

How stereotypes impair women's careers in science

Ernesto Reuben^a, Paola Sapienza^b, and Luigi Zingales^{c,1}

^aColumbia Business School, Columbia University, New York, NY 10027; ^bKellogg School of Management, Northwestern University, Evanston, IL 60208; and ^cBooth School of Business, University of Chicago, Chicago, IL 60637

Edited by Anthony G. Greenwald, University of Washington, Seattle, WA, and accepted by the Editorial Board January 31, 2014 (received for review August 5, 2013)

Women outnumber men in undergraduate enrollments, but they are much less likely than men to major in mathematics or science or to choose a profession in these fields. This outcome often is attributed to the effects of negative sex-based stereotypes. We studied the effect of such stereotypes in an experimental market, where subjects were hired to perform an arithmetic task that, on average, both genders perform equally well. We find that without any information other than a candidate's appearance (which makes sex clear), both male and female subjects are twice more likely to hire a man than a woman. The discrimination survives if performance on the arithmetic task is self-reported, because men tend to boast about their performance, whereas women generally underreport it. The discrimination is reduced, but not eliminated, by providing full information about previous performance on the task. By using the Implicit Association Test, we show that implicit stereotypes are responsible for the initial average bias in sex-related beliefs and for a bias in updating expectations when performance information is self-reported. That is, employers biased against women are less likely to take into account the fact that men, on average, boast more than women about their future performance, leading to suboptimal hiring choices that remain biased in favor of men.

gender stereotypes | science education | diversity | science workforce

Why does the proportion of women in science, technology, engineering, and mathematics (STEM)-related professions fail to reflect the interest girls demonstrate for mathematics and science courses in early school years? In high schools in the United States, girls and boys take mathematics and science courses in roughly equal numbers. Standardized-test results suggest that in high school girls are as prepared as boys to pursue science and engineering majors in college. However, from their first year in college, women are much less likely than men to choose a STEM major. College-graduate men outnumber women in nearly every science and engineering field (1). The sex-based disparity in STEM fields is even greater at the graduate-school level (2). In a controversial speech, Larry Summers (3), then President of Harvard University, advanced three hypotheses for this underrepresentation of women in science: different innate aptitudes among men and women at the high end of science-based fields; different career-related preferences among men and women; and discrimination. Although there is mounting evidence against the aptitude-based hypothesis (4–6), it is difficult to show the existence of discrimination if we allow for the possibility of a sex difference in preference; that is, if women truly prefer fields outside of mathematics and science, then their lower proportions in STEM domains may result not from discrimination but merely from preference. That possibility aside, it remains important from a policy point of view to determine whether discrimination exists and, if it does, what can be done to reduce it. For this reason, we designed an experiment in which supply-side considerations did not apply (job candidates were chosen randomly and could not opt out), and thus possible differences in preference could not lead to differences in performance quality (and thus qualification). We used a simple mathematics-related task for which there were no sex differences in performance (7–9).

An important part of our experimental design is that we directly elicited subjects' expectations for job candidates' performance.

This design allowed us to test not only whether performance-related expectations were indeed biased by sex and therefore were the driving force behind any observed exclusion of women but also whether there was an additional bias in the way subjects updated their expectations as they received more information concerning the performance of job candidates and what factors might lead to less biased updating. Last, to understand better the source of expectation biases, we investigated whether associations captured with the Implicit Association Test (IAT) (10) correlated with biases in subjects' initial beliefs and with biases in their updating process when performance-related information was provided by the experimenter or by the candidates themselves.

In our setting, when the employer had no information other than candidates' physical appearance, women were only half as likely to be hired as men, because they were (erroneously) perceived as less talented for the arithmetic task: Both men and women expected women to perform worse. When we allowed candidates to self-report their performance, women were chosen at equally low rates, even though better candidates were chosen on average. The reason is that men are more likely to boast about their performance, whereas women tend to underestimate it. Employers, especially employers with strong implicit stereotypes about women and mathematics, as measured by the IAT, tended not to take this bias into account. The sex gap in hiring was reduced, but not eliminated, by providing the employer with information about candidates' previous performance on the task.

The initial bias in employers' beliefs correlated with implicit stereotypes about women and mathematics, as measured by the IAT. These stereotypes also were partially responsible for the subsequent lack of complete Bayesian updating. Interestingly, we

Significance

Does discrimination contribute to the low percentage of women in mathematics and science careers? We designed an experiment to isolate discrimination's potential effect. Without provision of information about candidates other than their appearance, men are twice more likely to be hired for a mathematical task than women. If ability is self-reported, women still are discriminated against, because employers do not fully account for men's tendency to boast about performance. Providing full information about candidates' past performance reduces discrimination but does not eliminate it. We show that implicit stereotypes (as measured by the Implicit Association Test) predict not only the initial bias in beliefs but also the suboptimal updating of gender-related expectations when performance-related information comes from the subjects themselves.

Author contributions: E.R., P.S., and L.Z. designed research; E.R. performed research; E.R., P.S., and L.Z. analyzed data; and E.R., P.S., and L.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.G.G. is a guest editor invited by the Editorial Board.

¹To whom correspondence should be addressed. E-mail: luigi.zingales@chicagobooth.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1314788111/-DCSupplemental.

documented an important pattern related to the updating process. When the information was “objective” (i.e., provided by the experimenter), the updating, although not complete, was not biased by the preexisting stereotype (as measured by the IAT). In contrast, when the information was provided by the subjects themselves, employers biased against women were less likely to realize that, on average, men boast more about their performance than women do, leading to a biased and suboptimal choice in favor of men.

Methods

We used a laboratory experiment in which subjects were “hired” to perform an arithmetic task: correctly summing as many sets of four two-digit numbers as possible over a period of 4 min. We chose this task because of the strong evidence that it is performed equally well by men and women (7–9). Nevertheless, it belongs to an area—mathematics—about which there is a pervasive stereotype that men perform better (11–13).

First, all subjects performed the task and were informed of their performance (the number of problems they solved correctly). Subsequently, two subjects were selected randomly to be candidates; the remaining ones were to act as “employers,” hiring one of the candidates from the pair to perform a second arithmetic task of the same type as the original. Although the employers chose candidates from pairs representing any combination of genders, including same-sex pairs (e.g., two women), we analyzed data only from instances in which the two candidates in the pair were of different genders (one woman, one man). We did so to avoid making sex overly salient as a factor in the employers’ decisions. Employers provided two responses for each pair of candidates they evaluated: (i) choosing one of the two candidates as their “employee” and (ii) estimating the number of sums each candidate would complete correctly on a second arithmetic task. Candidates earned more money in the experiment if they were chosen by the employer. Employers earned more if they chose the candidate who performed better than the other candidate in the pair on the second arithmetic task.

We implemented four different treatments by varying the information available to employers when they chose between candidates, and we offered some employers the ability to update their choices after additional information about the candidates was provided. Each subject was assigned randomly to one of the four treatments described below and participated in multiple repetitions of the experiment within that treatment. The exact number of repetitions for a given subject depended on the total number of subjects in a particular session and the number represented by each sex. In every treatment, subjects assigned to act as employers first saw the pair of candidates from which they were to choose, allowing them to identify the candidates’ sex. In the first treatment, which we label “Cheap Talk,” each candidate in the pair communicated to the employer their expected performance on the second arithmetic task before the employer chose one of the pair as employee. In the second treatment, which we label “Past Performance,” employers were told the actual performance of each candidate in the first arithmetic task (the number of problems solved correctly) before choosing one candidate as employee. In the third treatment, labeled “Decision Then Cheap Talk,” employers first chose a candidate to hire without

information other than the candidates’ appearance—a departure from the previous two treatments, in which, before making a hiring decision, employers both saw the candidates and received information about their performance on the task from the experimenter or from the candidates themselves. After making their choice (and estimating how both candidates in the pair would perform on the task), employers in this treatment saw the candidates’ self-reported expected performance and were asked to update their choice of candidate and estimates of performance, thus providing a second set of responses. Similarly, in the fourth treatment, “Decision Then Past Performance,” employers made their initial decisions based only on the candidates’ appearance and then updated their decisions after being informed (by the experimenter) of the candidates’ actual performance on the original arithmetic task. Table 1 summarizes the characteristics of each of the four treatments and provides the number of employers and mixed-sex candidate pairs in each treatment.

As a final step, we asked all subjects to complete an IAT associating sex with science-related abilities (10). The IAT is a computer-based behavioral measure in which subjects rapidly place words and pictures that they observe on their screen into categories; easier pairings (as indicated by faster responses) are interpreted as more strongly associated in memory than more difficult pairings (as indicated by slower responses). In socially sensitive domains, the IAT is more reliable than self-reported measures because it bypasses the influence of the subjects’ social desirability bias on responses (14). For our setting, we used an IAT that required subjects to associate words/pictures with the categories “male,” “female,” “math and science,” and “liberal arts.” In one condition, subjects used the same key to categorize items representing male (e.g., a picture of a man) and math/science (e.g., the word “calculus”) and another key to categorize items representing female (e.g., a picture of a woman) and liberal arts (e.g., the word “literature”). In the other condition, subjects categorized the same words/pictures, but the words and pictures were paired differently: Male and liberal arts appeared together, and female and math/science items appeared together. Most people categorize the words/pictures faster and more accurately in the male-math/science condition than the female-math/science condition. This difference is interpreted as reflecting an implicit sex-math/science stereotype such that males are seen as more capable in these fields. All the data from the experiment, including the subjects’ decisions, expectations, and IAT scores, are available in [Dataset S1](#).

Results

Our results revealed a strong bias among subjects to hire male candidates for the arithmetic task. This bias was present among both male and female employers, related to their expectations of candidate performance by sex (as suggested by IAT scores), and remained undiminished by candidates’ self-reports of expected performance, largely because males tended to overestimate future performance. Objective information about past performance (how subjects actually performed on the task) attenuated sex-biased decision-making in this context but failed to eliminate it, especially in employers who showed a stronger implicit sex

Table 1. Characteristics and available information in each treatment of the laboratory experiment

	Cheap Talk	Past Performance	Decision Then Cheap Talk	Decision Then Past Performance
Number of employers	38	49	51	53
Number of mixed-sex candidate pairs	15	23	18	20
Mean number of mixed-sex candidate pairs per employer	4.21	5.41	5.27	4.49
Number of picking decisions	160	265	269	265
Information available for initial guesses and pick	Appearance and expected performance	Appearance and past performance	Appearance	Appearance
Additional information given for subsequent guesses and pick	N/A	N/A	Expected performance	Past performance

For each treatment, the table presents the number of subjects who acted as employers when a mixed-sex alternative was presented to them, the total number of mixed-sex candidate pairs, the mean number of decisions per employer in the mixed-sex pair, the total number of picking decisions across all sessions, and the type of information available to the employers in each treatment. We also use data when employers have no information on the candidates. Those data are collected before the Decision Then Cheap Talk and Decision Then Past Performance treatments, and the corresponding observations are the sum of the two treatments (total picking decisions, $n = 507$). For a detailed description of each session, see [SI Appendix, Table S1](#).

bias as revealed by the IAT. Detailed versions of these results are presented in the sections below.

Initial Hiring Decisions and Sex-Related Beliefs. Because employers were rewarded based on the quality of their picks, we expected that their choice of candidate would be guided by their beliefs about who would perform best. An objective of this paper, then, is to show that these performance-related beliefs were biased based on sex. To measure the extent of this distortion, we needed a benchmark the depended on the information available to the employer. We considered two extreme benchmarks: complete ignorance and perfect information. A completely uninformed prior (i.e., no information about the candidates in question) assigns equal probability to either the man or the woman being superior on the task. This prior is consistent with our in-sample performance (*SI Appendix*) and with the existing literature (11–13). In contrast, the full-information prior assumes employers know the actual future performance of the two candidates. Note that the employers in our study did not have this information, because at best they learned the candidates' performance on the first arithmetic task, which was highly predictive (Pearson's $r = 0.845$, $P < 0.001$) but was not identical to the candidate's actual performance on the second arithmetic task.

We started by analyzing employers' initial hiring decisions under the different treatments. For this purpose, we pooled together the initial decisions in the Decision Then Cheap Talk and Decision Then Past Performance treatments, in which subjects had no information about the candidates' performance, thus creating a No Information condition. As a result, initial hiring decisions are compared across three conditions, rather than our original four.

We found substantial discrimination against female candidates across conditions (Fig. 1). When employers had no information beyond appearance, they were twice more likely to choose male candidates than female candidates. Regression analyses (*SI Appendix*, Table S4) show that the fractions of female candidates chosen in the No Information and Cheap Talk conditions were almost identical (0.2 percentage points less in the Cheap Talk condition, $P = 0.972$), whereas the proportion was significantly higher in the Past Performance condition (9.1 percentage points more than in the No Information condition, $P = 0.004$; 9.3 percentage points more than in the Cheap Talk condition, $P = 0.076$). However, in all three conditions the proportion of female candidates was significantly

less than 50% ($P < 0.003$), the fraction that would have been chosen if there were no discrimination.

The cost of this discrimination pattern for employers and candidates varies by condition. In the No Information case, discrimination is not very costly for employers. If we remove the anti-women bias in expectations, employers would earn only 0.1% more in compensation. If, instead, we were to impose a random choice on employers, their earnings would drop by 11.4%, because employers do gain some relevant information from the appearance of the candidates, and this information allows them to make better-than-random choices (as can be seen in Fig. 1, which shows that employers in this condition choose the higher-performing candidate 55% of the time). Imposing a random choice would take away the benefit of this information. Still, although the cost for employers in this context is low, the cost for women is high: In the No Information condition the expected earnings of female candidates is 19.4% less than that of their male counterparts.

Moreover, our ex post analyses show that employers made suboptimal hiring decisions across conditions, with the worst decision-making in the No Information condition. A strength of our experimental design is that, in addition to detecting sex biases in the overall hiring decisions, it allows us to determine the degree to which decisions were suboptimal ex post (i.e., cases in which the candidate with the lower performance is chosen) and whether suboptimal decisions were biased in favor of men. The highest fraction of suboptimal decisions occurred in the No Information condition, in which almost half of the hiring decisions were suboptimal (Fig. 1). Regression analysis (*SI Appendix*, Table S5) showed that employers made the suboptimal decision significantly less often in the Cheap Talk condition than in No Information condition (by 13.1 percentage points, $P = 0.004$), suggesting that the candidates' statements about future performance contained useful information. Employers made even fewer suboptimal picks in the Past Performance condition (25.0 percentage points less than in the Cheap Talk condition, $P = 0.031$). In all three conditions, the higher-performing candidate was picked significantly more often than would have occurred by chance (by at least 4.6 percentage points, $P < 0.010$). However, hiring decisions were still far from optimal. For instance, if employers in the Past Performance condition based their choice solely on candidates' relative past performance (i.e., always choosing

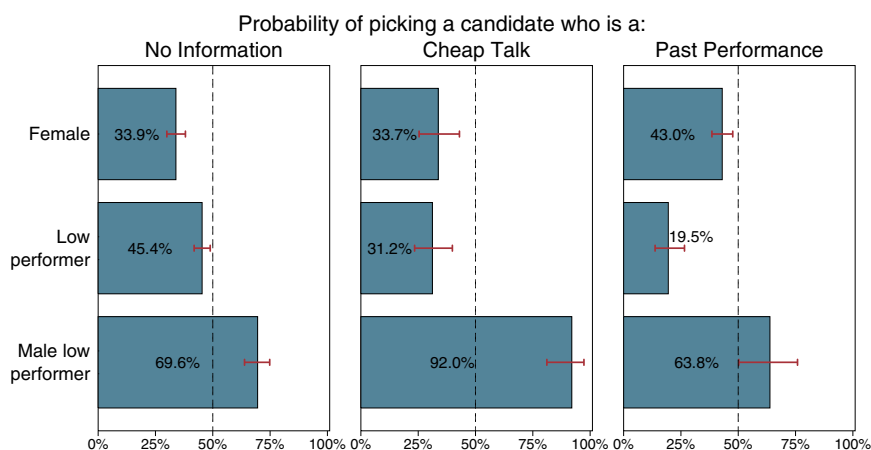


Fig. 1. The top bars show the percentages of female candidates that were picked, and the middle bars show the percentages of times the lower-performing candidate in the pair was picked. This percentage is computed using all the hiring decisions made in each treatment: 507 in the No Information condition, 160 in the Cheap Talk condition, and 265 in the Past Performance condition. The bottom bars show the percentage of times that the chosen candidate was male, conditional on the lower-performing candidate in the pair being chosen (230 cases in the No Information condition, 50 in the Cheap Talk condition, and 47 in the Past Performance condition). Error bars correspond to 95% confidence intervals calculated with regression analysis clustering SEs on employer (*SI Appendix*, Tables S4–S6).

the candidate with better past performance), they would have made the suboptimal choice only 3.4% instead of 8.9% of the time, boosting their earnings by 5.5% (0.198 SDs). In the Cheap Talk condition employers would have earned 7.3% more (0.294 SDs) if they had updated their prior in an unbiased way (optimal updating row in Table 2). Both improvements in earnings are statistically significant ($P < 0.009$) (SI Appendix, Table S17).

Suboptimal hiring decisions were associated strongly with sex bias. If hiring decisions were sex-neutral, the fraction of suboptimal decisions in which a lower-performing male was chosen over a higher-performing female would be close to 50%. We can see that this is not the case (Fig. 1). In all our conditions, suboptimal decisions were made in favor of the male candidate significantly more often than in favor of the female candidate (by at least 13.8 percentage points, $P < 0.046$ based on regression analysis; SI Appendix, Table S6), particularly in the Cheap Talk condition, in which 9 of 10 mistakes were cases in which a lower-performing man was selected over a higher-performing woman.

Hiring choices were consistent with employers' expectations regarding the performance of female and male candidates. Employers overwhelmingly chose the candidate for whom they had higher expectations, irrespective of candidates' sex (SI Appendix, Table S10). Hence, if employers did not have biased expectations in favor of men, there would be no noticeable sex gap in hiring decisions (SI Appendix, Fig. S3). Only on the rare occasions where employers have identical expectations about the performance of the male and female candidates would they tend to favor the male candidate.

Stereotypes and Biased Beliefs. In line with the last finding noted above, we studied how employers' biased expectations were related to stereotype-based prejudices against women. Specifically, we examined the link between employers' hiring biases and their IAT scores. First, we concentrate on employers' expectations when they had no information about candidates other than appearance. Subsequently we present results related to the updating process.

Our IAT-based results show that employers of both sexes associated women less strongly with mathematics than men. Positive scores on our IAT indicate that subjects associate women less with science/math than men; negative scores would suggest the opposite. The mean IAT scores for the men (0.35) and women (0.42) in our sample indicate that employers of both genders had more difficulty associating women with science/math than men. The scores were significantly different from zero

for both genders (t tests, $P < 0.001$). For both men and women, we found a positive correlation between the subjects' own performance in the arithmetic task and their IAT ($r = 0.190$, $P = 0.085$ for men and $r = 0.166$, $P = 0.087$ for women). In other words, both high-performing men and high-performing women associate science/math more with men than with women. Additional analysis of IAT scores is available in SI Appendix.

IAT scores also were related to employers' expectations of candidate performance, with higher scores associated with lower expectations for female candidates. We used regression analysis to test the relationship between employers' expectations about candidates' performance and employers' IAT scores (SI Appendix, Table S12). We found a positive relation between employers' IAT scores and their average expectation of the performance of all evaluated male candidates ($\beta = 1.08$, $P = 0.079$) and a negative relation between IAT scores and the average expected performance of all evaluated female candidates ($\beta = -0.92$, $P = 0.034$). As a result, there was a positive, highly significant relationship between IAT scores and the average expected difference in performance between the evaluated male and female candidates ($\beta = 1.99$, $P = 0.005$). This relationship is plotted in Fig. 2. Interestingly, even individuals with an IAT score of zero display biased expectations. Namely, their expected difference in the performance of men and women is predicted to be positive (biased toward men) and significantly different from zero (by 1.28 sums, $P = 0.002$). This result suggests that the IAT actually may underestimate the level of sex bias. Note, however, that subjects' own IAT scores were not significantly correlated with how much they overestimated their own future performance, for both men ($r = 0.034$, $P = 0.816$) and women ($r = 0.171$, $P = 0.216$).

Updated Beliefs and Subsequent Decisions. People do not rely only on their priors but try to integrate them with any additional relevant information available for decision-making. Hence, we studied the updating process by looking at the employers' subsequent beliefs and choices in the two treatments that allowed the integration of additional information after an initial decision had been made: Decision Then Cheap Talk and Decision Then Past Performance.

To evaluate how employers incorporate new information into their beliefs, we constructed a variable that measures the degree to which an employer i updated expectations about a candidate j after receiving new information about j 's performance: $\varphi_{ij} = (\mu_{ij} - b_{ij}) / (s_j - b_{ij})$. The numerator of φ_{ij} equals i 's expected performance of

Table 2. Degree to which employers update their expectations

	Male candidate		Female candidate		Difference	
	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Decision Then Past Performance						
All employers	0.735	0.038	0.696	0.049	0.038	0.050
Employers with low IAT scores	0.742	0.058	0.715	0.060	0.027	0.055
Employers with high IAT scores	0.732	0.050	0.674	0.077	0.058	0.081
Optimal updating	0.960	0.030	0.901	0.018	0.059	0.038
Decision Then Cheap Talk						
All employers	0.478	0.048	0.620	0.049	-0.142	0.055
Employers with low IAT scores	0.385	0.065	0.617	0.066	-0.232	0.070
Employers with high IAT scores	0.560	0.060	0.610	0.075	-0.050	0.075
Optimal updating	0.884	0.017	1.093	0.046	-0.209	0.048

The degree to which an employer i updates expectations about the performance of a candidate j as measured by $\varphi_{ij} = (\mu_{ij} - b_{ij}) / (s_j - b_{ij})$, where μ_{ij} is i 's updated belief of j 's performance, b_{ij} is i 's prior belief of j 's performance, and s_j is j 's claimed future performance in Decision Then Cheap Talk and j 's past performance in Decision Then Past Performance. The table presents the mean values of φ_{ij} depending on whether candidate j is male or female and the difference between these two values (estimated using regression analysis, see SI Appendix, Tables S14 and S15). The mean values of φ_{ij} are estimated separately for all employers, employers with low IAT scores (below average), and employers with high IAT scores (above average). The mean value of φ_{ij} that corresponds to optimal updating (i.e., the φ_{ij} for which i 's updated belief matches j 's subsequent performance) is also estimated.

Supplementary Information Appendix

Summary of the Important Results

- In all information conditions, there is substantial discrimination against female candidates and this bias is equally present regardless of whether the hiring is done by a man or a woman.
- Employers often make suboptimal hiring decisions across conditions, with the worst decision-making occurring when employers have no information other than the candidates' physical appearance.
- The employers' suboptimal hiring decisions usually occur in favor of a low-performing male candidate at the expense of a high-performing female candidate.
- The cost of discrimination against women is substantial when employers have additional information about the candidates' performance but is negligible if employers observe only the candidates physical appearance.
- Hiring choices are consistent with employers' expectations regarding the performance of female and male candidates, and therefore the gender gap in hiring decisions is due to a systematic underestimation of the performance of women compared to men.
- According to their IAT scores, employers of both genders associate women less strongly with math and science than men.
- There is a positive and highly significant relation between IAT scores and the average expected difference in performance between male and female candidates.
- Employers find candidates' past performance a more reliable signal, and hence more useful information for decision-making, than their self-reported expectation of future performance, but still weight prior beliefs excessively.
- The magnitude of updating of employers' beliefs is not biased by candidate gender when information on past performance is provided by the experimenter—including for employers with high IAT scores.
- Men tend to overestimate their future performance on the arithmetic task, while women underestimate it—a gender difference taken partially into account by employers' updating.
- Employers with a stronger implicit bias against women are more willing to believe men's overestimated expectations of their future performance.

Materials and Methods

Methods: Description of the experiment. The computerized experiment was conducted in 2012 in the laboratory of the Columbia Business School. It was approved by and conducted according to the guidelines of the Institutional Review Board of Columbia University. Subjects were recruited through the schools SONA recruitment website and the experiment was programmed with z-Tree (Fischbacher, 2007). Each session in the experiment lasted 45 minutes.

Upon arrival to the laboratory, subjects read and signed the study's consent form as well as answered a few questions about their demographics, including their race and gender. Thereafter, they were given the experiment's first set of instructions. Subjects were told that the experiment consisted of various parts and that they would be paid their earnings from one randomly-selected part. The total number of parts, P , depended on the number of subjects in the session. Specifically, if there were N subjects in the session, there were $P = (N + 4) / 2$ parts if N was even and $P = (N + 3) / 2$ parts if N was odd.

At this point, subjects read the instructions for part 1. This part consisted of performing sums of four two-digit numbers for four minutes (e.g., $14 + 25 + 79 + 84$). The numbers were randomly generated in the range $[11, 99]$ and the same sequence of random numbers was used for everyone in a session. The subjects' earnings in this task depended on the number of sums they answer correctly. Specifically, they earned \$0 for 5 or fewer sums, \$1 for 6 to 8 sums, \$2, for 9 to 11 sums, \$4 for 12 to 14 sums, \$7 for 15 to 17 sums \$11 for 18 to 20 sums, \$16 for 21 to 23 sums, and \$22 for 24 or more sums.

Once part 1 was complete and subjects were informed of the number of sums they answered correctly, they received the instructions for the remaining parts. In these instructions, subjects were told that they will perform the arithmetic task once again as the last part of the experiment. Moreover, they were told that they will be asked to indicate their expected performance (i.e., number of correct sums) in that task and that their earnings will not be affected by the accuracy of their expected performance. The remaining instructions concerned the intermediate parts of the experiment (i.e., parts 2 to $P-1$). The intermediate parts were identical and are described below. After reading these instructions, we asked subjects to answer a series of questions to ensure their understanding. Once everyone finished answering the control questions, subjects indicated their expected performance in the arithmetic task in the last part of the experiment. Subjects were reminded of their performance in the arithmetic task in the first part of the

experiment when answering this question. Subsequently, subjects completed the intermediate parts of the experiment.

At the beginning of each intermediate part, the computer program selected a pair of subjects to be the candidates in that part, which leaves the remaining subjects with the role of employers (in the instructions we referred to candidates as “contenders” and to employers as “observers”). A subject was a candidate at most once. If the number of subjects in the session was even then everyone got to be a candidate, otherwise one subject was not selected to be a candidate. To form the candidate pairs we used a matching procedure designed to maximize the number of pairs consisting of a randomly selected man and a randomly selected woman. However, since most sessions did not have exactly fifty percent of each gender, some candidate pairs consisted of subjects of the same gender. In other words, if a session consisted of N^M male subjects and N^F female subjects then the number of mixed-gender candidate pairs was $\min\{N^M, N^F\}$, the number of same-gender candidate pairs was $\max\{N^M, N^F\} - \min\{N^M, N^F\}$, and the total number of picking decisions in mixed-gender candidate pairs was $(N^M + N^F - 2) \times \min\{N^M, N^F\}$ if $N^M + N^F$ was even and $(N^M + N^F - 1) \times \min\{N^M, N^F\}$ if $N^M + N^F$ was odd. To avoid priming subjects about gender discrimination, we did not inform them of the precise details of the pairing procedure.

Candidates were randomly assigned to a sign that reads “Contender A” or “Contender B” and were asked to hold their sign in the front of the room. Employers were asked to look at the candidates before making their decisions. Employers made two decisions in the Cheap Talk and Past Performance treatments and four decisions in the Decision Then Cheap Talk and Decision Then Past Performance treatments. The first two decisions were made simultaneously on the screen as where the third and fourth decisions in the latter treatments. Subjects never received feedback concerning the choices of others.

The first and third decisions consisted of picking one of the two candidates. The second and fourth decisions consisted of guessing the number of sums each candidate will answer correctly when they perform the arithmetic task in the last part of the experiment. If a given part was selected for payment, earnings were determined as follows. The earnings of candidates depended on the choice of one randomly selected employer. Specifically, the candidate picked by the employer earns \$8 whereas the other candidate earns \$4. In order to avoid hedging between decisions, the earnings of employers were determined by randomly selecting one of their

Table S1. For each session, the table shows the number of subjects, the number of mixed-gender candidate pairs, the number of employer observations in mixed-gender candidate pairs, and the treatment they participated in.

	Subjects	Mixed-gender candidate pairs	Picking decisions in mixed-gender candidate pairs	Treatment
Session 1	18	5	80	Decision Then Cheap Talk
Session 2	18	5	80	Cheap Talk
Session 3	18	6	96	Decision Then Past Performance
Session 4	19	9	153	Past Performance
Session 5	17	7	105	Decision Then Cheap Talk
Session 6	10	4	32	Decision Then Past Performance
Session 7	10	4	32	Decision Then Past Performance
Session 8	10	5	40	Past Performance
Session 9	10	4	32	Past Performance
Session 10	10	5	40	Cheap Talk
Session 11	16	6	84	Decision Then Cheap Talk
Session 12	15	6	78	Decision Then Past Performance
Session 13	10	5	40	Cheap Talk
Session 14	10	5	40	Past Performance

decisions. If the first or third decision was selected then their earnings depend on the performance of the candidate they picked in the second arithmetic task (they earned \$0 for 5 or fewer sums, \$1 for 6 to 8 sums, \$2, for 9 to 11 sums, \$4 for 12 to 14 sums, \$7 for 15 to 17 sums \$11 for 18 to 20 sums, \$16 for 21 to 23 sums, and \$22 for 24 or more sums). If the second or fourth decision was selected then employers earned between \$0 and \$9 depending on how accurately they estimated the candidates' performance. For each guess, employers earned \$4.50 if the absolute difference between the their guess and the candidate's actual performance was 0 sums, \$4.38 if this difference was 1 sum, \$4.00 if it was 2 sums, \$3.38 if it was 3 sums, \$2.50 if it was 4 sums, \$1.38 if it was 5 sums, and \$0.00 if it was 6 or more sums (these payment schedule incentivizes a risk-neutral individual to reveal the mean of their distribution). Note that, by eliciting separately the employers' expectations from their candidate choice, we are able to observe whether employers have significant taste-based motivations for choosing a candidate—that is, they are willing to sacrifice their earnings by choosing the candidate with the lower expected performance in order to increase that candidate's expected earnings.

Once the intermediate parts had finished, subjects did the arithmetic task again as the last part of the experiment. Thereafter, we randomly selected a part to be paid. If the part to be paid was not the first or the last, we also randomly selected the decision to be paid. As a final step, we

Table S2. Sequence of blocks used in the IAT.

Block	Number of trials	Purpose	Left-key category-attribute	Right-key category-attribute
1	20	Practice	male	female
2	20	Practice	math and science	liberal arts
3	20	Practice	male-math and science	female-liberal arts
4	40	Test	male-math and science	female-liberal arts
5	20	Practice	female	male
6	20	Practice	female-math and science	male-liberal arts
7	40	Test	female-math and science	male-liberal arts

asked all subjects to complete an Implicit Association Test (IAT) between gender and science and math (see the description below). Thereafter, they were paid their earnings and dismissed.

In total, 191 undergraduate students (83 men and 108 women) participated in 14 sessions. We have 94 pairs of candidates, of which 76 are mixed-gender pairs (subjects observed an average of 4.88 mixed-gender pairs). For each session, Table S1 presents the number of subjects, the number of mixed-gender candidate pairs, the number of employer observations of mixed-gender candidate pairs, and the treatment they participated in. None of the subjects had participated in a similar experiment. Average earnings, including the \$8 show-up fee, were approximately \$20.

Methods: Implicit association test. We used the IAT (10) as an indirect measure of associations between the categories “male” and “female” and the attributes “math and science” and “liberal arts.” Specifically, subjects observed a screen where either a picture or a word appears and were asked to respond rapidly by pressing a right-hand key if the picture/word corresponded to one category or attribute (e.g., “male” and “liberal arts”) and a left-hand key if the picture/word corresponded to the other category or attribute (e.g., “female” and “math and science”). The words used for “math and science” were “physics,” “engineering,” “chemistry,” “biology,” “statistics,” “geometry,” “calculus,” and “algebra,” and the words used for “liberal arts” were “literature,” “music,” “philosophy,” “writing,” “history,” “arts,” “civics,” and “humanities.” Pictures are not reproduced here due to copyright but are available upon request. Subjects performed various trials of this task under different side-category-attribute combinations (see Table S2). Fig. S1 provides a sample screenshot of the IAT.



Fig. S1. Screenshot of the IAT.

The IAT score of each subject was constructed by comparing response times in the classification task. The IAT score is interpreted as a measure of association strengths by assuming that subjects respond more rapidly when the category and attribute on a given side are strongly associated than when they are weakly associated. For example, subjects that were faster when they have to press the same key for male faces and math/science words than when they have to press the same key for female faces and math/science words were classified as having an implicit association between math/science and males relative to females.

We computed the IAT score of each subject according to the scoring algorithm described in (Greenwald, Nosek, and Banaji, 2003). In short, first, we dropped the trials in which the response time is either too short (less than 0.1 seconds) or too long (more than 10 seconds). Of all the subjects, 97% (93%) answered 119 (120) of the 120 IAT trials within the suggested response times. Our results remain unaffected if we drop from the statistical analysis the few subjects with less than 119 trials. Second, we calculated the mean difference in response times between trials in blocks 6 and 3, $DIFF_{6-3}$, and between trials in blocks 7 and 4, $DIFF_{7-4}$. Third, we calculated the standard deviation in response times for all trials in blocks 3 and 6, SD_{6+3} , and in blocks 4 and 7, SD_{7+4} . A subject's IAT score is given by $\frac{1}{2}(DIFF_{6-3}/SD_{6+3} + DIFF_{7-4}/SD_{7+4})$, which results in a number between -2 and 2 . A positive score indicates an association of "male" with "math and science" and "female" with "liberal arts." Conversely, a negative score indicates an association of "female" with "math and science" and "male" with "liberal arts."

Materials: Instructions for the experiment. We provide the instructions of the Decision Then Cheap Talk treatment. The instructions of other treatments are available upon request. Subjects completed the first part of the experiment before they received the rest of the instructions.

Welcome

Thank you for participating in today's study. The study will last around 45 minutes. You are not allowed to communicate with other participants. If you have a question, raise your hand and we will gladly help you. For your participation you will receive an \$8 *show-up fee*. In addition, you will be able to earn more money. How you do this is described in these instructions. Please read them carefully.

The study is divided into *various parts*, none of which takes more than 5 minutes. At the end of the study we will randomly select *one* of the parts and pay you based on your performance in that part. Before each part starts, we will describe in detail how your payment is determined in that part.

Instructions for part 1

In part 1, you can earn money by performing a series of sums of four randomly-chosen two-digit numbers (e.g., $15 + 73 + 49 + 30$). Calculators are not allowed. You will have *four minutes* to answer as many sums as possible. The computer will record the number of sums that you answer correctly to determine your earnings. Your earnings do not decrease if you provide an incorrect answer to a sum.

The screen where you do the sums looks like the one below. You submit your answer by clicking on Submit. As soon as you submit your answer you will be told if it was correct or incorrect. You can also see the total number of sums you have answered correctly. At the bottom, you see how many seconds you have left. In order to familiarize yourself with the screen you will have a 30 second trial period in which you can practice adding sums. The trial period does not affect your earnings.

Sum 6: $41 + 80 + 80 + 66$

Your last answer was: **Correct**

Number of correct answers: **5**

Seconds left: **196**

Note that everyone in the room receives the same sequence of randomly generated sums. That is, everyone faces the same level of difficulty. If part 1 is the part randomly selected for payment, then your earnings are given by the table below.

Number of sums you answered correctly	Your earnings
less than 5 sums	\$0.00
between 6 and 8 sums	\$1.00
between 9 and 11 sums	\$2.00
between 12 and 14 sums	\$4.00
between 15 and 17 sums	\$7.00
between 18 and 20 sums	\$11.00
between 21 and 23 sums	\$16.00
more than 24 sums	\$22.00

If you have any questions please raise your hand. Otherwise you can click the button on your screen.

Instructions for the last part of the study

For reasons that will be obvious, the last part of the study is described now. The last part of the study is identical to part 1. That is, you will have another four minutes to answer sums. The computer will record the number of sums that you answer correctly. Your payment does not decrease if you provide an incorrect answer to a sum. If the last part of the study is the part randomly selected for payment, then your earnings are given by the same table as in part 1.

Stating your expected performance

Your first task after reading these instructions will be to provide an answer to the following question: “*Indicate the number of sums you expect to answer correctly when you perform in the last part of the study.*” You can answer the question with any number. Moreover, your earnings in the study will not be affected by the accuracy of the submitted number.

Instructions for the remaining parts

The remaining parts of the study are all identical. At the beginning of each part, two participants in the room will be selected by the computer through a random procedure. We will refer to these two participants as *contender A* and *contender B*. We will refer to the rest of you as *observers*. Each participant gets to be a contender at most once during the study. Contenders will be asked to stand up and hold a piece of paper indicating their label (A or B).

Observers

In each part, observers make *four decisions*. Decisions consist of either: (i) accurately guessing the number of sums that each contender will answer correctly, or (ii) picking one of the contenders.

If a given part is selected for payment, *one of the four decisions* in that part will be picked at random to determine your final payment. Each decision is explained in detail below.

Decisions 1 and 2

If you are an observer, you will make decisions 1 and 2 on the following screen:

Decision 1

The number of sums that **contender A** will answer correctly is:

The number of sums that **contender B** will answer correctly is:

If this decision is selected for payment, you will earn money depending on the accuracy of these guesses.

Decision 2

My pick for decision 2 is: ☐ Contender A
☐ Contender B

If this decision is selected for payment, you will earn money depending on the performance of your pick.

Submit

On the top part of the screen, you make *decision 1*. This decision consists of guessing the number of sums that each contender will answer correctly when they take part in the *last part* of the study. Your earnings depend on the accuracy of your guesses according to the table below.

Difference between your guess and the number of sums answered correctly	Earnings for your guess (per contender)
0 sums away (exact answer)	\$4.50
1 sum away	\$4.38
2 sums away	\$4.00
3 sums away	\$3.38
4 sums away	\$2.50
5 sums away	\$1.38
6 sums away or more	\$0.00
0 sums away (exact answer)	\$4.50

On the bottom part of the screen, you make *decision 2*. This decision consists of *picking one of the two contenders*. Your earnings depend on the performance in the *last part* of the study of

the contender that you picked. Specifically, your earnings are given by the same table as in part 1, which we reproduce below for your convenience.

Number of sums answered correctly by the contender you pick	Your earnings
less than 5 sums	\$0.00
between 6 and 8 sums	\$1.00
between 9 and 11 sums	\$2.00
between 12 and 14 sums	\$4.00
between 15 and 17 sums	\$7.00
between 18 and 20 sums	\$11.00
between 21 and 23 sums	\$16.00
more than 24 sums	\$22.00

Decisions 3 and 4

If you are an observer, you will make decisions 3 and 4 on the screen below.

On the top part of the screen, you make *decision 3*. You are asked once again to guess the number of sums that each contender will answer correctly when they take part in the last part of the study. Your earnings depend on the accuracy of your guesses according to the same table as in decision 1. Note that, unlike in decision 1, you can also see the *answers submitted by each contender to the question asking for their expected performance*.

On the bottom part of the screen, you make decision 4. Again, you are asked to *pick one* of the two contenders, and your earnings depend on the performance in the *last part* of the study of the contender that you picked according to the same table as in decision 2 (and part 1).

Decision 3

Contender A estimates he/she will answer 12 sums correctly.

The number of sums that **contender A** will answer correctly is:

Contender B estimates he/she will answer 12 sums correctly.

The number of sums that **contender B** will answer correctly is:

If this decision is selected for payment, you will earn money depending on the accuracy of these guesses.

Decision 4

My pick for decision 4 is: ☐ Contender A
☐ Contender B

If this decision is selected for payment, you will earn money depending on the performance of your pick.

Earnings of Contenders

The earnings of contenders in these remaining parts of the study depend on whether they are picked by observers. Specifically, *one observer* will be selected at random to determine the earnings of the contenders. If the observer picked contender A, then contender A earns \$8.00 and contender B earns \$4.00, and conversely, if the observer picked contender B, then contender A earns \$4.00 and contender B earns \$8.00. Lastly, if decisions 1 or 2 are used to determine payments then the earnings of the contenders are determined by decision 2 and if decisions 3 or 4 are used for payment then the earnings of contenders are determined by decision 4.

Example of how to calculate earnings

Suppose that you are an observer in the part that is picked for payment. Furthermore, in decision 1 you guessed that contender A will answer 10 sums correctly and contender B will answer 15 sums correctly. In decision 2 you picked contender B.

If it turns out that contender A answered 8 sums correctly and contender B answered 12 sums correctly, then:

- If decision 1 is selected for payment, your earnings would be: \$4.00 for your guess of A's performance + \$3.38 for your guess of B's performance + the \$8.00 show-up fee = \$15.38.
- If decision 2 is selected for payment, your earnings would be: \$4.00 for picking a contender that answered 12 sums + the \$8.00 show-up fee = \$12.00.
- For the earnings of contenders, suppose that you are the observer chosen to determine the contenders' earnings. In this case, Contender B's earnings would be: \$8.00 for being picked by you + the \$8.00 show-up fee = \$16.00, and contender A's earnings would be: \$4.00 for not being picked by you + the \$8.00 show-up fee = \$12.00.

Final note

Note that when they perform the sums in the last part of the study, contenders will *not know* how many observers have picked them. This will be revealed after they finished answering sums. Moreover, contenders will not know at any point what the guesses of the observers were.

If you have any questions please raise your hand. Otherwise you can click the button on your screen.

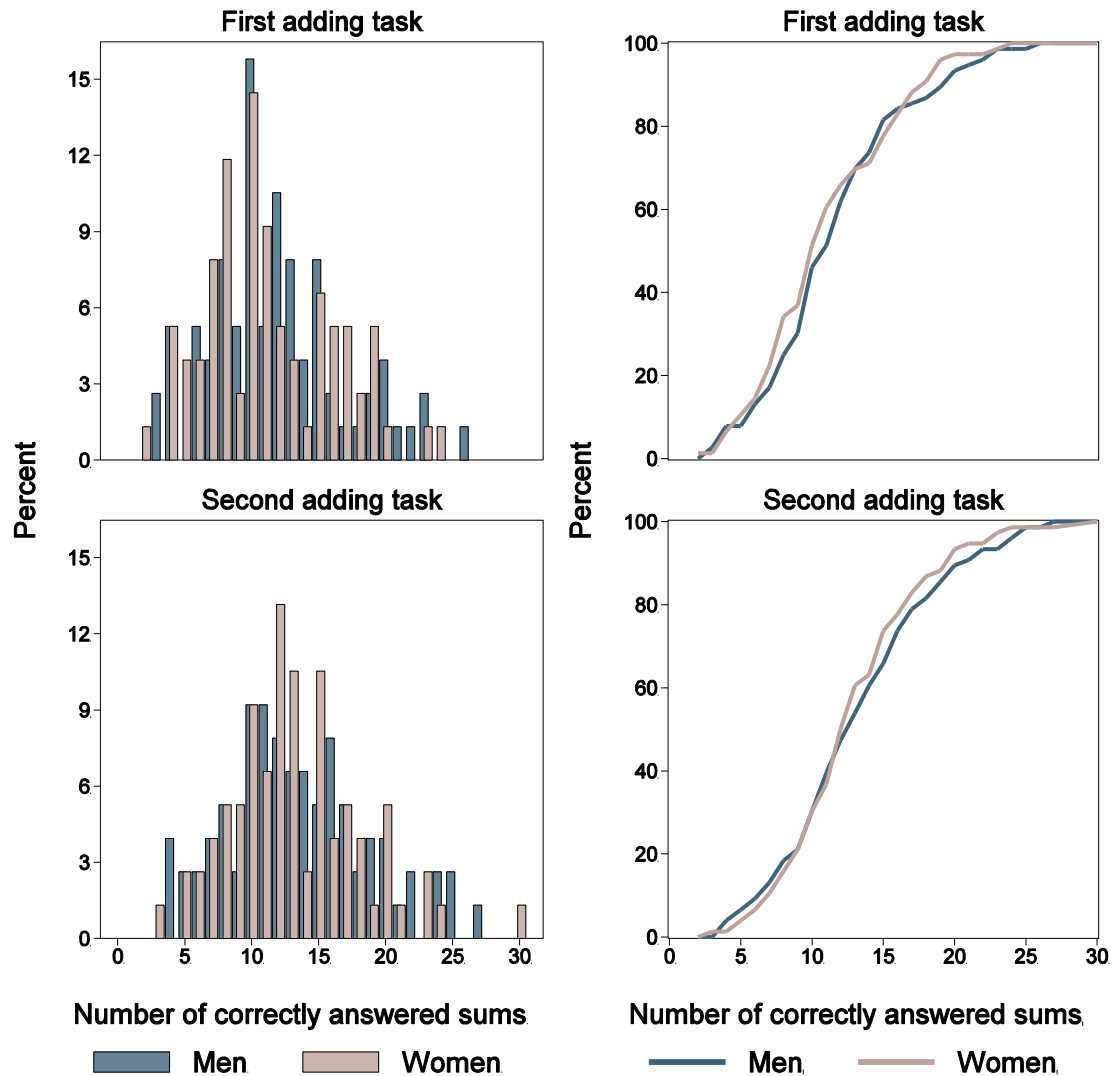


Fig. S2 The bars show the distribution of the subjects' performance in the two arithmetic tasks depending on their gender. The lines show the corresponding cumulative distributions.

Supplementary Data Analysis

Here, we provide the statistical analysis supporting the claims in the main body of the paper. Note that all *P*-values in the main body of the paper and in this document are from two-tailed tests. The data analysis was done with the statistics software STATA version 13.1. The executable file that performs the analysis as well as the dataset (in excel format) is available with these supplementary materials.

Performance in the arithmetic tasks. Figure S2 shows the similarity between the distributions of the men's and women's performance. In the first arithmetic task, the average number of

Table S3. Means, by information condition and treatment, for the: fraction of picked candidates that are female, fraction of picked candidates that had the lower performance, and fraction of picked candidates with the lower performance that are male.

	Probability of picking a:		
	Female	Low performer	Male low performer
No Information	0.339	0.454	0.696
Cheap Talk	0.338	0.313	0.920
Past Performance	0.430	0.196	0.638
Decision Then Cheap Talk	0.320	0.338	0.857
Decision Then Past Performance	0.391	0.118	0.821

correctly answered sums is 11.86 for men and 11.28 for women. We do not reject the null hypothesis that the distributions of men and women significantly differ with a Mann-Whitney U test ($P = 0.464$) or a Kolmogorov-Smirnov test ($P = 0.887$). The standard deviation in the performance of men is slightly higher (5.02 vs. 4.83), but the difference is not statistically significant (Conover's squared ranks test, $P = 0.724$). In the second arithmetic task, on average, men answered correctly 13.50 sums and women 13.17 (standard deviations equal 5.40 and 4.89, respectively). Once again, we do not find statistically significant differences between men and women (Mann-Whitney U test, $P = 0.727$; Kolmogorov-Smirnov test $P = 0.973$; Conover's squared ranks test, $P = 0.222$). Wilcoxon signed-rank tests indicate that both genders significantly improve their performance between the first and second arithmetic task ($P < 0.001$ for both men and women), but we do not find a significant difference between the men's improvement and the women's improvement (Mann-Whitney U test, $P = 0.563$).

Statistical analysis of the employers' decision. In this section, we use regression analysis to compare the employers' decisions across the different conditions. We compare three different variables. The first is the fraction of picked candidates that are female, the second is the fraction of picked candidates that had the lower performance in the second arithmetic task, and the third is the fraction of picked candidates with lower performance that are male. Table S3 contains the mean for each of these three variables in each information condition and treatment.

We use regression analysis to make the statistical comparisons. Since the three variables are binary, employers make multiple decisions, and they are randomly assigned to treatments, we use probit regressions with employer random effects. In all regressions, we use dummies indicating the *information conditions* as independent variables, using the No Information condition as the omitted group, and robust standard errors clustered on individual employers. We run four different regressions for each dependent variable. In the first regression, labeled

Table S4. Probit regressions with picking a female candidate as the dependent variable. The top panel reports marginal effects, robust standard errors in parenthesis. All regressions contain employer random effects. *, **, and *** denote significance at the 10%, 5%, and 1% level. The middle panel reports *P*-values from various hypotheses tests. The bottom panel indicates the number of observations and employers.

	Between	Within	Between II	Within II
Cheap talk	−0.002 (0.051)	−0.020 (0.026)	0.000 (0.051)	−0.020 (0.026)
Past performance	0.091*** (0.031)	0.051* (0.028)	0.095*** (0.032)	0.051* (0.028)
Female employer			0.031 (0.031)	0.001 (0.028)
(a) P(Cheap talk) = P(Past performance)	0.076	0.017	0.067	0.016
(b) P(No information) = 0.5	0.000	0.000	0.000	0.000
(c) P(Cheap talk) = 0.5	0.001	0.000	0.001	0.000
(d) P(Past performance) = 0.5	0.003	0.000	0.005	0.000
(e) Joint significance of all variables	0.009	0.050	0.023	0.103
Number of observations	932	1014	932	1014
Number of employers	191	104	191	104

“Between,” we make between-subjects comparisons. In other words, the regressions are run with the data from the Cheap Talk and Past Performance treatments plus the data from the No Information condition in the Decision Then Cheap Talk and Decision Then Past Performance treatments. In the second regression, labeled “Within,” we make within-subjects comparisons. In other words, the regressions are run with all the picking decisions in the Decision Then Cheap Talk and Decision Then Past Performance treatments. The third and fourth regressions, labeled “Between II” and “Within II,” mirror the first two expect that, in addition to the information conditions, we control for the gender of the employer. Besides the estimated marginal effects, we also report the *P*-values of the following hypotheses tests: (a) whether the coefficient of Cheap Talk equals that of Past Performance; (b)-(d) in each condition, whether the predicted probability for the dependent variable equals the benchmark of fifty percent; and lastly, (e) whether all independent variables are jointly significant.

Table S4 presents estimated marginal effects when the dependent variable is 0 if a male candidate is picked and 1 if a female candidate is picked. In all regressions in Table S4, the probability of picking a female candidate is almost identical between the No Information and Cheap Talk conditions and is significantly higher in Past Performance. Moreover, in all three conditions, the probability of picking a female candidate is significantly less than the no-discrimination benchmark of fifty percent. Note that we use fifty percent as the benchmark

Table S5. Probit regressions with picking the low performing candidate as the dependent variable. The top panel reports marginal effects, robust standard errors in parenthesis. All regressions contain employer random effects. *, **, and *** denote significance at the 10%, 5%, and 1% level. The middle panel reports *P*-values from various hypotheses tests. The bottom panel indicates the number of observations and employers.

	Between	Within	Between II	Within II
Cheap talk	-0.131*** (0.045)	-0.105*** (0.032)	-0.130*** (0.045)	-0.105*** (0.032)
Past performance	-0.250*** (0.043)	-0.320*** (0.039)	-0.249*** (0.044)	-0.319*** (0.039)
Female employer			0.012 (0.034)	0.022 (0.030)
(a) P(Cheap talk) = P(Past performance)	0.031	0.000	0.033	0.000
(b) P(No information) = 0.5	0.010	0.010	0.010	0.010
(c) P(Cheap talk) = 0.5	0.000	0.000	0.000	0.000
(d) P(Past performance) = 0.5	0.000	0.000	0.000	0.000
(e) Joint significance of all variables	0.000	0.000	0.000	0.000
Number of observations	932	1014	932	1014
Number of employers	191	104	191	104

because that is the ratio one obtains if there is no discrimination. However, one could argue that the right benchmark is the probability that a randomly chosen woman performs better than randomly chosen man. Using the distribution from the second arithmetic task to calculate this probability gives 48.4%. In all regressions, the fraction of female candidates is significantly less than this probability (Wald tests, $P < 0.022$). These results are robust to controlling for the gender of the employer. Moreover, the probability of picking a female candidate is not significantly different for female employers.

Table S5 presents estimated marginal effects when the dependent variable is 0 if the candidate with the higher performance in the second arithmetic task is picked and 1 if the candidate with the lower performance is picked. In all regressions in Table S5, the probability of picking the low-performing candidate is significantly lower in No Information, followed by Cheap Talk, and is significantly higher in Past Performance. In all three conditions, the probability of picking the low-performing candidate is significantly less than 50%. These results are robust to controlling for the gender of the employer and that the probability of picking the low-performing candidate is not significantly different for female employers.

Table S6 presents estimated marginal effects when the dependent variable is 0 if the female candidate is picked and 1 if the male candidate is picked and the data is restricted to the decisions where the employer picked the low-performing candidate. In all regressions in Table S6, the

Table S6. Probit regressions with picking a male candidate given that the low performing candidate was picked as the dependent variable. The top panel reports marginal effects, robust standard errors in parenthesis. All regressions contain employer random effects. *, **, and *** denote significance at the 10%, 5%, and 1% level. The middle panel reports *P*-values from various hypotheses tests. The bottom panel indicates the number of observations and employers.

	Between	Within	Between II	Within II
Cheap talk	0.229*** (0.071)	0.155*** (0.045)	0.228*** (0.072)	0.155*** (0.045)
Past performance	-0.054 (0.066)	0.112* (0.065)	-0.055 (0.067)	0.113* (0.065)
Female employer			-0.015 (0.050)	0.007 (0.053)
(a) P(Cheap talk) = P(Past performance)	0.001	0.640	0.001	0.640
(b) P(No information) = 0.5	0.000	0.000	0.000	0.000
(c) P(Cheap talk) = 0.5	0.000	0.000	0.000	0.000
(d) P(Past performance) = 0.5	0.046	0.000	0.049	0.000
(e) Joint significance of all variables	0.003	0.001	0.007	0.002
Number of observations	327	349	327	349
Number of employers	158	103	158	103

probability of picking a male low-performing candidate is significantly higher in Cheap Talk than in No Information. More importantly, in all three conditions the probability of picking a male low-performing candidate is significantly more than the no-discrimination benchmark of 50%. These results are robust to controlling for the gender of the employer and that the probability of picking a male low-performing candidate is not significantly different for female employers.

Next, we demonstrate that we obtain very similar results with nonparametric tests. To run the non-parametric tests we first calculate the mean per employer for each of the three dependent variables then use these means as observations. Table S7 presents the *P*-values of: (a)-(c) pairwise comparisons between the various conditions using Mann-Whitney U tests for between-subjects comparisons and Wilcoxon signed-rank tests for within-subject comparisons; (d)-(f) for each condition, a comparison with the 50% benchmark using Wilcoxon signed-rank tests; and (g) a Kruskal-Wallis equality-of-populations rank test. Table 1 in the main body of the paper shows the number of independent observations in each condition (i.e., the number of subjects per treatment).

Analysis of the employers' expectations. Here, we evaluate whether discrimination against female candidates in the picking decision is explained by biases in the employers' expectations. Table S8 presents descriptive statistics of the following two variables: employer *i*'s expected

Table S7. *P*-values of: (a)-(c) pairwise comparisons between conditions using Mann-Whitney U tests for between-subjects comparisons and Wilcoxon signed-rank tests for within-subject comparisons; (d)-(f) comparisons to the 50% benchmark using Wilcoxon signed-rank tests; and (g) a Kruskal-Wallis equality-of-populations rank test.

	Probability of picking a:					
	Female		Low performer		Male low performer	
	Between	Within	Between	Within	Between	Within
(a) No Information = Cheap talk	0.641	0.710	0.001	0.006	0.003	0.020
(b) No Information = Past performance	0.061	0.210	0.001	0.001	0.922	0.447
(c) Cheap talk = Past performance	0.074	0.007	0.054	0.001	0.032	0.951
(d) No information = 0.5	0.001		0.063		0.001	
(e) Cheap talk = 0.5	0.002	0.001	0.001	0.001	0.001	0.001
(f) Past performance = 0.5	0.001	0.003	0.001	0.001	0.003	0.002
(g) Equality of populations	0.033	0.046	0.008	0.001	0.002	0.002

performance of candidate j , denoted as e_{ij} , depending on whether j is male or female; and the fraction of times i expects j will perform better than the other candidate k , denoted as $e_{ij} > e_{ik}$ (recall that j and k are always of different gender).

To test whether there is a significant difference between male and female candidates, we run regressions with employer \times treatment fixed effects. We use with a dummy variable indicating the gender of the candidate interacted with dummies indicating the *information conditions* as independent variables and robust standard errors clustered on individual employers. We run a regression for each variable in Table S8 (a GLS regression for the first variables and a logit regression for the second). The results are presented in Table S9. In all information conditions, male candidates are expected to outperform female candidates significantly more often than the converse ($P < 0.023$). These results remain unaffected if we use nonparametric tests (available upon request).

Table S10 describes the relation between the employers' expectations and their picking choice. For a pair of candidates j and k , it shows the fraction of times j is picked given that j is expected to perform better, equal, or worse than k . Employers overwhelmingly pick candidates who they think will have a strictly higher performance irrespective of their gender. It is only in cases where there is a tie in expected performance that we see employers favoring male candidates. However, given that ties are not expected very often, this effect is bound to be relatively minor in explaining the gender gap in picking decisions compared to the bias in expectations.

Table S8. Descriptive statistics of (e_{ij}) employer i 's expected performance of candidate j and $(e_{ij} > e_{ik})$ the fraction of times i expects j will perform better than the other candidate k .

Variable	Statistic	No Information		Cheap Talk		Past Performance		Decision Then Cheap Talk		Decision Then Past Performance	
		Male	Female	Male	Male	Male	Female	Female	Female	Male	Female
e_{ij}	mean	13.002	10.941	11.625	12.219	12.642	11.581	13.736	11.703	12.000	11.571
	median	13.000	11.000	12.000	11.000	12.000	11.000	13.000	12.000	12.000	11.000
	std. dev.	5.073	5.066	3.077	4.319	5.340	3.597	5.611	4.226	4.576	4.680
	Cohen's d	0.407		-0.159		0.233		0.410		0.093	
$e_{ij} > e_{ik}$	mean	0.625	0.318	0.575	0.356	0.521	0.392	0.602	0.316	0.563	0.391
	std. dev.	0.485	0.466	0.496	0.480	0.500	0.489	0.490	0.466	0.497	0.489
	Cohen's d	0.648		0.449		0.260		0.600		0.350	

Table S9. Regressions with the following dependent variables: (e_{ij}) employer i 's expected performance of candidate j ; and $(e_{ij} > e_{ik})$ the fraction of times i expects j will perform better than the other candidate k . GLS (first variable) and logit (last variable) regressions with employer \times treatment fixed effects. Robust standard errors in parenthesis. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Independent variables	Dependent variable	
	e_{ij}	$e_{ij} > e_{ik}$
No Information \times female	-2.061*** (0.309)	-1.154*** (0.163)
Cheap Talk \times female	0.594 (0.548)	-0.795** (0.348)
Past Performance \times female	-1.060*** (0.321)	-0.481** (0.206)
Decision Then Cheap Talk \times female	-2.033*** (0.438)	-1.080*** (0.192)
Decision Then Past Performance \times female	-0.429 (0.279)	-0.624*** (0.162)
F or χ^2 statistic	12.740***	83.635***
Number of observations	2878	2878
Number of employers	191	191

Table S10. Fraction of times a candidate j is picked given that j is expected to perform better ($e_{ij} > e_{ik}$), equal ($e_{ij} = e_{ik}$), or worse ($e_{ij} < e_{ik}$) than the other candidate in the pair k .

Expectations	No Information		Cheap Talk		Past Performance	
	j is male	j is female	j is male	j is female	j is male	j is female
$e_{ij} > e_{ik}$	98.4%	96.3%	96.7%	87.7%	92.0%	91.3%
$e_{ij} = e_{ik}$	58.6%	41.4%	90.9%	9.1%	65.2%	34.8%
$e_{ij} < e_{ik}$	3.7%	1.6%	12.3%	3.3%	8.7%	8.0%

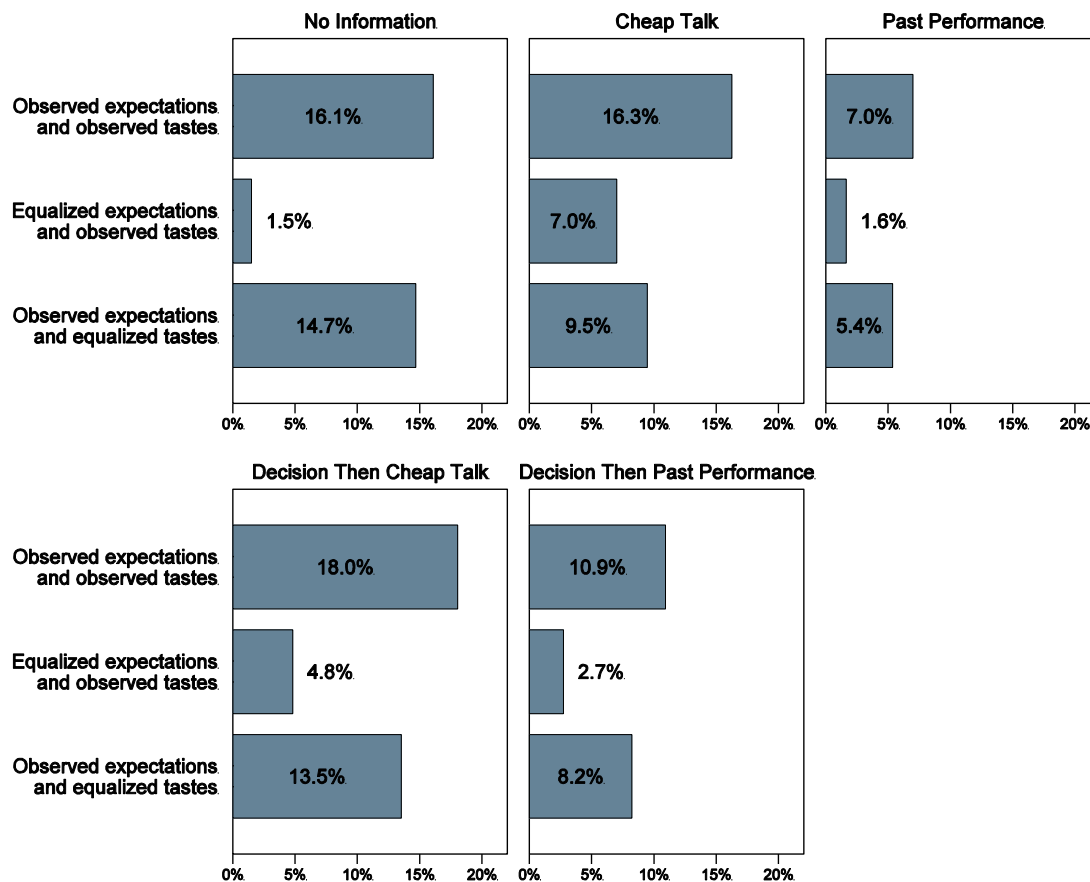


Fig. S3 Fraction of picked male candidates minus the fraction of picked female candidates in each information condition if: (top) there are gender differences in expectations and in picking; (middle) there are no gender differences in expectations but there are differences in picking; and (bottom) there are no gender differences in picking but there are differences in expectations.

To illustrate the impact of expectations on the picking decision, we perform the following exercise. In each information condition, we simulate what the gender gap in picking decisions would be in the following two scenarios: (a) employers assign male and female candidates the same probability of being the higher performer, but for a given belief, they pick male and female candidates based on the observed frequencies in Table S10; and (b) employers pick male and female candidates with the same probability for a given belief, but their belief of which candidate is the higher performer is given by the observed frequencies in Table S9. The results are displayed in Figure S3. The top bars show the observed gender gap in picking decisions. The middle bars show the gender gap if there are no gender differences in expectations but there are differences in picking (scenario a), while the bottom bars show the gender gap if there are no gender differences in picking but there are differences in expectations (scenario b). In all

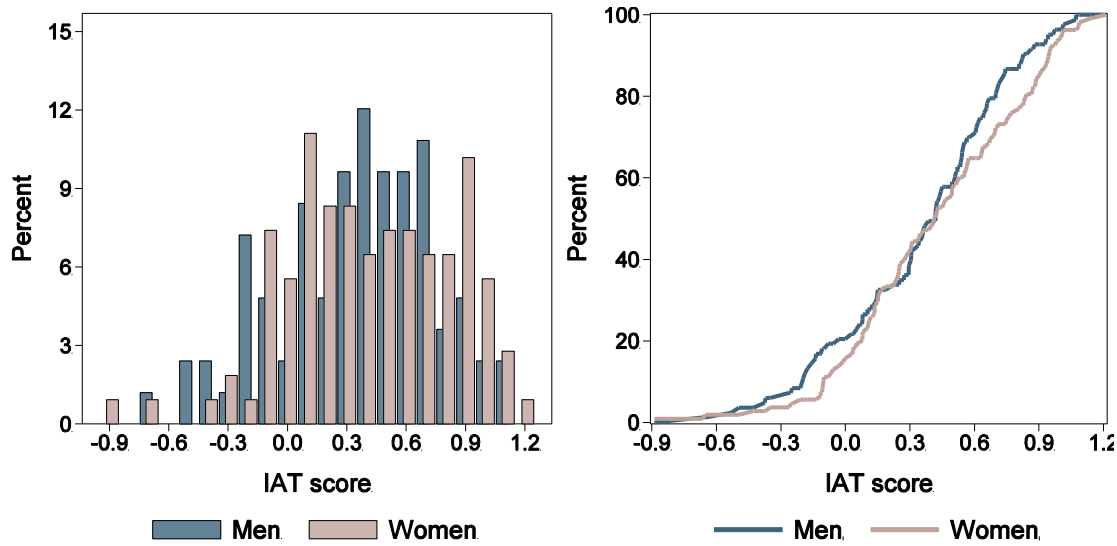


Fig. S4 The bars show the distribution of the employers' IAT score depending on their gender. The lines show the corresponding cumulative distributions.

information conditions, eliminating differences in expectations substantially decreases the gender gap in picking decisions. By contrast, eliminating differences in picking has a noticeable effect only in Cheap Talk and it does not affect the existence of a substantial gender gap in all information conditions. In other words, discrimination against female candidates is largely driven by differences in their expected performance.

Analysis of IAT scores and the pickers' prior beliefs. Figure S4 displays the distribution of the subjects' IAT scores by gender (descriptive statics are available in Table S11). We do not reject the null hypothesis that the distributions of men and women significantly differ with a two-sample t test ($P = 0.267$) or a Kolmogorov-Smirnov test ($P = 0.312$). A variance ratio test finds no significant difference in standard deviations ($P = 0.725$). One-sample t tests indicate that the mean IAT is significantly higher than zero ($P < 0.001$ for both men and women), indicating a stronger association of males with math and science than females.

Table S12 contains the OLS regressions associating the employers' IAT score with their prior beliefs. Specifically, in the first regression, the dependent variable is employer i 's mean expected performance of all male candidates in the No Information condition. In the second regression, the dependent variable is i 's mean expected performance of all female candidates in the No Information condition. In the third regression, the dependent variable is i 's mean difference in the expected performance of the male and female candidates across all pairs in the No Information condition. All regressions use i 's IAT score as the independent variable and

Table S11. Descriptive statistics of the subjects' IAT scores.

Statistic	All	Male	Female
mean	0.387	0.350	0.416
median	0.419	0.420	0.416
std. dev.	0.409	0.400	0.415
Cohen's d		-0.162	

Table S12. OLS regressions with the following dependent variables: the mean expected performance of male candidates in the No Information condition; the mean expected performance of female candidates in the No Information condition; and the mean difference in performance between male and female candidates. All regressions display robust standard errors in parenthesis. *, **, and *** denote significance at the 10%, 5%, and 1% level.

Independent variables	Dependent variable		
	Males	Females	Males – Females
IAT score	1.120* (0.674)	-0.896** (0.419)	2.016*** (0.710)
Constant	12.521*** (0.422)	11.321*** (0.273)	1.200** (0.466)
R ²	0.031	0.029	0.070
Number of observations	104	104	104

robust standard errors. The predicted association between IAT scores and prior beliefs of the first two regressions is visualized in Figure S5, and the predicted association of the third regression is visualized in Figure 2 in the main body of the paper. We obtain similar results by calculating correlation coefficients between the employers' IAT score and their mean expected performance for male candidates ($r = 0.177$, $P = 0.072$), female candidates ($r = -0.170$, $P = 0.084$), and the difference between male and female candidates ($r = 0.265$, $P = 0.007$).

Analysis of the candidates' expectations. Table S13 presents descriptive statistics of the following three variables: candidate j 's expected performance in the second arithmetic task, denoted as e_{2j} , depending on whether j is male or female; the difference between j 's expectation and j 's performance in the first arithmetic task, denoted as $e_{2j} - y_{1j}$; and the difference between j 's expectation and j 's performance in the second arithmetic task, denoted as $e_{2j} - y_{2j}$. Table S14 presents P -values from Mann-Whitney U tests comparing the distributions of male and female candidates for each of these variables and in each treatment.

Analysis of how employers update their expectations. Here, we evaluate how employers update their expectations depending on the candidates' gender and on the employers' IAT score. To do so we construct two variables. The first variable captures the news received by employer i

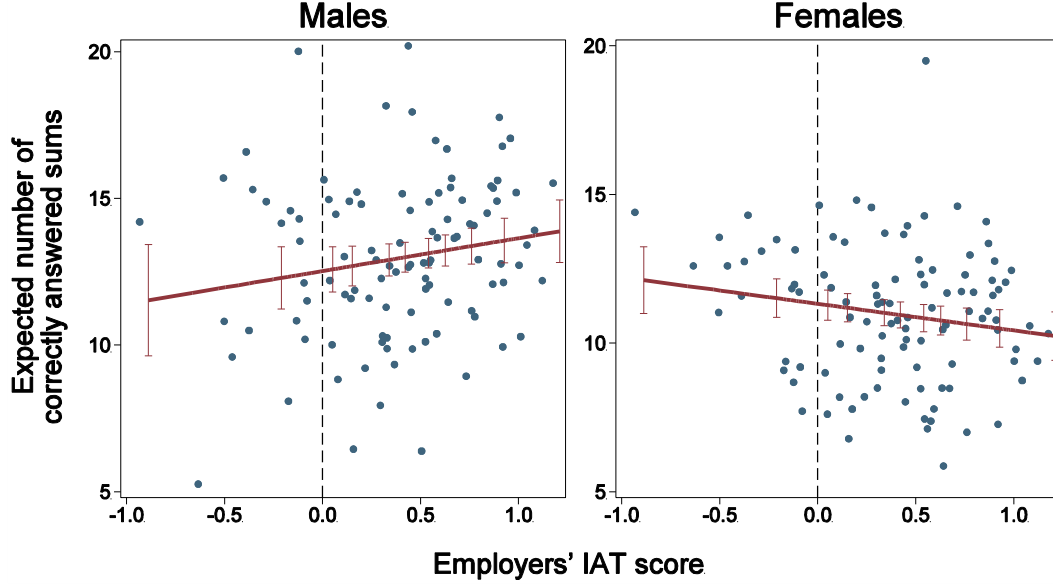


Fig. S5 Association between IAT scores and the expected performance of male and female candidates in the addition task. Each dot corresponds to an employer's IAT score and the mean expected performance of all the male (left panel) and female (right panel) candidates faced by that employer. The lines and 95% confidence intervals are calculated by regressing the employers' mean expected performance of either male (left panel) or female (right panel) candidates on the employer's IAT score in the No Information condition (using robust standard errors, see Table S12).

concerning the performance of candidate j : $\sigma_{ij} = s_{ij} - b_{ij}$, where b_{ij} is i 's expected performance of j when i has no information other than j 's appearance (i.e., i 's prior belief) and s_{ij} is the “signal” i observes about j 's performance (i.e., j 's announced future performance in Decision Then Cheap Talk or j 's past performance in Decision Then Past Performance). The second variable is the amount by which i updates her expectations after receiving the news σ_{ij} : $\theta_{ij} = \mu_{ij} - b_{ij}$, where μ_{ij} is i 's expected performance of j after observing s_{ij} . Note that the degree to which i updates her expectations, as defined in the main body of the paper is $\phi_{ij} = \theta_{ij} / \sigma_{ij}$.

We study differences in the updating process by regressing θ_{ij} on σ_{ij} . Since $\phi_{ij} \times \sigma_{ij} = \theta_{ij}$, in the regression of θ_{ij} on σ_{ij} , the coefficient of σ_{ij} provides us with an estimate for the mean value of ϕ_{ij} . We ran a separate regression for each treatment using linear estimates with picker fixed effects and robust standard errors clustered on individual employers. We excluded observations where θ_{ij} and σ_{ij} have opposite signs because these employers seem to have updated irrationally (i.e., they updated in the wrong direction). Less than 9.1% of all observations correspond to this case. Moreover, our results are unaffected if we include them. The resulting estimates are

Table S13. Descriptive statistics of candidate j 's expected performance in the second arithmetic task (e_{2j}) and the difference between j 's expectation and j 's performance in the first ($e_{2j} - y_{1j}$) and second ($e_{2j} - y_{2j}$) arithmetic tasks, depending on the gender of j .

Variable	Statistic	Cheap Talk		Past Performance		Decision Then Cheap Talk		Decision Then Past Performance	
		Male	Female	Male	Female	Male	Female	Male	Female
e_{2j}	mean	13.867	12.333	12.609	11.174	15.444	10.444	12.300	11.750
	median	13.000	11.000	13.000	10.000	14.000	9.500	12.000	11.500
	std. dev.	4.086	4.746	5.141	3.639	7.006	4.047	4.269	5.200
	Cohen's d	0.358		0.329		0.899		0.119	
$e_{2j} - y_{1j}$	mean	2.467	-0.067	0.652	0.043	3.333	0.444	0.450	0.000
	median	2.000	0.000	1.000	0.000	2.500	0.000	0.000	0.000
	std. dev.	1.922	2.604	0.647	0.638	3.710	1.247	0.605	0.649
	Cohen's d	1.146		0.968		1.074		0.736	
$e_{2j} - y_{2j}$	mean	0.800	-1.933	-0.957	-2.348	2.278	-1.167	-1.750	-1.600
	median	1.000	-2.000	-1.000	-2.000	3.000	0.000	-2.000	-1.000
	std. dev.	2.624	2.314	1.745	2.516	3.025	3.930	3.782	2.500
	Cohen's d	1.144		0.657		1.011		-0.048	

Table S14. P -values from Mann-Whitney U tests comparing the distributions of male and female candidates for the candidate's expected performance in the second arithmetic task (e_{2j}) and the difference between their expectation and their actual performance in the first ($e_{2j} - y_{1j}$) and second ($e_{2j} - y_{2j}$) arithmetic tasks.

Treatment	Variable		
	e_{2j}	$e_{2j} - y_{1j}$	$e_{2j} - y_{2j}$
Cheap Talk	0.196	0.001	0.007
Past Performance	0.433	0.004	0.012
Decision Then Cheap Talk	0.027	0.005	0.008
Decision Then Past Performance	0.786	0.041	0.923

presented in Table S15. In order not to make the table overly long, we simply report the coefficients that estimate the mean value of ϕ_{ij} .

Columns I and IV show the estimated mean values of ϕ_{ij} in Decision Then Past Performance and Decision Then Cheap Talk. They are both positive and are significantly different from zero and from one (Wald tests, $P < 0.001$ in all cases). Thus, employers update, but they do not update as much as Bayesian model with diffuse prior would predict.

In columns II and V, we interact σ_{ij} with a dummy variable indicating the gender of candidate j , which gives us a separate estimate of the mean value of ϕ_{ij} for male and female candidates. In Decision Then Past Performance, the coefficients are or similar value. By contrast, in Decision Then Cheap Talk, employers seem to update more when the candidate is a woman. In the middle panel of Table S15, we test whether these gender differences are significant. It

Table S15. GLS regressions with θ_{ij} as the dependent variable and σ_{ij} , interacted with various dummy variables as independent variables. $\theta_{ij} = \mu_{ij} - b_{ij}$, where b_{ij} and μ_{ij} are i 's prior and updated expectations of j 's performance. $\sigma_{ij} = s_{ij} - b_{ij}$, where s_{ij} is either j 's announced performance or j 's past performance. The top panel reports the estimated coefficients with robust standard errors in parenthesis. All regressions contain employer fixed effects. *, **, and *** denote significance at the 10%, 5%, and 1% level. The middle panel reports the coefficients and robust standard errors of various hypotheses tests. The bottom panel indicates the number of observations, number of employers, and the R^2 .

	Past Performance			Cheap Talk		
	I	II	III	IV	V	VI
σ_{ij}	0.712*** (0.037)			0.517*** (0.042)		
$\sigma_{ij} \times \text{female}$		0.696*** (0.049)			0.620*** (0.049)	
$\sigma_{ij} \times \text{male}$		0.735*** (0.038)			0.478*** (0.048)	
$\sigma_{ij} \times \text{female} \times \text{low IAT}$			0.715*** (0.060)			0.617*** (0.066)
$\sigma_{ij} \times \text{male} \times \text{low IAT}$			0.742*** (0.058)			0.385*** (0.065)
$\sigma_{ij} \times \text{female} \times \text{high IAT}$			0.674*** (0.077)			0.610*** (0.075)
$\sigma_{ij} \times \text{male} \times \text{high IAT}$			0.732*** (0.050)			0.560*** (0.060)
female – male		–0.038 (0.050)			0.142** (0.055)	
female \times low IAT – male \times low IAT			–0.027 (0.055)			0.232*** (0.070)
female \times high IAT – male \times high IAT			–0.058 (0.081)			0.050 (0.075)
(female \times low IAT – male \times low IAT) – (female \times high IAT – male \times high IAT)			0.031 (0.098)			0.182* (0.102)
Number of observations	446	446	446	476	476	476
Number of employers	53	53	53	51	51	51
R^2	0.701	0.702	0.700	0.543	0.556	0.572

reports the coefficient and standard error of the difference in the estimated values of φ_{ij} between females and males. We confirm that employers update similarly in Decision Then Past Performance ($P = 0.444$) and update significantly more for female candidates compared male candidates in Decision Then Cheap Talk ($P = 0.013$).

In columns III and VI, we interact σ_{ij} with a dummy variable indicating the gender of candidate j and a dummy variable indicating whether employer i 's IAT score is below average (labeled as low) or above average (labeled as high). As before, we test whether there are gender

Table S16. GLS regressions with ω_{ij} as the dependent variable and σ_{ij} , interacted with various dummy variables as independent variables. $\omega_{ij} = y_{2j} - b_{ij}$, where b_{ij} is i 's prior expectation of j 's performance and y_{2j} is j 's actual performance in the second arithmetic task. $\sigma_{ij} = s_{ij} - b_{ij}$, where s_{ij} is either j 's announced performance or j 's past performance. The top panel reports the estimated coefficients with robust standard errors in parenthesis. All regressions contain employer fixed effects. *, **, and *** denote significance at the 10%, 5%, and 1% level. The middle panel reports the coefficients and robust standard errors of various hypotheses tests. The bottom panel indicates the number of observations, number of employers, and the R^2 .

	Past Performance		Cheap Talk	
	I	II	III	IV
σ_{ij}	0.921*** (0.014)		0.907*** (0.018)	
$\sigma_{ij} \times \text{female}$		0.901*** (0.018)		1.093*** (0.046)
$\sigma_{ij} \times \text{male}$		0.960*** (0.030)		0.884*** (0.017)
female – male		–0.059 (0.038)		0.209*** (0.048)
Number of observations	476	476	538	538
Number of employers	53	53	51	51
R^2	0.764	0.766	0.759	0.813

differences in updating in the middle panel of the table. In Decision Then Past Performance, the estimated mean values of ϕ_{ij} between females and males are very similar irrespective of the employer's IAT score ($P = 0.625$ for low IAT scores and $P = 0.478$ for high IAT scores). In Decision Then Cheap Talk, employers with low IAT scores update considerably more when the candidate is a woman ($P = 0.002$) whereas employers with high IAT scores do not make this distinction ($P = 0.509$). If we test whether the difference in updating is significantly different between employers with low and high IAT scores we obtain a P -value of $P = 0.081$. Thus, stereotypes do not affect the updating process when the information provided is objective but do so when the information is self-reported.

As a last exercise, we evaluate how the updating of employers compares to the optimal amount of updating according to a perfect information benchmark. Namely, instead of regressing θ_{ij} on σ_{ij} , we regress ω_{ij} on σ_{ij} , where $\omega_{ij} = y_{2j} - b_{ij}$ and y_{2j} is j 's actual performance in the second arithmetic task. In other words, ω_{ij} indicates by how much i would have had to have updated her expectations to correctly guess j 's performance. We use regressions with the same characteristics as those in Table S15. The resulting estimates are presented in Table S16. Comparing the estimated coefficients in Table S15 to those in Table S16, we see that, in both Decision Then

Past Performance and Decision Then Cheap Talk, employers give too much credence to their uninformed prior beliefs as they update too little compared to the perfect information benchmark. From column II, we can see that the candidate's past performance is an equally reliable indicator of their future performance for both genders, i.e., the coefficients are not significantly different. By contrast, from column IV, we see that optimal updating implies giving more weight to the announcements of female candidates than those of male candidates, i.e., the coefficients are significantly different. In fact, if we look at the difference between these coefficients in Table S16, i.e. 0.209, we see that it is very close to the difference in the corresponding coefficients in Table S15 for employers with low IAT scores, i.e. 0.232, and is substantially larger than the difference for employers with high IAT scores, i.e. 0.050. In other words, employers that are less prejudiced against women anticipate the gender difference in the reliability of the candidates' announcements whereas employers that are more prejudiced do not.

Analysis of the costs of discrimination. Here, we evaluate the monetary cost of the employers' biases in beliefs to both the candidates and the employers themselves. First, we consider the earnings of candidates. Candidates that are picked earn \$8 whereas candidates that are not picked earn \$4. Therefore, the gender gap in picking decisions analyzed in Table S4 translates into a difference in the expected earnings of male and female candidates. Specifically, in the No Information condition the expected earnings of male candidates equal \$6.64 whereas that of female candidates equal \$5.36 (19.4% less), in Cheap Talk the expected earnings of males equal \$6.65 and that of females \$5.35 (19.5% less), and in Past Performance the expected earnings of males equal \$6.28 and that of females \$5.72 (8.9% less). Note that all the statistical comparisons in Table S4 apply to the candidates' expected earnings.

More interesting is to calculate the cost of the employers' biases to the employers themselves. To do so, we construct two measures of earnings. The first measure of earnings equals the employers' earnings given their pick, normalized by the maximum earnings they could have obtained. That is, if employer i picks candidate j over candidate k then the first earnings measure equals $\pi_j / \max[\pi_j, \pi_k]$, where π_j and π_k equal the earnings implied by the performance of j and k in the second arithmetic task (the correspondence between earnings and performance is available in the Materials and Methods section). For our second measure of earnings, we concentrate solely on the effect of *biases in beliefs*. To do so, we use i 's expected performance of j and k to determine which candidate i should pick (assuming i picks: j if $e_{ij} > e_{ik}$,

k if $e_{ij} < e_{ik}$, and randomizes if $e_{ij} = e_{ik}$), and then we use this pick to once again determine i 's normalized earnings, $\pi_j / \max[\pi_j, \pi_k]$.

We compare these earnings measures to four benchmarks. For our first benchmark, we calculate normalized earnings if employers were to pick one of the two candidates at random. For our second benchmark, we calculate normalized earnings if employers were to pick the candidate who performed better in the first arithmetic task (note that this information was available to the employers only in the Past Performance condition). For our third benchmark, we use information concerning the degree to which the employers' *initial beliefs* are biased to attempt to arrive to an unbiased pick. Specifically, we calculate the mean difference between the performance of candidates in the second arithmetic task and the employers' initial beliefs for both male and female candidates (on average, employers underestimate the performance of men by 0.434 sums and the performance of women by 1.361 sums). Then, we use these means to adjust the employers' initial beliefs and use the "unbiased" initial beliefs to calculate which candidate should be chosen by each employer and what the corresponding normalized earnings are. In the No Information condition, this is straightforward. For the subsequent decisions in Decision Then Cheap Talk and Decision Then Past Performance, we need to make extra assumptions about the employer's updating process, which we assume is Bayesian updating according to the coefficients of regressions II and V of Table S15. That is, we calculate each employer i 's posterior belief μ_{ij} of candidate j 's performance given j 's gender and the "signal" i observes about j 's performance (s_{ij}) as $\mu_{ij} = \sigma \times (s_{ij} - b_{ij}^U) + b_{ij}^U$, where b_{ij}^U is i 's "unbiased" initial belief and σ is the appropriate updating coefficient of Table S15 (e.g., in Decision Then Cheap Talk, $\mu_{ij} = 0.620 \times (s_{ij} - b_{ij}^U) + b_{ij}^U$ if j is female and $\mu_{ij} = 0.478 \times (s_{ij} - b_{ij}^U) + b_{ij}^U$ if j is male). In other words, this benchmark reduces the bias in initial beliefs but ignores any biases in the belief updating process. For our fourth benchmark, we use information concerning the degree to which the employers' *belief updating process* is biased to attempt to arrive to an unbiased pick. Specifically, we take each employer i 's initial expectations of candidate j (b_{ij}) as given and then use the coefficients from regressions II and IV of Table S16 to calculate what i 's optimal posterior belief μ_{ij}^U is given j 's gender and the "signal" i observes about j 's performance (s_{ij}) (e.g., in Decision Then Cheap Talk $\mu_{ij}^U = 1.093 \times (s_{ij} - b_{ij}) + b_{ij}$ if j is female and $\mu_{ij}^U = 0.884 \times (s_{ij} - b_{ij}) + b_{ij}$ if j is male). We then use the "unbiased" posterior beliefs to calculate which candidate should be chosen by each employer and what the corresponding normalized earnings

Table S17. Mean earnings of employers according to the performance of: (a) the candidate picked by the employer, (b) the candidate expected to perform best with actual beliefs, (c) a randomly chosen candidate, (d) the candidate with the higher past performance, (e) the candidate expected to perform best with unbiased initial beliefs, and (f) the candidate expected to perform best with unbiased posterior beliefs.

Earnings according to the:	No Information	Decision Then Cheap Talk	Decision Then Past Performance
(a) Employers' picks	79.2%	90.4%	94.6%
(b) Employers' expectations	78.8%	88.0%	92.0%
(c) Random picking	73.8%	76.0%	71.4%
(d) Candidates' past performance	100.0%	100.0%	100.0%
(e) Unbiased initial beliefs	78.9%	89.8%	95.4%
(f) Unbiased posterior beliefs	78.8%	94.1%	100.0%

are. In other words, this benchmark leaves the bias in initial beliefs but removes biases in the belief updating process.

The means for the two measures of normalized earnings and the four benchmarks are available Table S17. As one would expect, earnings are higher when employers have more information about the candidates (compare No Information with subsequent decisions in Decision Then Cheap Talk or Decision Then Past Performance), and the more so the better the quality of the information is (compare Decision Then Cheap Talk with Decision Then Past Performance). Interestingly, employers seem to gain some useful information from the appearance of the candidates as their earnings are higher in the No Information condition compared to the random-choice benchmark (difference is significant with both earnings measures with Wilcoxon signed-ranked tests, $P < 0.001$). We can also see that correcting initial beliefs to take into account the employers' relative underestimation of the performance of female candidates has a negligible effect in No Information (an improvement of 0.1%, $P = 0.485$ with a Wilcoxon signed-ranked test). Moreover, although adjusting the employers' initial beliefs (leaving untouched the updating process) produces modest gains after employers' update their expectations (an improvement of 1.8% in Decision Then Cheap Talk and 3.4% Decision Then Past Performance, respectively $P = 0.798$ and $P = 0.097$ with Wilcoxon signed-ranked tests), a considerably bigger improvement is obtained if we adjust the updating process, (which produces an improvement of 6.1% in Decision Then Cheap Talk and one of 8.0% in Decision Then Past Performance, respectively $P = 0.029$ and $P < 0.001$ with a Wilcoxon signed-ranked tests).

Supplementary Information References

Fischbacher U (2007) z-tree: Zurich toolbox for ready-made economic experiments. *Exp Econ* 10(2): 171-178.

Greenwald AG, Nosek BA, Banaji MR (2003) Understanding and using the implicit association test: I. An improved scoring algorithm. *J Pers Soc Psychol* 85(2): 197-216.