***	−0.119***	−0.122***
	(0.007)	(0.008)	(0.022)	(0.021)
N	20,000	2,000	10,000	1,000
Clustered SE	Participant	No	Participant	No

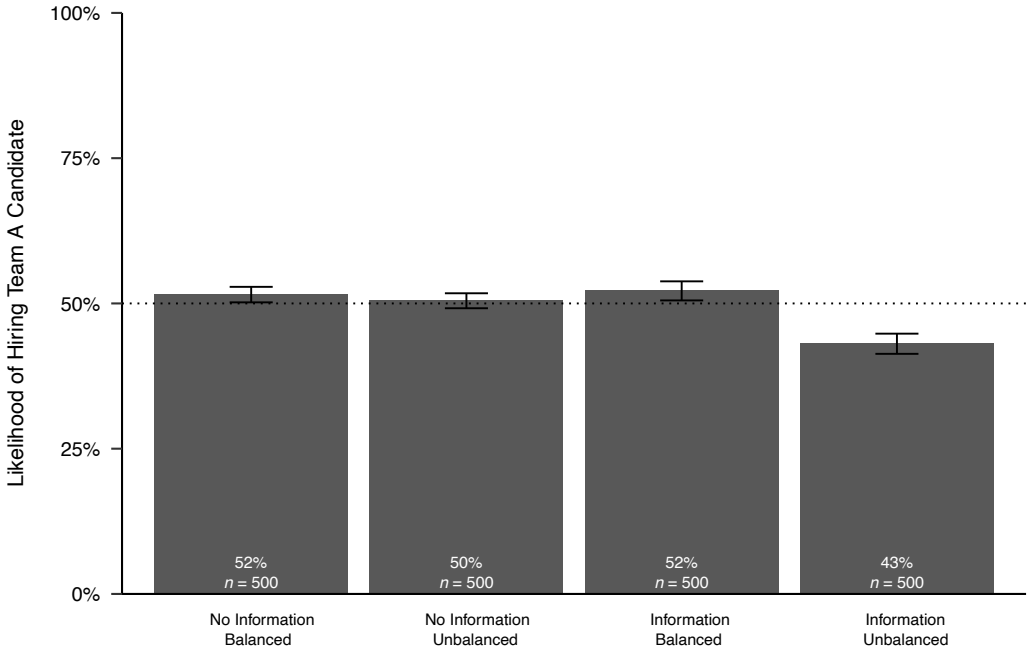
Notes. Receiving information about the number of team A members among the top performers in the candidate pool decreases the likelihood of hiring a member of team A if that pool has more members of team B (column (1)) and the estimated probability of superiority (column (2)). Participants are responsive to information about the number of team A members among the top five performers, but they adjust insufficiently to the pool imbalance in their hiring decision (column (3)) and not at all in their probability of superiority estimate (column (4)). We subtract 0.50 from all constant terms such that a positive coefficient implies a preference for hiring the candidate from team A or a probability of superiority estimate greater than equivalence. Parentheses show standard errors that are clustered at the participant level for hiring choices and robust standard errors for the probability of superiority estimate.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

consequences for how the composition of the worst performers is interpreted. Specifically, if most participants in a pool are men, then not only are most of the top performers expected to be male but also most of the bottom performers, and, of course, most of the average performers, too.

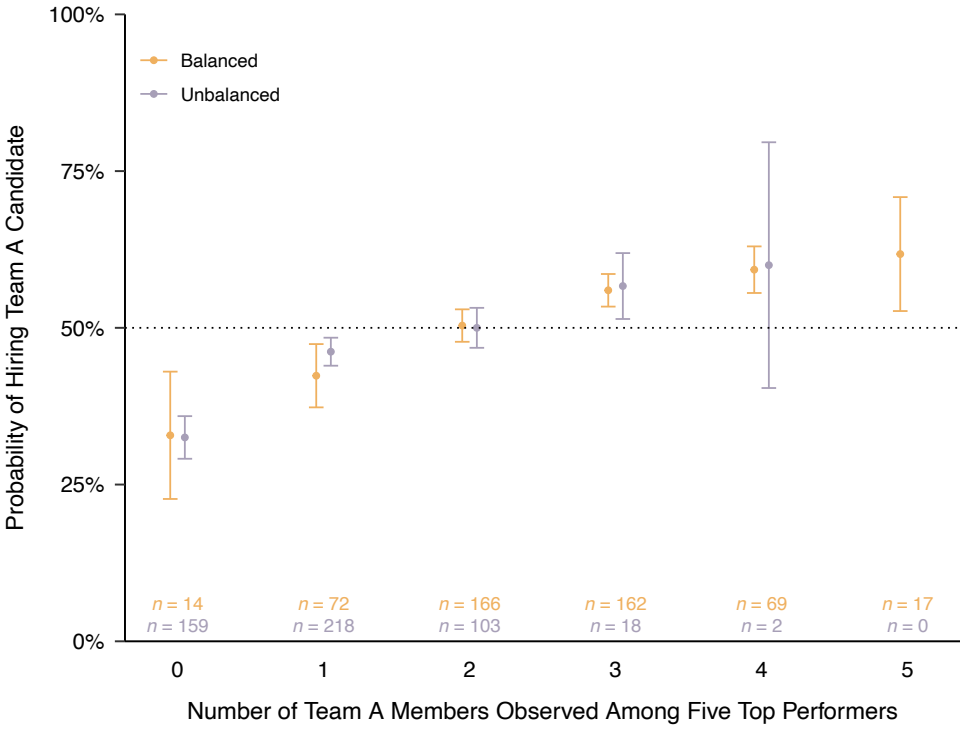
We thus extend our design to a manipulation of the *availability* of different types of information, which participants can then choose to view. Specifically, participants in the *Information* treatments have the option to see the demographics of the top five, middle five, and/or bottom five performers out of a pool of 100 workers who completed the string reversal task. They can select one or multiple groups, or none at all. Notably, seeking information is free of cost to decision makers in this setting. If information was costly, as it

Figure 5. Hiring Decisions in Study 2



Notes. Participants who receive information about the team membership of the top performers in a pool that is unbalanced toward Team B are less likely to hire candidates of Team A. Error bars show 95% confidence intervals.

Figure 6. (Color online) Likelihood of Hiring the Team A Candidate in the Pair as a Function of Information Received



Notes. Participants who learn that there are more members of Team A among the top performers in their pool are more likely to hire the Team A candidate. The rate at which Team A members are hired is largely invariant to whether the information stems from a balanced or unbalanced candidate pool. Error bars show 95% confidence intervals and differ in size as a result of how likely such information is, given the absence of a performance difference in the underlying data. Note that no participant in the unbalanced sample treatment observes five Team A members among the top five performers.

often is in real-world contexts (e.g., by allowing participants to only sample one of the three different types of performers or requiring some payment for viewing), these costs might further increase observed differences in information choice. It is possible that people place more weight on information they actively seek out, rather than information that is exogenously provided to them. This could strengthen the effect of the information treatment relative to the previous experiments. On the other hand, endogenous information choice may also weaken the effect when people choose to not (only) observe top performer information, allowing them to realize that men likewise make up the majority of middle and low performers. In any case, rendering information selection endogenous may strengthen the realism of the study to the extent that real hiring managers actively seek out information, rather than only receiving it passively.

Moreover, instead of highlighting the gender of these performers for participants who elect to receive information, we show a table conveying four pieces of demographic information identical to what is also displayed when participants make the hiring choice (gender, age, education, and ethnicity). Although this way of presenting the information makes our test more conservative, it also better reflects the richer information people would have available in the real world. We also do not display information in terms of relative (numeric) frequencies (i.e., the number of men and women in each group) but, rather, in a more naturalistic format where people see the profiles of multiple employees. They may then try to discover possible patterns by themselves, a design feature that we would also expect to attenuate our information effect. To the extent that showing the frequency of one specific characteristic among the top five performers may have evoked demand effects in the prior studies (despite the incentivization of all decisions), we expect the much more subtle presentation of information in this study to decrease that risk. We further target participants who report having hiring experience as part of their job in our recruitment and report separate analyses for this group.

We predict that participants attend more to information about the top performers than the other groups. Therefore, even though participants could look at all groups and learn that in the gender-unbalanced treatment, men are, on average, overrepresented among all of them, we anticipate that our interaction effect persists in this setting of endogenous information choice.[10]

Experimental Design. We recruit 2,000 participants from Prolific, targeting participants who had previously

responded "yes" to the question "Do you have any experience in making hiring decisions (i.e., have you been responsible for hiring job candidates)?" using a filter provided by the platform. However, after more than two weeks, we filled only 1,407 spots, thus falling short of our preregistered sample size. We then opened the survey to all Prolific participants to fill the remaining spots.

All participants are first informed that they will take the role of a hiring manager for a company and that they will see information about real workers that were previously recruited via Amazon Mechanical Turk. We then randomly assign them to one of four treatments in a 2×2 design. We vary whether the company they are hiring for currently employs a gender-balanced pool of employees (50 women and 50 men) or whether the pool is unbalanced (20 women and 80 men). We further vary whether they have the opportunity to receive information about the performance of existing employees at the company. After passing three comprehension check questions, including one on the gender composition of the employee pool, participants in the *Information* treatment are then informed that they can receive demographic information (gender, age, education, and ethnicity) of three different groups of current employees at the company: the top five performers, the middle five performers, or the bottom five performers. They can select one group, multiple groups, or none of them. After making their choice, they observe the demographics of the selected groups (see Figure 7). We randomize the order in which the groups are presented both when choosing the information (except "None," which is always displayed as the last option) and when observing the information. That is, participants who select all groups may see the demographics of the top five employees first, second, or third, thus counterbalancing potential order effects in the presentation of the information. As before, we generate these groups by randomly sampling 100 workers conditional on the gender composition matching their treatment (either 50 women and 50 men, or 20 women and 80 men) and order them according to their performance, with ties broken randomly. Participants are informed of this procedure and are also informed that the candidates they could hire are not included in this current employee pool. That is, they cannot try to match the demographics of the workers they choose to observe against the demographics they will later observe in their hiring choice.

Following the design of Study 1, participants then make 10 pairwise hiring choices between male and female candidates and estimate the probability of superiority. We incentivize them with a five-cent bonus for each string that one of their selected candidates completed, and with a 20-cent bonus if they are within five percentage points of the correct probability of superiority.

The survey concludes with basic demographic questions as well as a question identical to the Prolific screener asking about hiring experience used to recruit participants initially. In our survey, 70% report having hiring experience as part of their work. Notably, some people who pass the screener provided by Prolific for having hiring experience say "no" to this question (11%), whereas some who are recruited from the standard

Figure 7. (Color online) Screenshot of Information Displayed to Participants in Study 3

Below is the demographic information of the **top five performers** at the company.

	#1	#2	#3	#4	#5
Gender	Male	Male	Male	Male	Female
Age	65	29	42	39	29
Education	Professional or Masters degree	4 year degree	High school graduate	Professional or Masters degree	4 year degree
Ethnicity	Asian	Asian	White	White	Black or African American

→

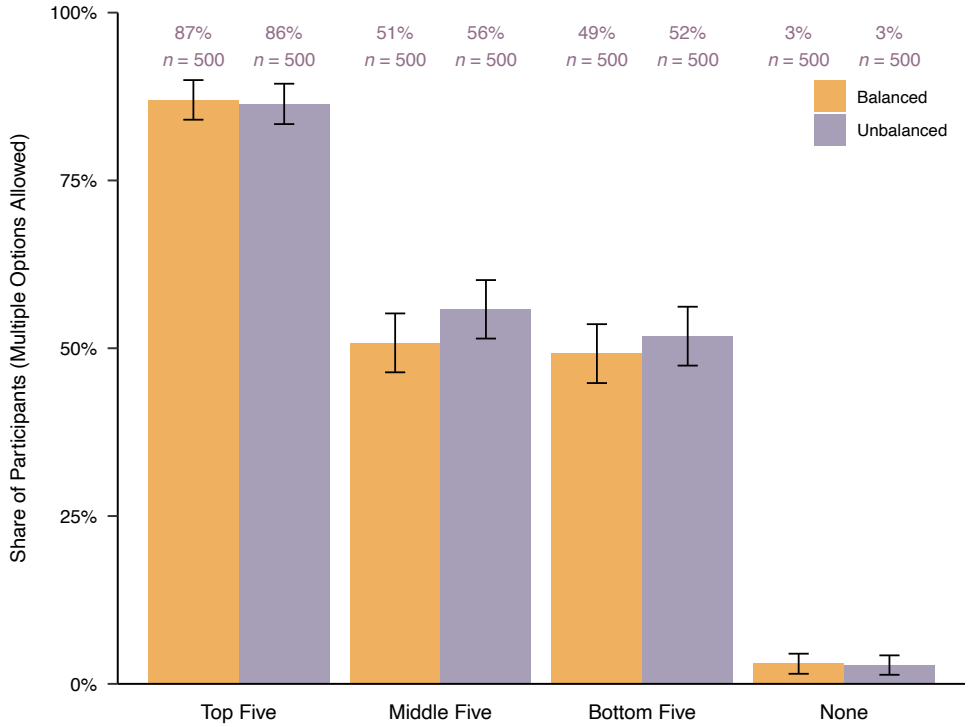
Notes. Participants could choose which performance group(s), if any, they want to observe. This example shows five top performers, taken from the *Unbalanced* sample treatment. The demographics shown differ across participants, as the pool of workers is randomly sampled for each participant.

sample of participants answer "yes" (26%). Following our preregistration, we analyze the data using all participants who complete the study and report separately the regression analyses for participants who report having hiring experience in our study.[11]

Results. A test of balance shows successful randomization (see Table S4 in the supplemental information), and Table S5 shows a comparison of participant characteristics based on whether they self-report having hiring experience. We begin our analyses by examining which group(s) of workers participants would like to receive information about (the top five, middle five, bottom five, or none). We report within-subject *t*-tests comparing the top five against the other three groups. Across the *Information* treatments (*Balanced* and *Unbalanced*), 87% of participants want to learn about the composition of the top five, whereas only 53% want to learn about the middle five ($t(999) = 17.30$, $p < 0.001$), and 51% want to learn about the bottom five, ($t(999) = 21.07$, $p < 0.001$). The difference between the middle five and bottom five is not statistically significant ($t(999) = 1.56$, $p = 0.119$). Fewer than 3% of participants ($n = 29$) elect not to look at any of the information, and 36% of participants look at all three groups. These percentages do not differ significantly by whether the worker pool is *Balanced* or *Unbalanced* (see Figure 8). Notably, however, participants with hiring experience are *less* likely to look at all three groups (33% versus 41%, $t(525.55) = 2.14$, $p = 0.033$) and more likely to seek out only information about top performers (31% versus 25%, $t(579.04) = -1.93$, $p = 0.054$).

Next, we show a linear probability model in which we use our experimental treatments and their interaction as predictors of the likelihood of hiring the female candidate in each pair (column (1) of Table 3). As in our previous studies, we cluster standard errors at the participant level and subtract 0.50 from the constant such that a positive constant term reflects a preference for hiring the woman in the *Balanced-NoInformation* baseline treatment. Merely learning that the pool is *Unbalanced* does not significantly affect hiring choices. However, having the option to look at information increases the rate at which women are hired in the *Balanced* treatment ($p < 0.05$). Moreover, and as we had predicted, we find a significant interaction effect: participants who have the opportunity to receive information about an unbalanced pool are less likely to hire the female candidate in any given pair than those who have the option to receive information about a balanced pool ($p < 0.01$)—a difference we do not see in the *NoInformation* treatments. We show these results graphically in Figure 9. In column (3), we show our results separately for the participants who report having hiring experience.

Figure 8. (Color online) Share of Participants Who Elect to See Different Groups of Participants Before Making Their Hiring Choices in Study 3



Notes. More participants look at the top five workers than any other group. 35.5% of participants choose to look at all three groups (not displayed).

Table 3. OLS Regressions for the Hiring Choice and Probability of Superiority Estimate in Study 3

	All participants		With hiring experience	
	P(Hire Female)	*P(F > M)*	*P(Hire Female)*	*P(F > M)*
Unbalanced	0.011	−0.032*	0.012	−0.026****
	(0.012)	(0.013)	(0.015)	(0.015)
Information	0.027*	0.002	0.020	0.009
	(0.012)	(0.012)	(0.014)	(0.014)
Unbalanced × Information	−0.045**	−0.055**	−0.035****	−0.071***
	(0.017)	(0.018)	(0.021)	(0.021)
Constant	0.020*	0.023**	0.026*	0.024*
	(0.009)	(0.008)	(0.010)	(0.010)
N	20,000	2,000	14,100	1,410
Clustered SE	Participant	None	Participant	None

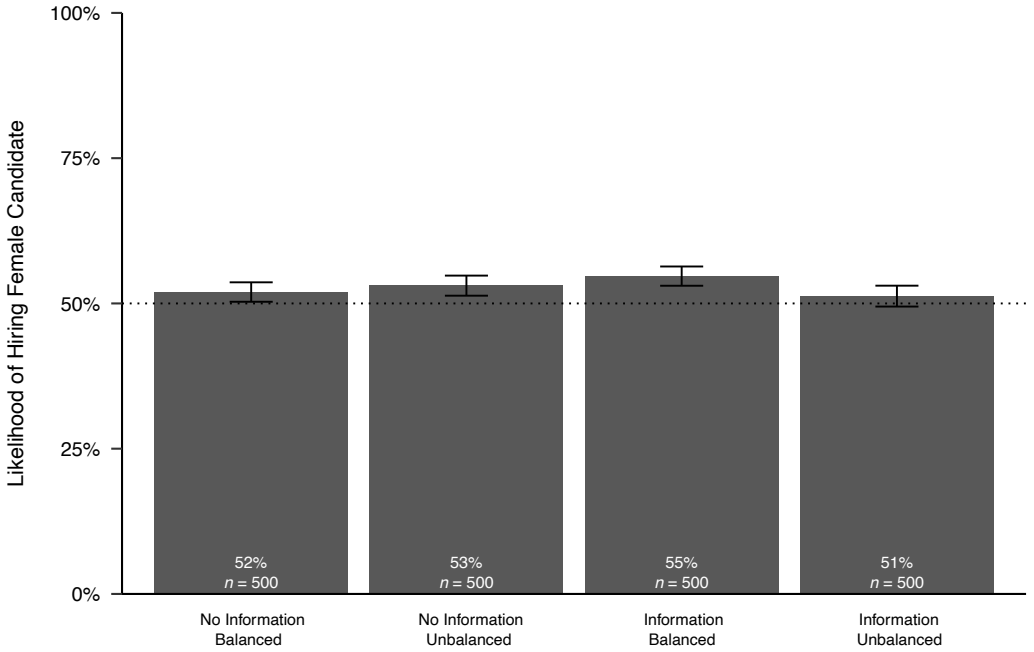
Notes. Receiving information about the number of women among the top performers in the candidate pool decreases the likelihood of hiring women if that pool is unbalanced toward male candidates (column (1)). Participants estimate a lower probability of superiority for women when the candidate pool is unbalanced and do so even more if they receive information about this unbalanced pool (column (2)). These findings replicate in the subsample of participants who report having hiring experience (columns (3) and (4)). Parentheses show standard errors that are clustered at the participant level for hiring choices and robust standard errors for the probability of superiority estimate.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.1$.

Notably, all coefficients go in the same direction as in the full sample, and we still find a marginally significant interaction in this smaller group ($p < 0.10$). Table C.1 in the supplemental information reports separate OLS regressions for participants conditional on their endogenous information choice. Our effect is driven by participants who looked only at the five top performers.

Finally, we look at the probability of superiority. Similarly to our previous analyses, we subtract 0.50 such that the constant term reflects a test of whether participants believe men and women complete the same number of tasks (columns (2) and (4) of Table 3 for all participants and for the subset with hiring experience, respectively). Consistent with our previous results and our preregistered prediction, we find a significant interaction such that participants who can obtain information about the *Unbalanced* sample believe women completed fewer tasks than did men, relative to those who could obtain information about the *Balanced* sample ($p < 0.01$ for all participants and $p < 0.001$ for

Figure 9. Hiring Decisions in Study 3



Notes. Participants who had the option to receive information about the demographics of the top performers in a pool that is unbalanced toward men are less likely to hire female candidates. Error bars show 95% confidence intervals.

participants with hiring experience)—a difference that we do not observe in the *NoInformation* conditions.

Taken together, these results suggest that participants seek out information about top performers more so than about other groups and that false beliefs emerge even when the information choice is endogenous, and participants, in theory, could obtain more information and realize that the larger group is overrepresented among all performers. Participants with hiring experience do not appear to be less vulnerable to this mistake and actually focus more on top performers than do participants overall.

Study 4

So far, we have focused on designs with two groups in which there are no performance differences; thus, participants incur no cost from their false beliefs. Moreover, though gender-imbalanced occupations that motivate the unbalanced samples in the previous experiments are widespread, it may be a sufficiently unnatural candidate composition that participants fail to attend to the consequences of the imbalance. If participants naturally default to gender base rates in the real-world population, they may discount the implications of the base rate information for their sample. Thus, in Study 4, we draw on characteristics that are naturally unbalanced at the population level and where we (unexpectedly) observe performance differences that could make it costly to engage in statistical discrimination based on incorrect beliefs.

According to the 2020 Census, approximately 76% of U.S. residents identify as White, 13% as Black or African American, and 6% as Asian.[12] As a result, a nationally representative group is severely unbalanced across racial groups: a group with 20 members, for example, has, on average, only three Black members and one Asian member. Moreover, in our raw performance data, we observe an unexpected performance difference across racial groups. Specifically, participants who self-identify as White completed, on average, 14.7 tasks, whereas those who identified as Black or Asian completed 21.2 tasks ($t(383) = 2.50$, $p = 0.013$).

If participants do not anticipate such performance differences by racial group, then, in a balanced pool that has the same number of members of each racial group, participants who receive information about the top-performing candidates should be more likely to hire a non-White candidate than participants who do not receive this information. That is, participants should adjust their hiring decisions to favor more non-Whites as they (on average) correctly learn that more non-Whites are among the top performers. However, in a nationally representative (and, hence, unbalanced) candidate pool, participants who receive information about the top-performing candidates could make the opposite inference about group differences in performance if they neglect the known base rates. That is, true and accurate information about the top-performing candidates would lead these participants to draw the wrong inference about group performance differences.

Our extension to race in Study 4 also allows us to examine whether our previous results are robust to people receiving distributional information for three rather than only two groups. For simplicity, our preregistered outcome measure is whether participants hire the non-White candidate in the pair. However, we report decisions involving Black and Asian candidates separately in an exploratory analysis. Because there are approximately twice as many Black candidates as there are Asian candidates in a nationally representative pool, we will be able to compare participants' reactions depending on treatment intensity (i.e., the degree to which the populations are unbalanced). To the extent that they react to treatment intensity, we would expect that Asian candidates are more (negatively) impacted by information about an unbalanced sample than are Black candidates.

Experimental Design. We recruit 2,000 new participants from Prolific and randomly assign them to one of four treatments in a 2×2 design. We vary whether participants receive a candidate pool that contains 20 White, 20 Black, and 20 Asian candidates (*Balanced*), or one that is nationally representative and contains 48 White, 8 Black, and 4 Asian candidates (*Representative*). We further vary whether participants receive information about the racial composition of the top five performers in their sample.

On average, participants in the *Balanced* treatment learn that there are 1.08 White, 1.66 Black, and 2.26 Asian candidates among the top five performers, suggestive of the true performance difference. Those in the *Representative* treatment receive information that, on its face and without accounting for base rates, suggests the reverse order of performance: 3.55 White, 0.86 Black, and 0.58 Asian candidates among the top five performers.[13]

After passing a three-item comprehension check, including one on the racial composition of their pool, participants see the same background information about the string reversal task and the identical hiring screen as in the previous studies.[14] Participants again make 10 decisions between pairs of candidates. Each participant receives their own randomly generated pool of candidates, and all decisions are between a White and a non-White candidate. The study concludes with basic demographic questions.

Our preregistered outcome measure is whether participants hire the non-White candidate in the pair. We omit the probability of superiority measure because of the complexity of extending it to three groups. Our key

prediction is a significant interaction such that participants who receive information about the representative (and hence unbalanced) sample will be less likely to hire a non-White candidate than those who receive information about a balanced sample and that this difference will not equally occur in the *NoInformation* treatments. Moreover, we anticipate that receiving information about a balanced sample may increase the likelihood of hiring a non-White candidate relative to the *NoInformation* treatments if participants do not expect the extent to which non-White candidates indeed performed better.

We again use the "quota" option in Qualtrics to limit advancement past the comprehension check to 500 participants in each of the four treatments and again observe differential dropout across the balanced and representative treatments. Specifically, 80 participants fail the comprehension check in the *Balanced* treatment, and 133 participants do so in the *Representative* treatment ($\chi^2(1, n = 2,218) = 11.39, p < 0.001$).[15]

Results. We show a test of balance across demographic variables in the supplemental information (Table S6). We begin with our primary preregistered analysis: a linear probability model for the decision to hire the non-White candidate in the pair. We use our experimental treatments, their interaction, and a binary variable capturing whether the non-White candidate in the pair is Asian to account for different perceptions of the two non-White groups unrelated to our experimental treatment (column (1) of Table 4). Because each participant makes 10 such binary decisions, we again cluster standard errors at the participant level. We subtract 0.50 from the constant such that the constant's estimate reflects a significance test for a preference for the Black candidate in the *Balanced-NoInformation* baseline treatment.

We find that receiving information about a balanced candidate pool increases the likelihood of hiring the non-White candidate by about five percentage points ($p < 0.001$). This increase suggests that participants do not ex ante anticipate a performance difference across the racial groups (or, at least, not as much as the true performance difference) and, when facing a balanced hiring pool, learn correctly based on top performer information that non-Whites completed more tasks.

Participants choosing between a White and a Black candidate from a balanced sample in the NoInformation treatment hire the latter 58% of the time (significantly greater than equal probability, $p < 0.001$). When choosing between a White and an Asian candidate, the latter is hired 68% of the time (significantly more likely than the Black candidate, $p < 0.001$). As in Studies 1 and 2, changing the sample composition (i.e., balanced versus representative) does not significantly affect hiring choices in the absence of information. However, as predicted, we find a significant interaction effect: receiving information about the top-performing candidates of a *Representative* pool strongly decreases (and even reverses) the effect of *Information* on hiring a non-White candidate compared with the *Balanced* pool. In other words, *Information* about top performers in the *Representative* treatment decreases participants' likelihood of hiring the non-White candidate in any given

Table 4. OLS Regressions for the Hiring Choice and Performance of the Hired Candidates in Study 4

	P(Hire non-White)	*P(Hire Black)*	*P(Hire Asian)*	*P(Hire non-White)*	*Performance*
Information	0.047***	0.059***	0.034*		0.910*
	(0.011)	(0.016)	(0.015)		(0.383)
Representative	0.006	0.000	0.027****	0.000	0.014
	(0.010)	(0.014)	(0.015)	(0.040)	(0.393)
Information × Representative	−0.129***	−0.114***	−0.174***		−1.049****
	(0.016)	(0.021)	(0.024)		(0.552)
White In Top 5				−0.078***	
				(0.010)	
White In Top 5 × Representative				0.015	
				(0.015)	
Asian Candidate	0.098***				2.184***
	(0.008)				(0.289)
Constant	0.078***	0.071***	0.183***	0.259***	17.317***
	(0.008)	(0.010)	(0.010)	(0.014)	(0.307)
N	20,000	11,650	8,350	10,000	20,000
Clustered SE	Participant	Participant	Participant	Participant	Participant

Notes. Receiving information about the representative pool, in which White candidates make up the majority of candidates, reduces the rate at which non-White candidates are being hired (column (1)). This effect holds whether the non-White candidate is Black (column (2)) or Asian (column (3)). Although participants are responsive to information about the number of White candidates among the top five, they fail to account that the implications for performance differ as a result of the pool's imbalance (column (4)). Finally, participants who receive information about a *Balanced* sample hire more productive candidates, but those who receive information about a *Representative* sample do not (column (5)). Parentheses show standard errors that are clustered at the participant level.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.1$.

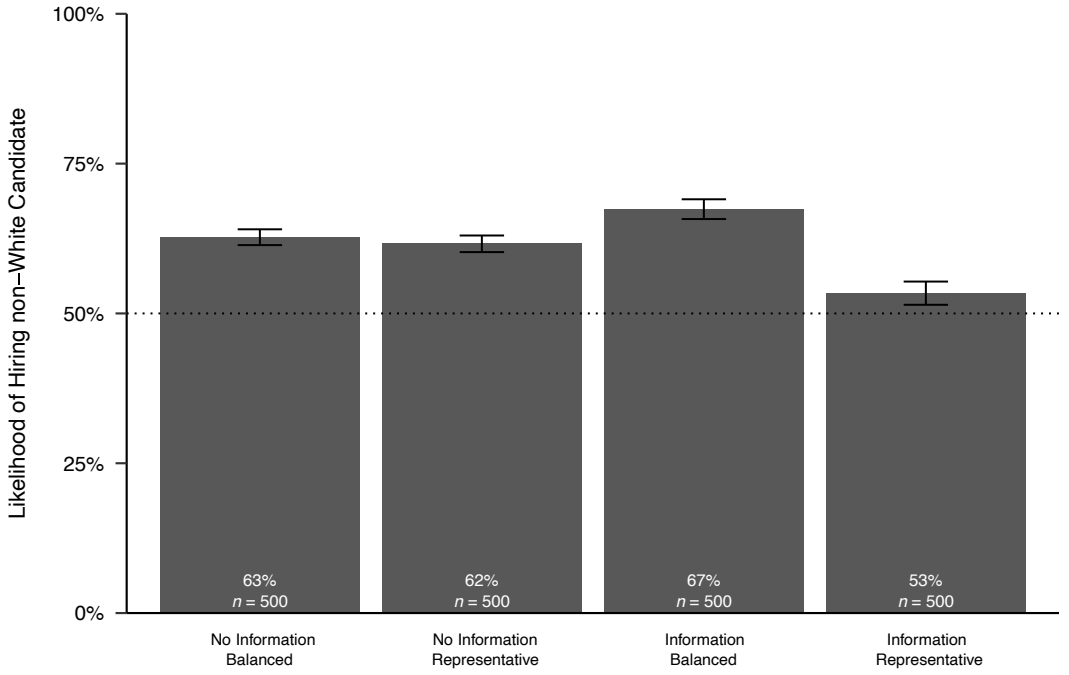
pair compared with receiving *Information* on the *Balanced* sample—a difference that we do not observe in the *NoInformation* treatments. The interaction term is sizable, decreasing the hiring probability by about 13 percentage points ($p < 0.001$). We show these results graphically in Figure 10.

It is possible that our effect is driven by decisions involving either Black or Asian candidates. Columns (2) and (3) of Table 4 thus show the regressions separately for hiring pairs in which the non-White candidate is Black or Asian. We observe that main effects and interactions replicate in both subsets. Notably, the magnitude of the interaction is about 50% greater for Asians than for Black candidates (17.4 percentage points versus 11.4 percentage points). This difference is consistent with a different intensity of the information treatment: a representative sample contains twice as many Black candidates as Asian candidates. Therefore, participants learn that there are even fewer Asian candidates among the top five performers.

As with the first two studies, we look closer at the effect of random variation in top performance information, comparing selection choices in the two *Information* treatments.[16] In Figure 11, we show how likely participants are to hire the non-White candidate in the pair, conditional on the information they receive about the race of the top performers. Participants who learn that there are more non-White participants among the top performers are again more likely to hire the non-White candidate and again fail to account for the size of demographic groups in their pool: their choices are invariant to whether the number of non-White participants comes from a race-balanced pool or one representative of the U.S. population on race. We show the corresponding OLS regression in column (4) of Table 4. Whereas we see a significant effect of the number of non-Whites among the top five candidates, participants do not take into account how many non-White candidates are in the pool, as reflected by the nonsignificant (and small) interaction term.

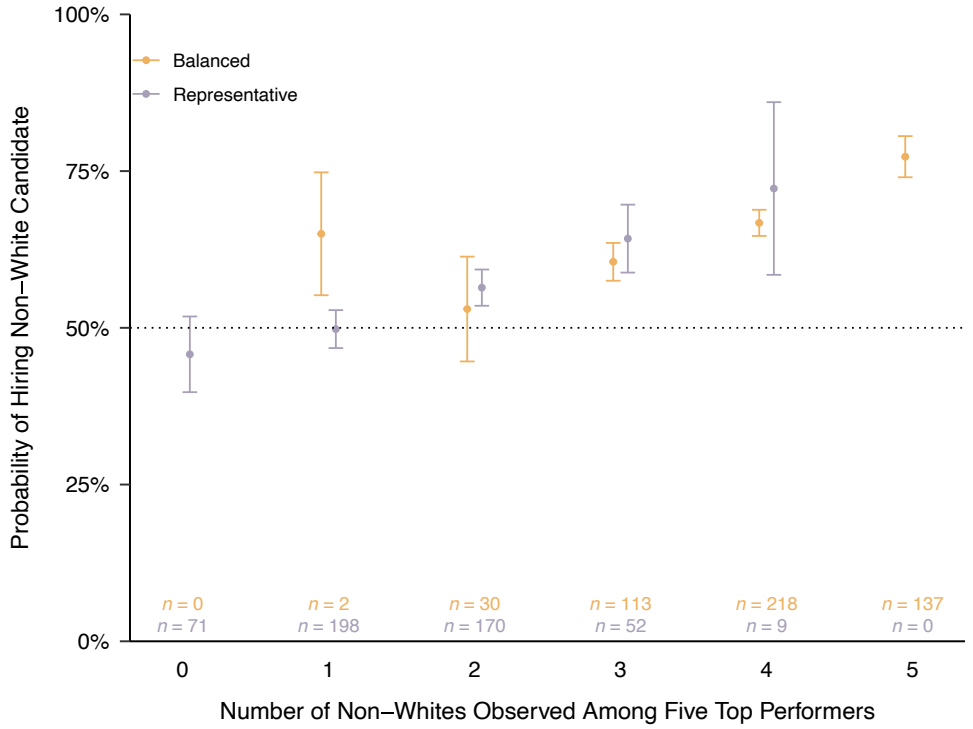
Finally, we look at the earnings of participants across the four treatments. Recall that White candidates completed the fewest tasks and Asians completed the most. Although we do not claim that such performance differences are robust across samples or tasks, participants receive information consistent with such a difference in performance. This raises the question of whether they are able to learn about performance differences in a situation where a failure to do so would potentially come at a direct cost to their experimental earnings. Column (5) of Table 4 shows the number of tasks completed by the chosen candidate. In the baseline *Balanced-NoInformation* treatment, the hired participant completed, on average, 18.4 tasks. When participants receive information about a balanced sample, they indeed hire more productive candidates: the selected candidates completed an average of 0.9 more tasks, which translates

Figure 10. Hiring Decisions in Study 4



Notes. Participants who receive information about the race of top performers in a balanced pool are more likely to hire non-White candidates than participants who receive no information. This difference is consistent with a true (but unexpected) difference in the underlying performance data. However, when they receive information about a pool that is representative of the U.S. population, which is 80% White, they draw the opposite inference and become less likely to hire the non-White candidate. Error bars show 95% confidence intervals.

Figure 11. (Color online) Likelihood of Hiring the Non-White Candidate in the Pair as a Function of Information Received



Notes. Participants who learn that there are more non-White candidates among the top performers in their pool are more likely to hire the non-White candidate. However, their decisions do not account for whether this count comes from the *Balanced* or *Representative* pool. Error bars show 95% confidence intervals and differ in size as a result of how likely such information is observed. Note that two participants observe only a single non-White participant in the top five with a balanced sample. Whereas these confidence intervals do not overlap, the estimates are in the opposite direction of what would be expected: observing an equivalent number of non-White participants in the *Balanced* vs. *Representative* treatment implies a higher performance of the non-White candidates for the representative pool.

into 5% higher performance. Participants who receive information about the top performers in the *Representative* sample, however, fail to choose a more productive candidate, picking someone who, on average, completed 17.9 tasks.

Discussion

Statistical discrimination hinges on beliefs about average differences in unobservable characteristics between different demographic or social groups. Such beliefs, however, can be false or exaggerated. In this paper, we propose base rate neglect as a cognitive driver of false beliefs leading to statistical discrimination. Four experiments present evidence consistent with this proposition. When people obtain information about the composition of top performers for unbalanced samples, they fail to adjust for the demographic base rates, leading them to form systematically false beliefs and discriminate against numerically smaller groups. When there are real performance differences between the groups, this error comes at an economic cost in that they fail to hire more productive candidates on average. Despite leading to false beliefs, when given the choice, people disproportionately seek out information on the composition of top performers as compared with those at the middle or bottom of the performance spectrum.

Notably, the choices resulting from this erroneous belief-based discrimination are indistinguishable from animus-based discrimination but suggest different policy implications. Whereas animus-based discrimination may be addressed with inclusion training, cognitive biases suggest a need for more statistics training or, in this specific case, interventions that highlight the implications of observing information about imbalanced populations. Improving peoples' understanding of this phenomenon would not only counter discrimination but also have the potential to enhance their own (or their organization's) economic outcomes when minority groups perform better.

Teaching people statistical reasoning or even reflection on intuitive answers, however, may be challenging (Meyer and Frederick 2023) and warrants a search for systemic solutions (Chater and Loewenstein 2022). Our results may offer a starting point that could be explored in future research. Throughout this paper, we referred to the *Balanced-NoInformation* treatments as the "baseline" treatments. However, in the real world,

people do observe performance information, and groups are frequently (and, at times, unavoidably) unbalanced. That is, it might be more natural to think of the *Unbalanced-Information* treatments as the real baseline. Conversely, the *Balanced-Information* treatments may then be viewed as a potential intervention to correct false beliefs. That is, organizations could try to purposely present performance information in a format that rebalances for base rate differences. Relatedly, as we have argued, media sources regularly provide information about outliers and top performers, and it is this information that, in our experiment, creates incorrect beliefs and leads to statistical discrimination against minority groups. Yet which of the many potential top performers the media emphasize is ultimately an editorial decision.[17] News could draw from a balanced sample to change the information people receive, even as the underlying population remains unbalanced, or they could create parallel lists that focus on different demographic groups. Our findings thus speak to the importance of representation within and beyond organizations, but especially in what information is presented to people. Future research could examine the effectiveness of such an intervention in a field setting (e.g., with performance information presented to hiring managers). An alternative approach, of course, would be to downplay the focus on top performers, reflecting our *NoInformation* treatments. We suspect, however, that demand for such lists will persist, and as our simulations show, they can convey valuable information when interpreted correctly.

Despite these insights, a limitation of our experimental studies is the artificial setting of the hiring task. Participants make pairwise choices between candidates based on demographic information. Hiring managers normally would have access to additional information that can be more predictive of performance. Also, they would often hire a single candidate from a larger pool, rather than making a single, pairwise decision. Although hiring decisions often come down to a choice among finalists that resembles a pairwise choice, the dynamics of a multiround process involving multiple hiring managers may differ and are beyond the scope of this paper. Moreover, the tasks workers in this study completed (typing strings in reverse) are stylized, and hiring managers may use different strategies to evaluate candidates for more realistic tasks. We used this task because we did not expect people to hold preexisting notions about differences in performance between people who differ in their gender or race. In more naturalistic tasks, stereotypes could exacerbate or narrow the effect of base rate neglect. For example, it is possible that observing many more men among top performers in a computer science task (which is more stereotypical of men) would not get people to pause and reflect on why they are observing more men. However, if they observed that about 40% of inmates in an American prison are non-White, they may cite this as evidence that the justice system is biased against non-White defendants, recognizing that White inmates are "underrepresented" compared with their share of the population. When people do and when they fail to take into account base rates in comparative judgment may be another interesting avenue for future research.

Participants made hiring decisions without observing outcomes. This resembles processes in many large organizations where people hiring workers differ from those who directly observe workers' performance. Organizations, however, can implement feedback systems that could allow hiring managers to learn about the performance of the candidates they had advanced. They might then learn from this information and, ultimately, be disappointed with the performance of the larger group. In many organizations, such feedback systems may not be in place, and we found that participants with hiring experience were even more likely to seek out information about top performers—that is, experience may not resolve this cognitive error when the feedback is not conducive to learning (for a similar example, see Meyer et al. 2018, who find in experiments with more than 14,000 participants that taking the cognitive reflection test up to 25 times did not improve average performance). Whereas our third study confirms that people disproportionately seek out information on the top performers, we did not explicitly investigate what drives this choice. Although we incentivized them for hiring the most productive candidates, it is possible that beyond the aim to identify factors predictive of performance, other motivations such as curiosity about those at the top or a desire to seek information that is positive rather than negative may also play a role (Golman et al. 2022). Future research may investigate what reasons other than a desire for accurate beliefs could drive information search that leads to incorrect beliefs.

Our study also contributes to nascent research on statistical discrimination based on false beliefs about different social groups. Prior studies often consider the source of false beliefs to be erroneous stereotypes (Bohren et al. 2019, Bursztyn et al. 2020) or exaggerated group differences (Bordalo et al. 2016, Coffman et al. 2021). The stereotypic characteristics of a social group (e.g., librarians are shy) are more heavily weighted than the base rate of the social group when assessing whether an individual with the characteristic belongs to the associated group (e.g., Kahneman and Tversky 1973, Lyon and Slovic 1976, Bar-Hillel 1980, Kahneman and Frederick 2002). Discrimination can occur because people tend to exaggerate intergroup differences on a trait while minimizing intragroup differences (Locksley et al. 1980, Nelson et al. 1990). Our paradigm, in contrast, presumes no a priori beliefs about social groups

but, instead, provides one clue for how these beliefs might form in the first place when groups are unevenly represented in populations. Thus, whereas prior research shows that false beliefs can be corrected with accurate performance information (Fershtman and Gneezy 2001, Jensen 2010), our research demonstrates that with unequally sized populations, accurate performance information can lead to the formation rather than correction of false beliefs.

Acknowledgments

The authors are grateful to Daniel Feiler, George Loewenstein, Maurice Schweitzer, and the participants of the CBDR Seminar at Carnegie Mellon University for their helpful comments.

Appendix A. Information Value of Top-Performer Characteristics

In our experiments, participants learn about the group membership of top performers on a task and have to infer whether (and if so, to what extent) to update their beliefs about group differences in performance on the task. We elicit incentivized estimates of the probability of superiority (i.e., the likelihood that a randomly selected member of one group performed better than a randomly selected member of the other group) in addition to the more complex hiring decision that may depend on multiple demographic factors. But to what extent can information about the characteristics of top performers help identify characteristics that are predictive of success? Here, we draw on simulations to establish a benchmark for the value of this information.

Suppose that a population consists of two groups (X and Y), where each member has their "performance" determined by draws from the same normal distribution $N(0,1)$.[18] From this population, we can create a balanced sample consisting of 50 members of each group: $S_B = \{x_1,\ldots,x_{50},y_1,\ldots,y_{50}\}$. We can also create an unbalanced sample, such as one consisting of 20 members of group X and 80 members of Group Y: $S_U = \{x_1,\ldots,x_{20},y_1,\ldots,y_{80}\}$. Within a given sample, group differences can emerge by chance. So, what information does the number of group X and group Y members among the top performers provide about the respective sample?

We begin by examining how many members of group X will be among the top performers for the balanced and unbalanced samples, respectively. We simulate a million balanced and a million unbalanced samples and report in Table A.1 how frequently each count of Group X members is observed among the five top performers ($Top_X = j \in [0,1,2,3,4,5]$). The challenge hiring managers now face is that they observe the top performers from their sample and want to make an inference as to whether and to what extent group membership is informative of performance in their sample. They want to infer $P(x_i > y_i|Top_X = j)$, that is, the probability that a randomly selected member of group X from their sample has a higher performance than a randomly selected member of group Y (the so-called probability of superiority), conditional on having observed j members of group X among the five top performers. This statistic for group differences, importantly, is not affected by base rates (Ruscio 2008). We also report this value in Table A.1. Notably, the table demonstrates that the number of Group X members among the five top performers is informative about the relative performance of the two groups within a given sample.

Table A.1. Predictiveness of Top 5 Information in Simulated Data

	$j = 0$	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
			Balanced sample			
$P(\text{TopX} = j)$	2.8	15.2	31.9	31.9	15.3	2.9
$P(x_i > y_i\|\text{TopX} = j)$	45.0	47.0	49.0	51.0	53.0	55.0
			Unbalanced sample			
$P(\text{TopX} = j)$	31.9	42.0	20.7	4.8	0.5	0.0
$P(x_i > y_i\|\text{TopX} = j)$	46.9	50.0	53.1	56.2	59.2	62.8

Notes. We draw the performance of members of two groups (X and Y) from identical normal distributions $N(0,1)$ and report for a balanced and an unbalanced sample how often each count of group X members occurs among the five top performers, $P(Top_X)$. Conditional on this count, we report the probability that a randomly selected member of group X outperforms a randomly selected member of Group Y, $P(x_i > y_i|Top_X = j)$. As we can see, even when the performance of both groups is drawn from the same distribution, information about top performers is informative for any given sample. All data are given as percentages.

Summing up across the different types of information observed by the hiring managers in the simulation would recover the true performance equality in the population, which we have by construction:

$$\begin{aligned} P(x_i > y_i) &= \sum_{j=0}^{5} P(Top_{X_{50}} = j) \times P(x_i > y_i|Top_{X_{50}} = j) \\ &= \sum_{j=0}^{5} P(Top_{X_{20}} = j) \times P(x_i > y_i|Top_{X_{20}} = j) \\ &= 50\%. \end{aligned}$$

Here, however, is where we propose that people make a systematic error when facing an unbalanced sample. We propose that people interpret the information as if it came from a balanced group, making them "too surprised" when they do not see a member of group X and "insufficiently surprised" when they see many members of group X. That is, they mistakenly use $P(x_i > y_i|\text{Top}_{X_{50}})$ instead of $P(x_i > y_i|\text{Top}_{X_{20}})$ when predicting performance differences between groups. When aggregating across all the samples, we thus get $\sum_{j=0}^{5} P(Top_{X_{20}} = j) \times P(x_i > y_i|Top_{X_{50}} = j) < 50\%$. As a result, participants in the aggregate underselect minority members even when, across all samples, the information they observe is consistent with no performance differences between the groups.

Whereas this simulation is illustrative of the phenomenon, it highlights a potential alternative mechanism for why people would be less likely to select a member of group X (the minority group) in the *Unbalanced* sample. Suppose that someone was perfectly aware of these simulated probabilities of superiority. In the *Balanced* sample, they would choose the member of group X if $j \geq 3$, which occurs 50% of

the time. In the *Unbalanced* sample, they would choose the member of group X if $j \geq 2$ and be indifferent between members of the two groups when $j = 1$. Thus, it is rationalizable that they break these ties in favor of group Y and thus select a member of group X only $20.7 + 4.8 + 0.5 + 0.0 = 26\%$ of the time. To rule this out as an alternative mechanism, we conduct additional simulations using real, rather than identically distributed, performance data, and present these results in Appendix B.[19]

Appendix B. Benchmark for Probability of Superiority Estimates

In Appendix A, we calculate the value of information about group membership of top performers under the assumption that the performance of both groups is drawn from identical distributions. The performance data for groups in our experiments, however, are not *exactly* identical, even in the first three studies in which performance between groups does not differ significantly. Small variations that may result from chance can (and do) perturb this precise equality. We report here simulations similar to those reported in Table A.1, but using the performance data from the workers shown to participants as candidates in our experiment.

Each of the following four tables shows—based on one million draws for the balanced and unbalanced samples—the frequency of observing a particular count of members of the smaller group among the five top performers, the probability of superiority (i.e., the likelihood that a member of the smaller group completed more tasks than a member of the larger group), participants' estimate of this probability of superiority (only for participants in the information treatment that observed this count), and t-tests comparing these estimates to the benchmarks from simulation. Note that participants did not make an estimate about the probability of superiority in Study 4 where they were presented with a

Table B.1. Comparing Simulated Results with Estimates from Participants in Study 1

	$j = 0$	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
			Balanced sample			
$P(\text{TopX} = j)$	2.4	14.3	32.1	33.0	15.5	2.7
$P(x_i > y_i \mid \text{TopX} = j)$	48.7	50.2	51.6	53.1	54.6	56.1
$\hat{P}(x_i > y_i \mid \text{TopX} = j)$	45.3	40.6	49.5	52.4	58.8	63.2
$\hat{P}$ vs. P	$t(15) = -0.87$, $p = 0.400$	$t(94) = -4.77$, $p < 0.001$	$t(254) = -1.84$, $p = 0.067$	$t(243) = -0.54$, $p = 0.587$	$t(121) = 2.18$, $p = 0.031$	$t(18) = 1.45$, $p = 0.165$
			Unbalanced sample			
$P(\text{TopX} = j)$	30.9	42.7	21.2	4.7	0.5	0.0
$P(x_i > y_i \mid \text{TopX} = j)$	50	52.4	54.8	57.3	59.8	60.9
$\hat{P}(x_i > y_i \mid \text{TopX} = j)$	35.4	42.7	46.2	49.9	49.5	—
$\hat{P}$ vs. P	$t(221) = -9.58$, $p < 0.001$	$t(326) = -7.85$, $p < 0.001$	$t(151) = -4.84$, $p < 0.001$	$t(44) = -2.00$, $p = 0.052$	$t(3) = -0.61$, $p = 0.585$	—

Notes. We show how frequently j women are among the five top performers, given the underlying performance data ($P(\text{TopX} = j)$). These numbers may not add up to exactly 100% because of rounding. Next, we report the probability that a randomly selected woman outperformed a randomly selected man, based on one million simulations ($P(x_i > y_i \mid \text{TopX} = j)$). We contrast this simulated value with the participants' average estimate thereof ($\hat{P}(x_i > y_i \mid \text{TopX} = j)$) and report the corresponding t-test statistic. All data are given as percentages.

Table B.2. Comparing Simulated Results with Estimates from Participants in Study 2

	$j = 0$	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
			Balanced sample			
$P(\text{TopA} = j)$	2.3	14.3	32.1	33.0	15.5	2.7
$P(a_i > b_i \mid \text{TopA} = j)$	44.1	45.5	47.0	48.5	50.0	51.5
$\hat{P}(a_i > b_i \mid \text{TopA} = j)$	37.4	41.0	43.8	49.2	50.4	51.5
$\hat{P}$ vs. P	$t(13) = -1.28$, $p = 0.222$	$t(71) = -1.97$, $p = 0.053$	$t(165) = -2.51$, $p = 0.013$	$t(161) = 0.54$, $p = 0.592$	$t(68) = 0.17$, $p = 0.869$	$t(16) = 0.00$, $p = 0.998$
			Unbalanced sample			
$P(\text{TopA} = j)$	31.1	42.6	21.1	4.7	0.5	0.0
$P(a_i > b_i \mid \text{TopA} = j)$	45.2	47.7	50.3	53.0	55.9	58.0
$\hat{P}(a_i > b_i \mid \text{TopA} = j)$	33.1	35.7	40.1	33.7	45.0	—
$\hat{P}$ vs. P	$t(158) = -7.29$, $p < 0.001$	$t(217) = -9.00$, $p < 0.001$	$t(102) = -4.89$, $p < 0.001$	$t(17) = -3.43$, $p = 0.003$	$t(1) = -0.43$, $p = 0.739$	—

Notes. We show how frequently j candidates from Team A are among the five top performers, given the underlying performance data ($P(\text{TopA} = j)$). These numbers may not add up to exactly 100% because of rounding. Next, we report the probability that a randomly selected Team A candidate outperformed a randomly selected Team B candidate, based on one million simulations ($P(a_i > b_i \mid \text{TopA} = j)$). Although we randomly assigned workers to one of the two teams, we find that members of Team B (the larger group) were overall more likely to complete more tasks than members of Team A.

Table B.3. Simulated Results from Performance Data for White and Non-White Candidates Used in Study 4

	$j = 0$	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
			Balanced sample			
$P(\text{TopX} = j)$	9.6	34.7	38.5	15.3	1.9	0.0
$P(x_i > y_i \mid \text{TopX} = j)$	39.0	39.8	40.7	41.7	42.7	44.2
			Unbalanced sample			
$P(\text{TopX} = j)$	34.6	46.2	17.1	2.0	0.1	0.0
$P(x_i > y_i \mid \text{TopX} = j)$	35.7	41.0	47.0	53.9	62.0	80.8

Notes. We show how frequently j non-White candidates are among the five top performers, given the underlying performance data ($P(\text{TopX} = j)$). These numbers may not add up to exactly 100% because of rounding. Next, we report the probability that a randomly selected non-White candidate outperformed a randomly selected White candidate, based on one million simulations ($P(x_i > y_i \mid \text{TopX} = j)$). In this study, we did not collect the probability of superiority because participants received information about three different groups (White, Black, and Asian).

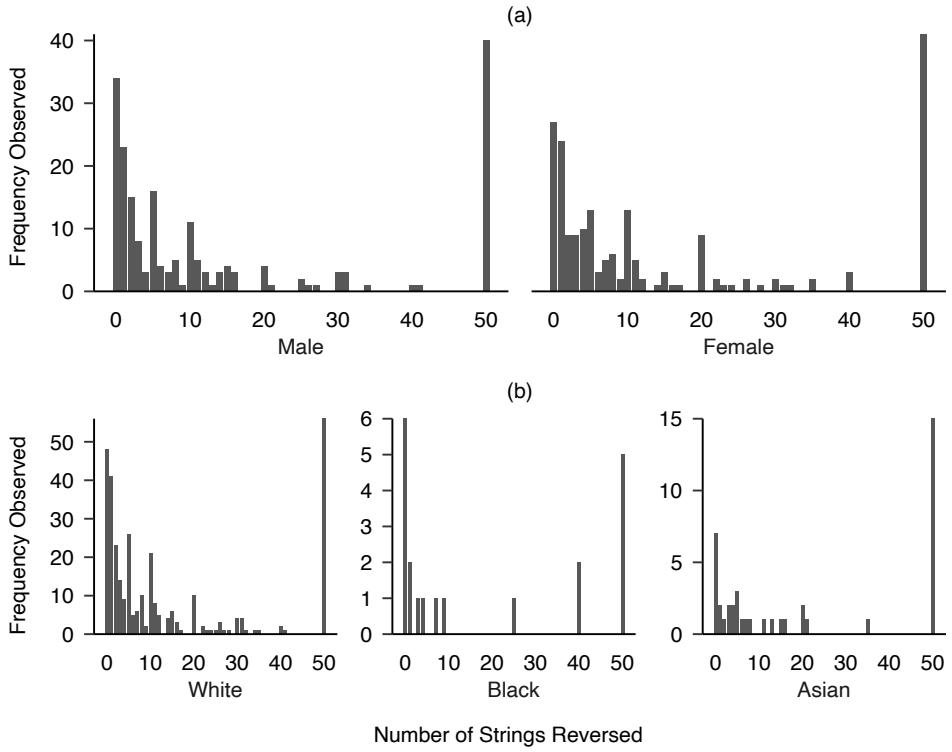
count for White, Black, and Asian participants (rather than a count of non-White participants, which is what we present here for comparability). We omit Study 3 because the underlying performance data are identical to those in Study 1 but information choice was endogenous.

If decision makers always chose the group whose probability of superiority exceeded 50%, conditional on observing a specific count and disregarding all other information, we find that similar to our simulations, the member of the smaller group would be less likely to be selected in the *Unbalanced* sample than in the *Balanced* sample in Studies 1 and 4. However, in Study 2, workers from the smaller team A would be selected more in the *Unbalanced* sample than in the *Balanced* sample. Because the pattern of participants' choices does not differ between Studies 1, 2, and 4, this alternative account does not explain our empirical results.

Note that participants consistently underestimate the probability of superiority for the smaller group in the *Unbalanced* sample. Whereas this finding is consistent with our account, we believe that it could also suggest a misunderstanding of how this measure is calculated. Thus, in the manuscript, we focus on the interaction of sample imbalance with the information treatment and the hiring choice outcome measure.

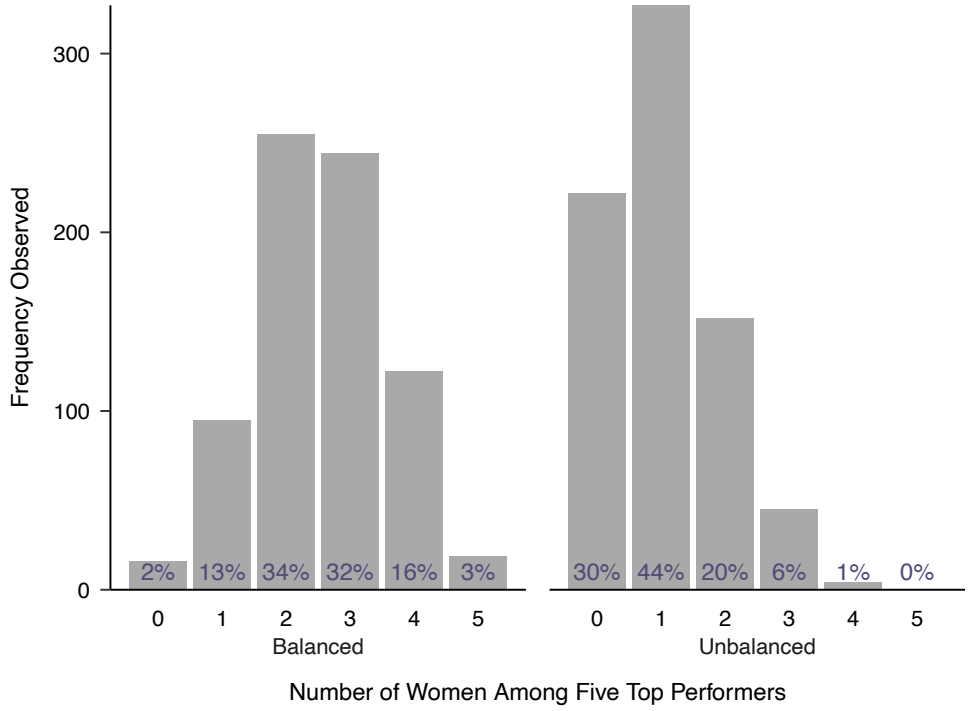
Appendix C. Additional Figures and Tables

Figure C.1. String Reversal Performance Data



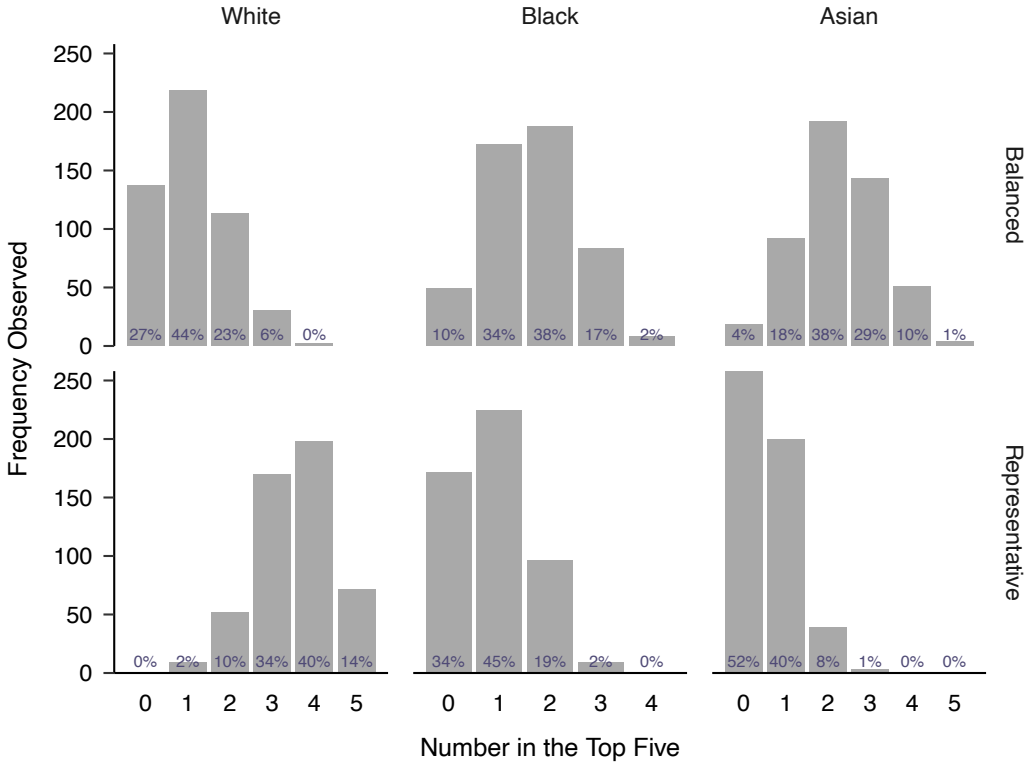
Note. The number of strings successfully reversed by men and women (a) and by White, Black, and Asian participants (b).

Figure C.2. (Color online) Information Shown to Participants in Study 1



Notes. In the *Balanced* treatment, participants observe on average an equal number of women and men among the five top performers. This is consistent with the absence of a performance difference in the underlying data. In the *Unbalanced* treatment, however, most of the top performers are male.

Figure C.3. (Color online) Information Shown to Participants in Study 4



Notes. In the *Balanced* treatment, participants observe more Black candidates than White candidates in the top five, and more Asian than Black candidates. This reflects the true performance difference in our data. In the *Representative* treatment, however, the order is reversed. More than half of all participants do not see a single Asian among the five top performers.

Table C.1. OLS Regressions for the Hiring Choice in Study 3 for Participants Who Select Information About Different Groups of Performers

	Endogenous information exposure: *P(Hire Female)*							
	Top only	Middle only	Bottom only	All three groups	Top + middle	Top + bottom	Middle + bottom	None
Unbalanced	−0.073**	0.004	0.127****	−0.006	−0.057	−0.075*	−0.250***	0.024
	(0.025)	(0.041)	(0.074)	(0.020)	(0.036)	(0.037)	(0.044)	(0.072)
Constant	0.049**	0.038	−0.020	0.028*	0.071**	0.094***	0.200***	0.033
	(0.016)	(0.029)	(0.050)	(0.014)	(0.024)	(0.023)	(0.000)	(0.056)
N	2,910	720	290	3,550	1,030	1,180	30	290
Clustered SE	Participant	Participant	Participant	Participant	Participant	Participant	Participant	Participant

Notes. Participants who look only at the top five performers are seven percentage points less likely to hire the female candidate in a pair when their pool of performers consisted mostly of men (column (1)). Participants who look only at the middle five performers and those who look at all three groups are unaffected by the composition of the sample. The small number of participants who look only at the worst five performers (n = 29) are marginally more likely to choose the majority candidate in the *Unbalanced* sample. Parentheses show standard errors that are clustered at the participant level.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.1$.

Table C.2. OLS Regressions with Additional Controls for the Hiring Choices in Study 1

	(1)	(2)	(3)	(4)	(5)
Information	0.006	0.006	0.004	0.003	0.000
	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)
Unbalanced	0.002	0.002	0.002	−0.001	−0.002
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)
Information × *Unbalanced*	−0.098***	−0.098***	−0.094***	−0.091***	−0.086***
	(0.015)	(0.015)	(0.015)	(0.014)	(0.014)
Constant	0.051***	0.056***	−0.019	0.175***	0.097***
	(0.007)	(0.008)	(0.016)	(0.016)	(0.020)
N	30,020	30,020	30,020	30,020	30,020
Clustered SE	Participant	Participant	Participant	Participant	Participant
Pair number FE	No	Yes	Yes	Yes	Yes
Participant demographics	No	No	Yes	No	Yes
Candidate demographics	No	No	No	Yes	Yes

Notes. The first column shows the regression from the main text of the manuscript. We conduct robustness checks with fixed effects for the pair number (zero to nine), participant demographics (age, gender, ethnicity, and education), and demographics of the male and female candidates (age, ethnicity, and education), respectively. The reference group for participants is male, White, with a four-year degree. We use the same reference group for ethnicity and education for the male and female candidates. Our predicted interaction effect remains significant and of similar magnitude across all specifications. Parentheses show standard errors that are clustered at the participant level. FE, fixed effects.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.1$.

Table C.3. OLS Regressions with Additional Controls for the Hiring Choices in Study 2

	(1)	(2)	(3)	(4)	(5)
Information	0.006	0.006	0.007	0.004	0.005
	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)
Unbalanced	−0.011	−0.011	−0.009	−0.014	−0.013
	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)
Information × *Unbalanced*	−0.080***	−0.080***	−0.081***	−0.078***	−0.079***
	(0.015)	(0.015)	(0.015)	(0.014)	(0.014)
Constant	0.015*	0.020*	0.037*	−0.037*	−0.021
	(0.007)	(0.009)	(0.016)	(0.019)	(0.023)
N	20,000	20,000	20,000	20,000	20,000
Clustered SE	Participant	Participant	Participant	Participant	Participant
Pair number FE	No	Yes	Yes	Yes	Yes
Participant demographics	No	No	Yes	No	Yes
Candidate demographics	No	No	No	Yes	Yes

Notes. The first column shows the regression from the main text of the manuscript. We conduct robustness checks with fixed effects for the pair number (zero to nine), participant demographics (age, gender, ethnicity, and education), and demographics of the team A and team B candidates (age, gender, ethnicity, and education). The reference groups for participants and candidates is male, White, with a four-year degree. Our predicted interaction effect remains significant and of similar magnitude across all specifications. Parentheses show standard errors that are clustered at the participant level.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.1$.

Table C.4. OLS Regressions with Additional Controls for the Hiring Choices in Study 3

	(1)	(2)	(3)	(4)	(5)
Information	0.027*	0.027*	0.024*	0.028*	0.025*
	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)
Unbalanced	0.011	0.011	0.009	0.010	0.008
	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)
Information × *Unbalanced*	−0.045**	−0.045**	−0.042*	−0.045**	−0.043**
	(0.017)	(0.017)	(0.017)	(0.017)	(0.016)
Constant	0.020*	0.031**	−0.026	0.138***	0.083***
	(0.009)	(0.010)	(0.018)	(0.020)	(0.024)
N	20,000	20,000	20,000	20,000	20,000
Clustered SE	Participant	Participant	Participant	Participant	Participant
Pair number FE	No	Yes	Yes	Yes	Yes
Participant demographics	No	No	Yes	No	Yes
Candidate demographics	No	No	No	Yes	Yes

Notes. The first column shows the regression from the main text of the manuscript. We conduct robustness checks with fixed effects for the pair number (zero to nine), participant demographics (age, gender, ethnicity, and education), and demographics of the male and female candidates (age, ethnicity, and education), respectively. The reference group for participants is male, White, with a four-year degree. We use the same reference group for ethnicity and education for the male and female candidates. Our predicted interaction effect remains significant and of similar magnitude across all specifications. Parentheses show standard errors that are clustered at the participant level.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.1$.

Table C.5. OLS Regressions with Additional Controls for the Hiring Choices in Study 4

	(1)	(2)	(3)	(4)	(5)
Information	0.047***	0.047***	0.047***	0.047***	0.047***
	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)
Representative	0.006	−0.011	−0.010	0.000	0.001
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)
Information × *Representative*	−0.129***	−0.129***	−0.129***	−0.128***	−0.127***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Constant	0.078***	0.124***	0.112***	0.035****	0.023
	(0.008)	(0.009)	(0.018)	(0.019)	(0.023)
N	20,000	20,000	20,000	20,000	20,000
Clustered SE	Participant	Participant	Participant	Participant	Participant
Pair number FE	No	Yes	Yes	Yes	Yes
Participant demographics	No	No	Yes	No	Yes
Candidate demographics	No	No	No	Yes	Yes

Notes. The first column shows the regression from the main text of the manuscript. We conduct robustness checks with fixed effects for the pair number (zero to nine), participant demographics (age, gender, ethnicity, and education), and demographics of the White and non-White candidates (age, gender, and education), respectively. The reference group for participants is male, White, with a four-year degree. We use the same reference group for gender and education for the White and non-White candidates. Our predicted interaction effect remains significant and of similar magnitude across all specifications. Parentheses show standard errors that are clustered at the participant level.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.1$.

Endnotes

[1] In many real-world settings, group sizes (such as race in the United States) would have to be inferred, and any errors underestimating size differences would further exacerbate our effect. Underestimation could, for instance, happen when base rates change over time.

[2] For example, regarding gender, the Current Population Survey by the U.S. Bureau of Labor Statistics notes that 90% of mechanical engineers are male, whereas 90% of nurse practitioners are female.

[3] We focus on the top performers in a distribution where higher is better. Notably, the same reasoning would apply to cases where scoring higher is worse (e.g., states with the most COVID-19 infections without accounting for population size) or when making inferences based on the bottom of a distribution (e.g., the states with the most low-performing schools). In these cases, we would expect the prevalence of the relevant characteristic to be underestimated for the larger groups.

[4] Appendix B (Tables B.1–B.3) shows that this pattern also holds for simulations based on the true performance data of our experiments, which we will discuss later. We also report participants' estimates of these probabilities of superiority and document systematic deviations from the simulation benchmark. We are grateful to an anonymous reviewer for proposing these simulations.

[5] OSF Repository: https://osf.io/vxct9. Preregistration for Study 1: https://aspredicted.org/db95-tpty.pdf. Preregistration for Study 2: https://aspredicted.org/d7fs-pthm.pdf. Preregistration for Study 3: https://aspredicted.org/sprt-kc7s.pdf. Preregistration for Study 4: https://aspredicted.org/t589-mzmx.pdf.

[6] We use Amazon Mechanical Turk because the platform does not enforce a minimum base payment. This allows us to conduct a study in which participants differ substantially in their completion times and in which they earn payment primarily through a variable bonus that depends on their effort.

[7] The strings were randomly generated and included small and capital letters and numbers. We excluded ambiguous characters ("I," "l," "0," and "O").

[8] This and the following analyses, reported in columns (3)–(5) of Table 1, were not preregistered.

[9] Note that holding fixed the number of women among the top five performers across treatments does not hold constant the informational content. For instance, observing two versus one woman in the unbalanced sample suggests an increase in women's performance that is higher than when observing the same increase for the balanced sample (see Table A.1). As a result, our test is conservative: even a positive interaction could be consistent with a partial neglect of the gender base rates in the candidate pool. Our results suggest that participants in our study fully ignore the relative proportion of women and men in the candidate pool. This finding is also evident from Figure 4.

[10] Recall that active information seeking is only one of the proposed reasons why information on top performers may affect hiring decisions. Such an effect could additionally work via increased salience of these performers or via hiring managers overweighting information about such workers.

[11] We use their self-reported response collected at the end of our experiment. Coefficient estimates are similar if we instead restrict our analyses to participants recruited with the Prolific filter for having hiring experience.

[12] The Census draws a distinction between "race" and "ethnicity," allowing a responder to identify, for example, as White and Hispanic. We focus here on race rather than ethnicity.

[13] We show the distributions of the top five draws in Appendix C (Figure C.3).

[14] Notably, we do not reorder the demographic characteristics. Thus, whereas the focal demographic was at the top in the previous studies, the focal characteristic of Study 4 (race) is displayed at the bottom (see Figure 2).

[15] We further exclude one participant because the worker demographics did not populate for them due to a technical issue and five participants who exited the study after passing the comprehension check but before completing the study.

[16] This and the following analysis of the performance of selected candidates were not preregistered.

[17] For example, Forbes published an article explaining how the lack of women at the top of the largest companies (and their decision to only look at leaders of the largest companies) had ultimately led to having only one woman on their list of 100 of "America's Most Innovative Leaders" (Lane 2019).

[18] We conducted robustness checks and find that the results do not change if the variance increases or if we instead model performance as bimodal or unimodal beta distributions. Importantly, across all these robustness checks, performance from the two groups is drawn from identical distributions. We return to this point in Appendix B, where we repeat this exercise with the real performance data underlying our experiments.

[19] Note that whereas knowing about characteristics of top performers does hold information value for the samples from which they emerge (and especially so for unbalanced samples), this should not inspire focusing *only* on such information (as when researchers sample on the dependent variable, McDermott 2023). The characteristics of top performers are informative only when accounting for base rates, which requires information about people who are not top performers.

References

Arrow K (1971) Some models of racial discrimination in the labor market. RAND Corporation. https://www.rand.org/pubs/research_memoranda/RM6253.html.

Bar-Hillel M (1980) The base-rate fallacy in probability judgments. *Acta Psychologica* 44(3):211–233.

Baum JAC, McKelvey B (2006) Analysis of extremes in management studies. Ketchen DJ, Bergh DD, eds. *Research Methodology in Strategy and Management*, vol. 3 (Emerald Group Publishing Limited, Leeds, UK), 123–196.

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

Benjamin D, Bodoh-Creed A, Rabin M (2019) Base-rate neglect: Foundations and implications. Working paper, Stony Brook Center for Game Theory, Stony Brook, NY.

Bertrand M, Duflo E (2017) Field experiments on discrimination. Banerjee AV, Duflo E, eds. *Handbook of Economic Field Experiments*, vol. 1 (North-Holland, Amsterdam), 309–393.

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

Bohren JA, Imas A, Rosenberg M (2019) The dynamics of discrimination: Theory and evidence. *Amer. Econom. Rev.* 109(10):3395–3436.

Bohren JA, Haggag K, Imas A, Pope DG (2023) Inaccurate statistical discrimination: An identification problem. *Rev. Econom. Statist.* 107(3):605–620.

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

Bursztyn L, González AL, Yanagizawa-Drott D (2020) Misperceived social norms: Women working outside the home in Saudi Arabia. *Amer. Econom. Rev.* 110(10):2997–3029.

Buser T, Niederle M, Oosterbeek H (2014) Gender, competitiveness, and career choices. *Quart. J. Econom.* 129(3):1409–1447.

Buser T, Peter N, Wolter SC (2017) Gender, competitiveness, and study choices in high school: Evidence from Switzerland. *Amer. Econom. Rev.* 107(5):125–130.

Chater N, Loewenstein G (2022) The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral Brain Sci.* 46:e147.

Coffman KB, Exley CL, Niederle M (2021) The role of beliefs in driving gender discrimination. *Management Sci.* 67(6): 3551–3569.

Collins JC (2001) *Good to Great: Why Some Companies Make the Leap … And Others Don't*, 1st ed. (HarperBusiness, New York).

Dannals JE, Miller DT (2017) Social norm perception in groups with outliers. *J. Experiment. Psych.: General* 146(9):1342–1359.

Exley CL, Kessler JB (2024) Motivated errors. *Amer. Econom. Rev.* 114(4):961–987.

Fershtman C, Gneezy U (2001) Discrimination in a segmented society: An experimental approach. *Quart. J. Econom.* 116(1):351–377.

Fiske S (1980) Attention and weight in person perception: The impact of negative and extreme behavior. *J. Personality Soc. Psych.* 38(6):889–906.

Forgues B (2012) Sampling on the dependent variable is not always that bad: Quantitative case-control designs for strategic organization research. *Strategic Organ.* 10(3):269–275.

Gladwell M (2008) *Outliers: The Story of Success* (Little, Brown and Company, New York).

Goldin C, Rouse C (2000) Orchestrating impartiality: The impact of "blind" auditions on female musicians. *Amer. Econom. Rev.* 90(4):715–741.

Golman R, Loewenstein G, Molnar A, Saccardo S (2022) The demand for, and avoidance of, information. *Management Sci.* 68(9):6454–6476.

Hedegaard MS, Tyran J-R (2018) The price of prejudice. *Amer. Econom. J.: Appl. Econom.* 10(1):40–63.

Hilton JL, von Hippel W (1996) Stereotypes. *Annual Rev. Psych.* 47(1):237–271.

Huck S, Szech N, Wenner LM (2015) More effort with less pay: On information avoidance, belief design and performance. Working paper series in economics, Karlsruher Institut für Technologie (KIT), Karlsruhe.

Isaac MS, Schindler RM (2014) The top-ten effect: Consumers' subjective categorization of ranked lists. *J. Consumer Res.* 40(6):1181–1202.

Jensen R (2010) The (perceived) returns to education and the demand for schooling. *Quart. J. Econom.* 125(2):515–548.

Kahneman D, Frederick S (2002) Representativeness revisited: Attribute substitution in intuitive judgment. Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge University Press, Cambridge, UK), 49–81.

Kahneman D, Tversky A (1973) On the psychology of prediction. *Psych. Rev.* 80(4):237–251.

Lane R (2019) Opportunity missed: Reflecting on the lack of women on our most innovative leaders list. *Forbes* (September 8), https://www.forbes.com/sites/randalllane/2019/09/08/opportunity-missed-reflecting-on-the-lack-of-women-on-our-most-innovative-leaders-list/.

Locksley A, Ortiz V, Hepburn C (1980) Social categorization and discriminatory behavior: Extinguishing the minimal intergroup discrimination effect. *J. Personality Soc. Psych.* 39(5):773–783.

Lyon D, Slovic P (1976) Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica* 40(4):287–298.

McDermott R (2023) On the scientific study of small samples: Challenges confronting quantitative and qualitative methodologies. *Leadership Quart.* 34(3):101675.

McKelvey B, Andriani P (2005) Why Gaussian statistics are mostly wrong for strategic organization. *Strategic Organ.* 3(2):219–228.

Meyer A, Frederick S (2023) The formation and revision of intuitions. *Cognition* 240:105380.

Meyer A, Zhou E, Frederick S (2018) The non-effects of repeated exposure to the cognitive reflection test. *Judgment Decision Making* 13(3):246–259.

Nelson TE, Biernat MR, Manis M (1990) Everyday base rates (sex stereotypes): Potent and resilient. *J. Personality Soc. Psych.* 59(4):664–675.

Pennycook G, Trippas D, Handley SJ, Thompson VA (2014) Base rates: Both neglected and intuitive. *J. Experiment. Psych. Learn. Memory Cognition* 40(2):544–554.

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

Ruscio J (2008) A probability-based measure of effect size: Robustness to base rates and other factors. *Psych. Methods* 13(1):19–30.

Samek A (2019) Gender differences in job entry decisions: A university-wide field experiment. *Management Sci.* 65(7):3272–3281.

Starbuck WH (2006) *The Production of Knowledge: The Challenge of Social Science Research* (Oxford University Press, Oxford, UK).

Stengård E, Juslin P, Hahn U, Van Den Berg R (2022) On the generality and cognitive basis of base-rate neglect. *Cognition* 226:105160.