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 forth 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 trails within the suggested response times. Our results remain unaffected if we drop from the statistical analysis the few subjects with less than 119 trails. Second, we calculated the mean difference in response times between trials in blocks 6 and 3, $DIFF_{6\text{-}3}$, and between trials in blocks 7 and 4, $DIFF_{7\text{-}4}$. 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 ½($DIFF_{6\text{-}3}$/ SD_{6+3} + $DIFF_{7\text{-}4}$/ 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

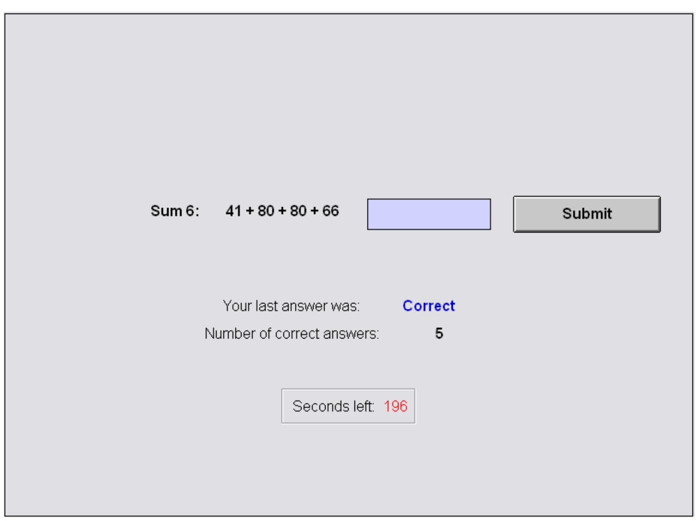
Thank you for participating in today's study. The study will last around 45 minutes. You are not allowed to communicate with other participants. If you have a question, raise your hand and we will gladly help you. For your participation you will receive an $8 *show-up fee*. In addition, you will be able 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.



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).

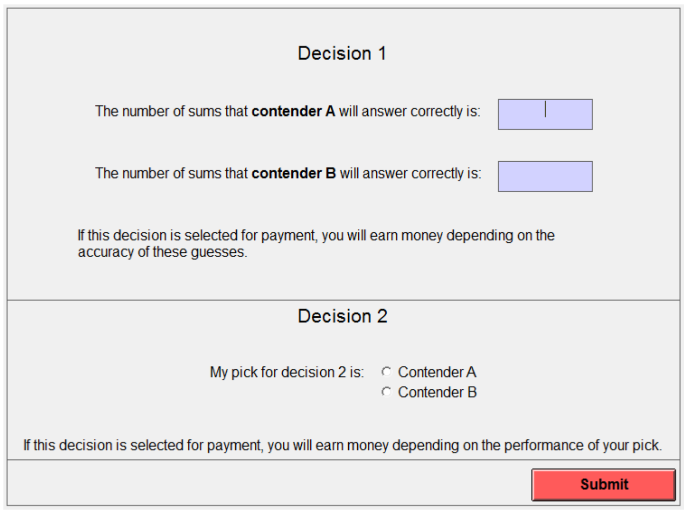
Observers

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:



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.

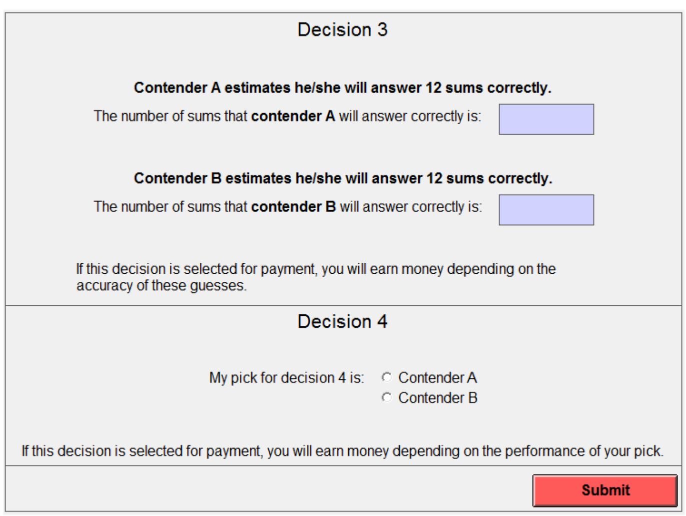
Number of sums answered correctly by the contender you pick	Your earnings
less than 5 sums	\$0.00
between 6 and 8 sums	\$1.00
between 9 and 11 sums	\$2.00
between 12 and 14 sums	\$4.00
between 15 and 17 sums	\$7.00
between 18 and 20 sums	\$11.00
between 21 and 23 sums	\$16.00
more than 24 sums	\$22.00

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).



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.

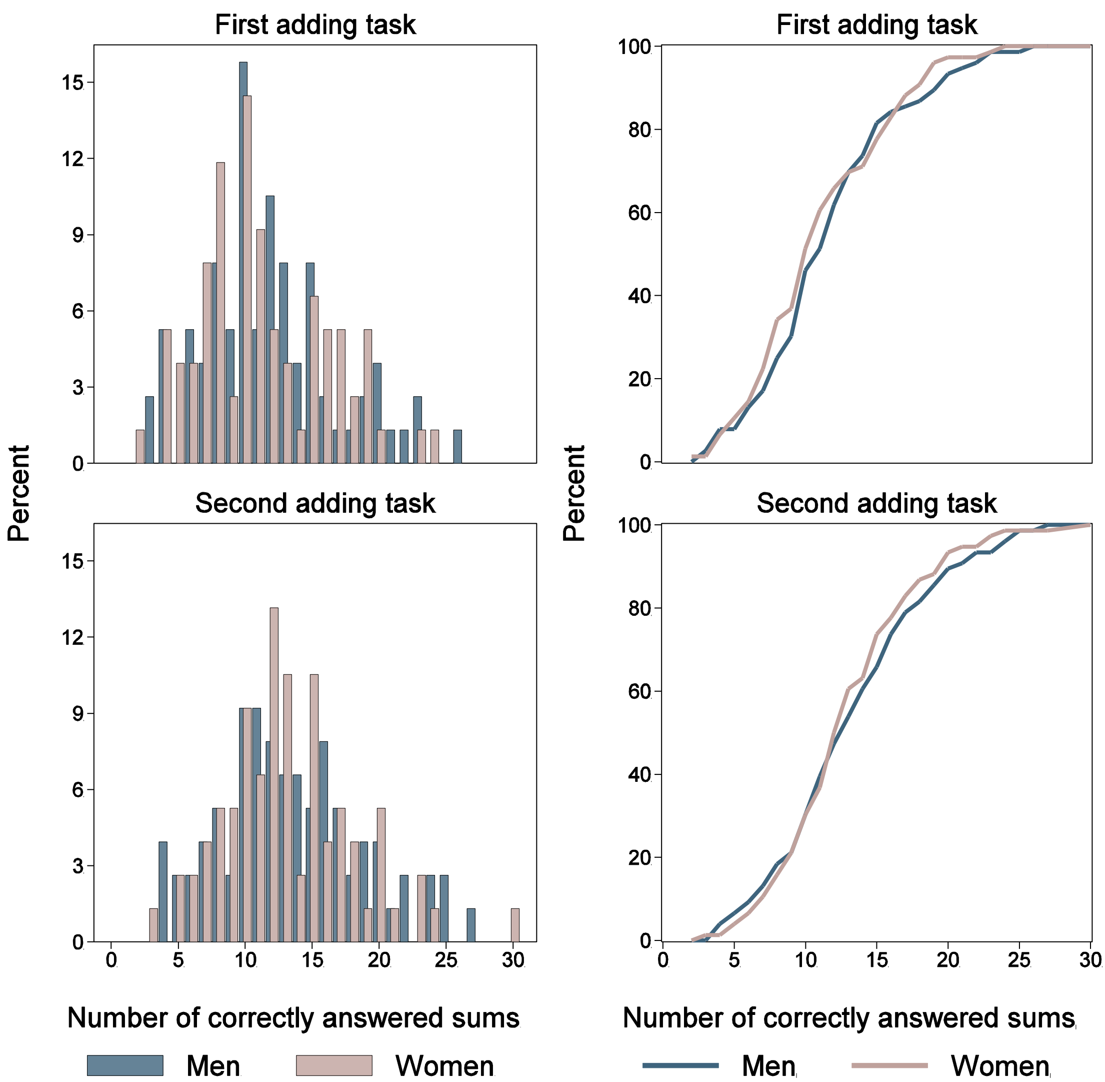


Fig. S2 The bars show the distribution of the subjects' performance in the two arithmetic tasks depending on their gender. The lines show the corresponding cumulative distributions.

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	–0.020	0.000	–0.020
	(0.051)	(0.026)	(0.051)	(0.026)
Past performance	0.091***	0.051*	0.095***	0.051*
	(0.031)	(0.028)	(0.032)	(0.028)
Female employer			0.031	0.001
			(0.031)	(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***	–0.105***	–0.130***	–0.105***
	(0.045)	(0.032)	(0.045)	(0.032)
Past performance	–0.250***	–0.320***	–0.249***	–0.319***
	(0.043)	(0.039)	(0.044)	(0.039)
Female employer			0.012	0.022
			(0.034)	(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***	0.155***	0.228***	0.155***
	(0.071)	(0.045)	(0.072)	(0.045)
Past performance	–0.054	0.112*	–0.055	0.113*
	(0.066)	(0.065)	(0.067)	(0.065)
Female employer			–0.015	0.007
			(0.050)	(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 × 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.

		No Information		Cheap Talk		Past Performance		Decision Then Cheap Talk		Decision Then Past Performance	
Variable	Statistic	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 × treatment fixed effects. Robust standard errors in parenthesis. *, **, and *** denote significance at the 10%, 5%, and 1% level.

	Dependent variable	
Independent variables	e_{ij}	$e_{ij} > e_{ik}$
No Information × female	-2.061^{***}	-1.154^{***}
	(0.309)	(0.163)
Cheap Talk × female	0.594	-0.795^{**}
	(0.548)	(0.348)
Past Performance × female	-1.060^{***}	-0.481^{**}
	(0.321)	(0.206)
Decision Then Cheap Talk × female	-2.033^{***}	-1.080^{***}
	(0.438)	(0.192)
Decision Then Past Performance × female	–0.429	-0.624^{***}
	(0.279)	(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.

	No Information		Cheap Talk		Past Performance	
Expectations	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%

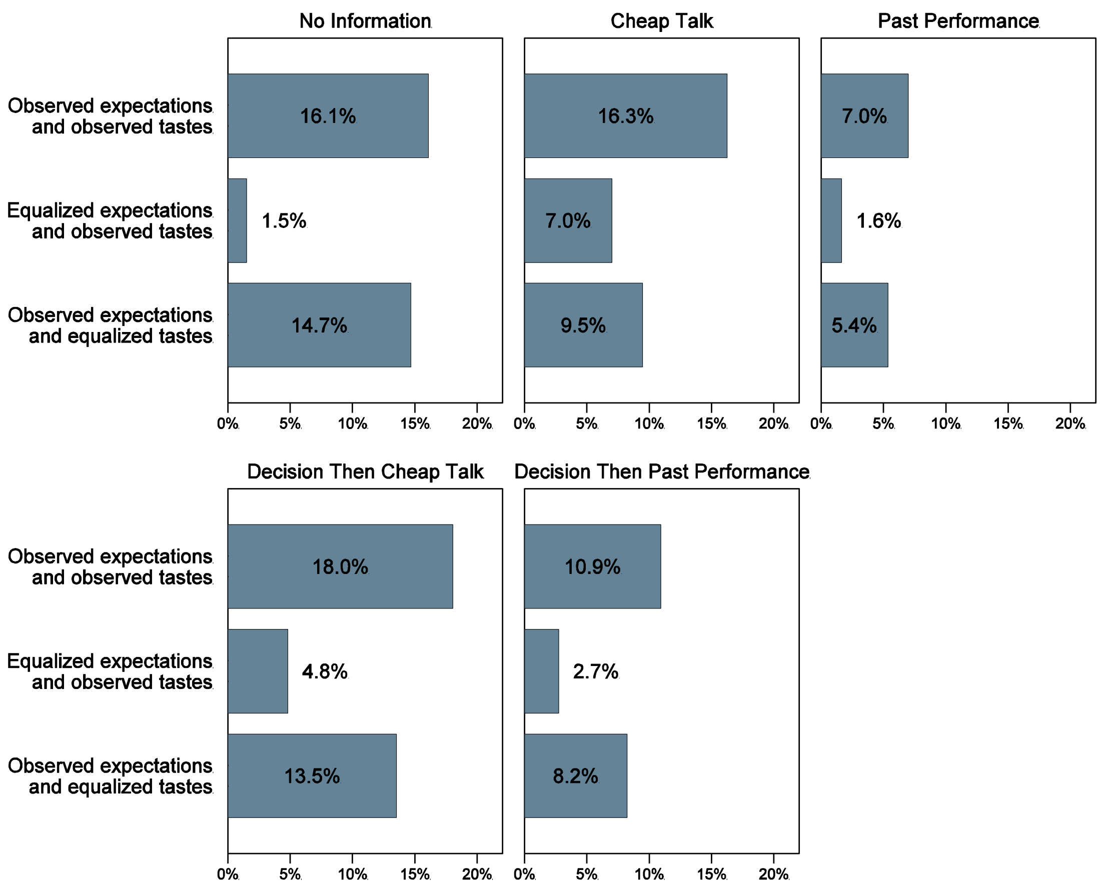


Fig. S3 Fraction of picked male candidates minus the fraction of picked female candidates in each information condition if: (top) there are gender differences in expectations and in picking; (middle) there are no gender differences in expectations but there are differences in picking; and (bottom) there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

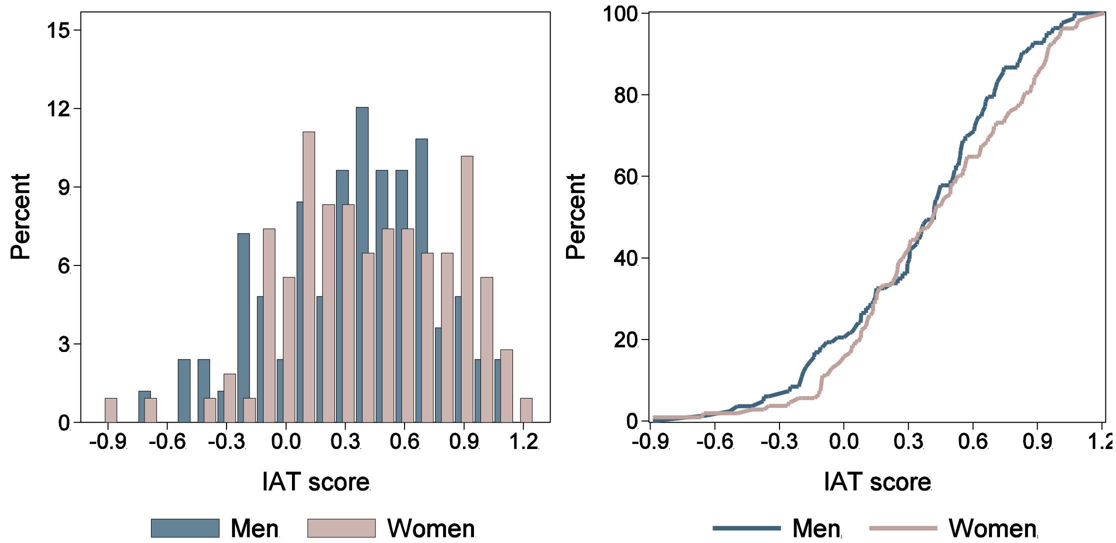


Fig. S4 The bars show the distribution of the employers' IAT score depending on their gender. The lines show the corresponding cumulative distributions.

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.

	Dependent variable		
Independent variables	Males	Females	Males – Females
IAT score	1.120^{*}	-0.896^{**}	2.016^{***}
	(0.674)	(0.419)	(0.710)
Constant	12.521^{***}	11.321^{***}	1.200^{**}
	(0.422)	(0.273)	(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

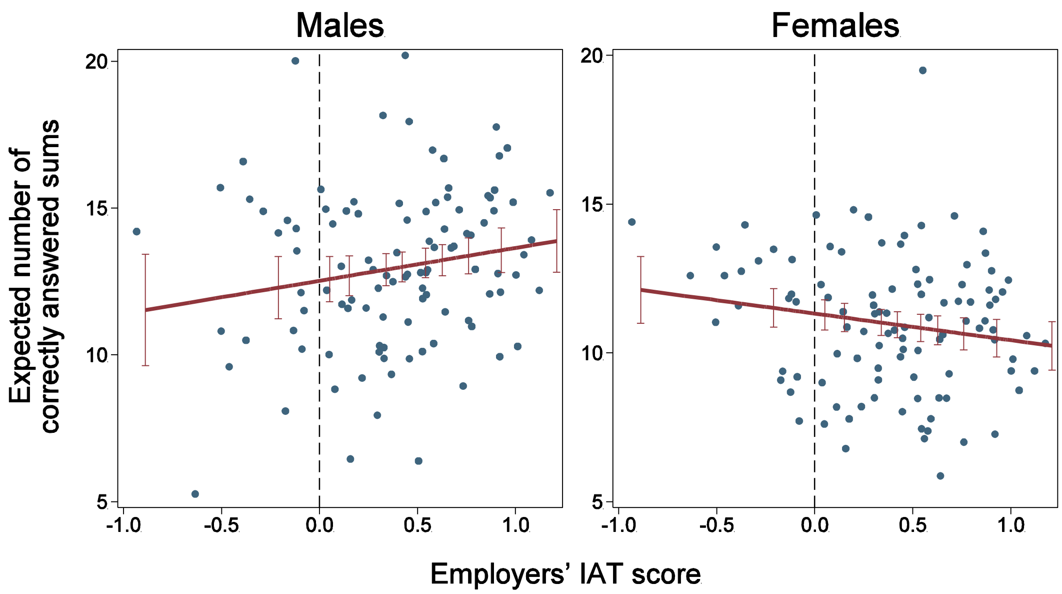


Fig. S5 Association between IAT scores and the expected performance of male and female candidates in the addition task. Each dot corresponds to an employer's IAT score and the mean expected performance of all the male (left panel) and female (right panel) candidates faced by that employer. The lines 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 $\varphi_{ij} = \theta_{ij} / \sigma_{ij}$.

We study differences in the updating process by regressing θ_{ij} on σ_{ij}. Since $\varphi_{ij} \times \sigma_{ij} = \theta_{ij}$, 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*.

		Cheap Talk		Past Performance		Decision Then Cheap Talk		Decision Then Past Performance	
Variable	Statistic	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.

	Variable		
Treatment	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***			0.517***		
	(0.037)			(0.042)		
σ_{ij} × female		0.696***			0.620***	
		(0.049)			(0.049)	
σ_{ij} × male		0.735***			0.478***	
		(0.038)			(0.048)	
σ_{ij} × female × low IAT			0.715***			0.617***
			(0.060)			(0.066)
σ_{ij} × male × low IAT			0.742***			0.385***
			(0.058)			(0.065)
σ_{ij} × female × high IAT			0.674***			0.610***
			(0.077)			(0.075)
σ_{ij} × male × high IAT			0.732***			0.560***
			(0.050)			(0.060)
female – male		–0.038			0.142**	
		(0.050)			(0.055)	
female × low IAT – male × low IAT			–0.027			0.232***
			(0.055)			(0.070)
female × high IAT – male × high IAT			–0.058			0.050
			(0.081)			(0.075)
(female × low IAT – male × low IAT) –			0.031			0.182*
(female × high IAT – male × high IAT)			(0.098)			(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***		0.907***	
	(0.014)		(0.018)	
σ_{ij} × female		0.901***		1.093***
		(0.018)		(0.046)
σ_{ij} × male		0.960***		0.884***
		(0.030)		(0.017)
female – male		–0.059		0.209***
		(0.038)		(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 π_j / max[π_j, π_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 $\mu_{ij} = \sigma \times (s_{ij} - b^U_{ij}) + b^U_{ij}$, where b^U_{ij} is i's "unbiased" initial belief and σ is the appropriate updating coefficient of Table S15 (e.g., in Decision Then Cheap Talk, $\mu_{ij} = 0.620 \times (s_{ij} - b^U_{ij}) + b^U_{ij}$ if j is female and $\mu_{ij} = 0.478 \times (s_{ij} - b^U_{ij}) + b^U_{ij}$ 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^U_{ij} is given j's gender and the "signal" i observes about j's performance (s_{ij}) (e.g., in Decision Then Cheap Talk $\mu^U_{ij} = 1.093 \times (s_{ij} - b_{ij}) + b_{ij}$ if j is female and $\mu^U_{ij} = 0.884 \times (s_{ij} - b_{ij}) + b_{ij}$ 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.