How stereotypes impair women's careers in science

Ernesto Reuben,
Paola Sapienza,
Luigi Zingales

How stereotypes impair women's careers in science

Ernesto Reuben

Columbia University, e-mail: ereuben@columbia.edu

Paola Sapienza

Northwestern University, e-mail: Paola-Sapienza@northwestern.edu

Luigi Zingales

University of Chicago, e-mail: Luigi.Zingales@chicagobooth.edu

ABSTRACT: 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 gender-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 gender 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 gender-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.

Keywords: gender stereotypes, science education, diversity, science workforce

Note: This is the authors' version of a work that was accepted for publication in the *Proceedings of the National Academy of Sciences*. Changes resulting from the publishing process may not be reflected in this document. A final version is published in http://dx.doi.org/10.1073/pnas.1314788111.

1. Introduction

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 (Zafar, 2013). The gender-based disparity in STEM fields is even greater at the graduate-school level (Hill et al., 2010). In a controversial speech, Larry Summers (Summers, 2005), 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 (Hyde and Mertz, 2009; Hyde et al., 2008; Guiso et al., 2008), it is difficult to show the existence of discrimination if we allow for the possibility of a gender 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 gender differences in performance (Hyde et al., 1990; Niederle and Vesterlund, 2007; Niederle et al., 2013).

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 gender 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) (Greenwald et al., 1998) 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 gender 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 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.

2. Experimental design

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 (Hyde et al., 1990; Niederle and Vesterlund, 2007; Niederle et al., 2013). Nevertheless, it belongs to an area—mathematics—about which there is a pervasive stereotype that men perform better (Correll, 2001; Rudman et al., 2001; Kiefer and Sekaquaptewa, 2007).

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-gender 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 gender 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 gender. 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' gender. 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-gender candidate pairs in each treatment.

As a final step, we asked all subjects to complete an IAT associating gender with science-related abilities (Greenwald et al., 1998). 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 (Greenwald et al., 2009). 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

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-gender candidates pairs	15	23	18	20
Mean number of mixed-gender 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

Note: For each treatment, the table presents the number of subjects who acted as employers when a mixed-gender alternative was presented to them, the total number of mixed-gender candidate pairs, the mean number of decisions per employer in the mixed-gender 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 Table S1.

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 gender-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 the Supplementary Information (Dataset S1).

3. 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 gender (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 gender-biased decision-making in this context but failed to eliminate it, especially in employers who showed a stronger implicit gender bias as revealed by the IAT. Detailed versions of these results are presented in the sections below. Statistical support for the results is presented in the Supplementary Information.

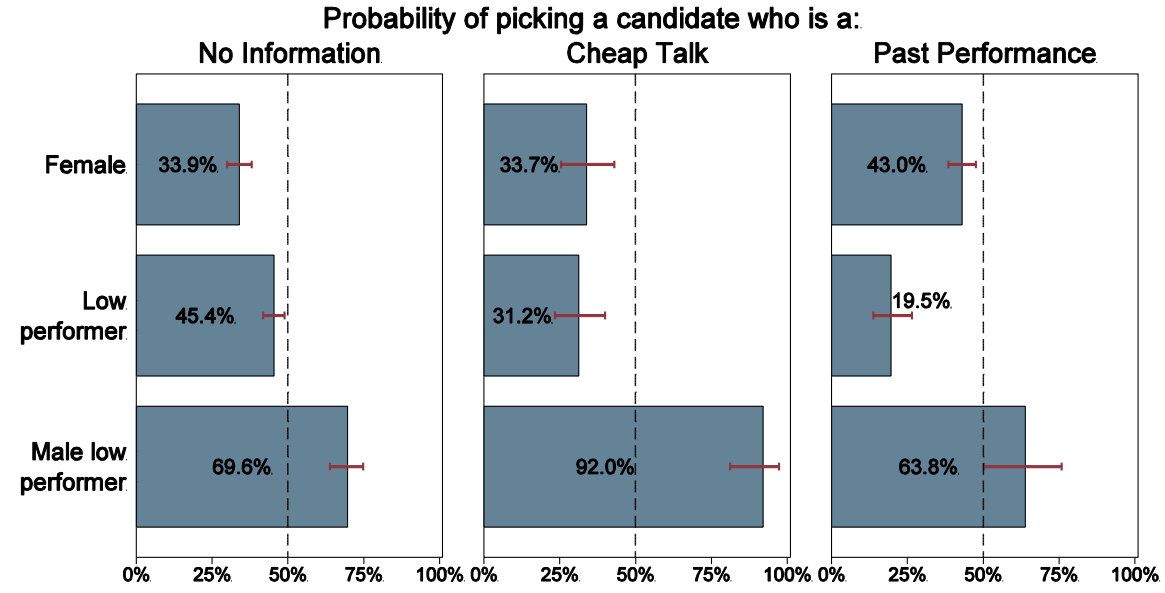
3.1. Initial Hiring Decisions and Gender-Related Beliefs

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 gender. 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 and with the existing literature (Correll, 2001; Rudman et al., 2001; Kiefer and Sekaquaptewa, 2007). 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 (Figure 1). When employers had no information beyond appearance, they were twice more likely to choose

Figure 1 – Initial picking decision depending on the available information



Note: 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 (Tables S4–S6).

male candidates than female candidates. Regression analyses (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 Figure 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 gender 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 (Figure 1). Regression analysis (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 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 ($p < 0.009$) (Table S17).

Suboptimal hiring decisions were associated strongly with gender bias. If hiring decisions were gender-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 (Figure 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; 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.

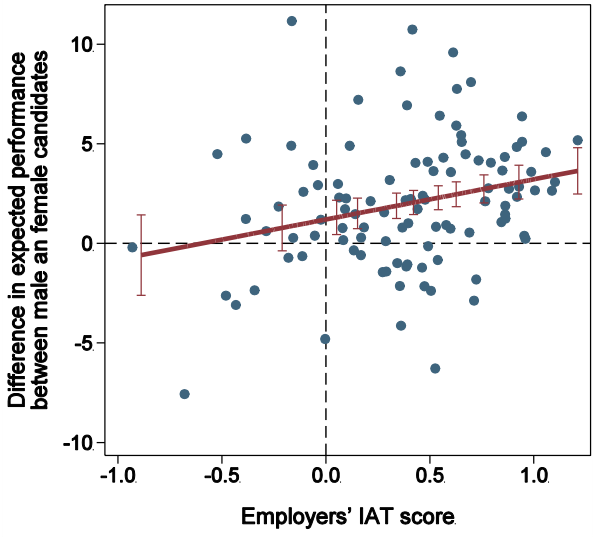
Hiring choices were consistent with employers' expectations regarding the performance of female and male candidates. Employers overwhelmingly chose the candidate for whom they had higher expectations, irrespective of candidates' gender (Table S10). Hence, if employers did not have biased expectations in favor of men, there would be no noticeable gender gap in hiring decisions (Figure 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.

3.2. 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 genders 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

Figure 2 – IAT score and the difference in expected performance between male and female candidates



Note: Association between IAT scores and the difference in expected performance of male and female candidates in the addition task in the No Information condition (n = 104). Each dot corresponds to an employer's IAT score and the difference between the expected performance of the male and the female candidate averaged across all mixed-gender pairs faced by that employer. The line and 95% confidence intervals are calculated by regressing employer i's difference between the expected performance of the male and the female candidate averaged across all mixed-gender pairs faced by employer i on i's IAT score (using robust SEs; see Table S12).

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 the Supplementary Information.

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 (Table S12). We found a positive relation between employers' IAT scores and their average expectation of the performance of all evaluated male candidates (β = 1.08, p = 0.079) and a negative relation between IAT scores and the average expected performance of all evaluated female candidates (β = −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 (β = 1.99, p = 0.005). This relationship is plotted in Figure 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 gender 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).

3.3. 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 j after receiving new information about j's performance (i's updated belief, μ_{ij}) minus i's expected performance of j before receiving any information (i's prior belief, b_{ij}). The denominator of φ_{ij} equals the "signal" s_j about candidate j's performance—s_j equals j's claimed future performance in the Decision Then Cheap Talk condition and j's past performance in the Decision Then Past Performance condition—minus i's prior expectation. Note that if i treats the signal s_j as completely uninformative, then the updated belief will be $\mu_{ij} = b_{ij}$ and $\varphi_{ij} = 0$. In contrast, if i treats the prior belief as completely uninformative (i.e., i has a diffuse prior), then the updated belief will be $\mu_{ij} = s_j$ and $\varphi_{ij} = 1$. In the Decision Then Cheap Talk condition, 20.7% of employers did not update their expectation ($\varphi_{ij} = 0$ when $s_j \neq b_{ij}$), and 34.6% updated as if their prior belief was completely uninformative ($\varphi_{ij} = 1$ when $s_j \neq b_{ij}$). In the Decision Then Past Performance condition, the respective numbers were 12.8% and 46.6%. We used regression analysis to estimate the mean value of φ_{ij} that best describes the employers' updating in the different information conditions (Table S15) as well as the mean value of φ ij that corresponds to optimal updating (Table S16), which is defined as the φ_{ij} for which i's updated belief matches j's subsequent performance.

Employers found candidates' past performance a more reliable signal, and hence more useful information for decision-making, than their self-reported expectation of future

Table 2 – Degree to which employers update their expectations

	Female candidate		Male candidate		Difference	
	estimate	std. err.	estimate	std. err.	estimate	std. err.
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)

Note: 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 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.

performance, but they still weighted prior beliefs excessively. In the Decision Then Past Performance condition, the estimated mean value of φ_{ij} was 0.712, whereas in Decision Then Cheap Talk condition it was 0.517. However, in both cases the estimated mean value of φ_{ij} was significantly lower than the mean values of φ_{ij} implied by optimal updating (i.e., 0.921 in the Decision Then Past Performance condition and 0.907 in the Decision Then Cheap Talk condition; Wald tests, $p < 0.001$); these values are very close to one, the value predicted by a Bayesian model with a diffuse (i.e., uninformative) prior. Thus, employers updated, but did so insufficiently, because they weighted their uninformed prior beliefs too heavily.

The magnitude of updating of employers' beliefs was not biased by candidate gender when information about past performance was provided by the experimenter, even for employers with higher IAT scores. We studied differences in the updating process by looking at how the mean value of φ_{ij} depended on whether the employer was updating expectations about a male or a female candidate and on the employers' implicit prejudices against women, as measured by the IAT. The results are available in Table 2. First, we studied the Decision Then Past

Performance treatment, in which the experimenter provided information about candidates' past performance. We estimated the mean value of φ ij depending on the candidate's gender. The mean values of φ_{ij} were very similar and were not statistically different (a difference of 0.04, $p = 0.444$). The lack of gender-biased updating in this treatment is in line with optimal updating, which assigns similar mean values of φ_{ij} to male and female candidates. Then, we reestimated the same regressions, splitting the sample on whether the employer's IAT score was below average (low) or above average (high). Once again, mean values of φ_{ij} were not statistically different (a difference of 0.03 for low IAT scorers, $p = 0.625$; a difference of 0.06 for high IAT scorers, $p = 0.479$). Thus, stereotypes did not seem to affect the updating process when the information was provided by a neutral third party.

Men tended to overestimate their future performance on the arithmetic task, and women tended to underestimate it—a gender difference taken partially into account by employers' updating. In the bottom rows of Table 2, we repeat the analysis described above for the Decision Then Cheap Talk treatment, in which performance-related information was provided by the candidates themselves. When asked about their future performance, both male and female candidates reported a number higher than their past performance. The difference between figures varied considerably by gender: Men reported 3.33 more correct sums, whereas women reported only 0.44 more correct sums. As a result, men's announcements overestimated their future performance by 2.28 sums, and women's underestimated their future performance by −1.17 sums (significantly different with a Mann–Whitney U test, $p = 0.008$). This behavior is consistent with existing research reporting that women underestimate their performance and show more modesty than men in self-promotion (Beyer, 1990; Reuben et al., 2012). Thus, because men overestimate their future performance, and women underestimate it, optimal updating would require compensating for these biases by giving less weight to the announcements of men than those of women, leading to a significantly lower φ_{ij} for men (by −0.21, $p = 0.001$). The left columns in the lower rows of Table 2 show that employers do anticipate a difference between the announcements of men and women, as the estimated mean value of φ_{ij} is significantly lower for male candidates than for female candidates (by −0.14, $p =$

0.013). Nonetheless, the difference in the mean values of φ_{ij} was not as large as the difference that would be seen with optimal updating.

Employers with a stronger implicit bias against women were more willing to believe men's overestimated expectations of their future performance. We reestimated the mean value of φ_{ij} depending on the level of stereotype-based beliefs held by employers. Less-biased employers (with low IAT scores) made a stark distinction between self-reported performance levels based on the candidates' gender (a difference in the mean value of φ_{ij} of −0.23, $p = 0.002$, which is very close to the optimal difference in the mean values of φ_{ij}). In contrast, more biased employers (with high IAT scores) put more weight on the male candidates' announcements and, as a result, did not differentiate significantly between the self-reports of male and female candidates (a difference in the mean values of φ_{ij} of −0.05, $p = 0.509$). Thus, the same stereotype that made employers discriminate against women on the basis of an incorrect belief in the first place prevented them from filtering candidates' self-reported information optimally. Employers who were more implicitly biased against women were more willing to believe men's inflated expectations about their performance, despite well-established evidence of overestimation in this regard.

Employers' subsequent hiring choices were consistent with their updated beliefs but still resulted in the hiring of fewer female candidates than male candidates. When employers received objective information about candidates' past performance, female candidates still were chosen significantly less often than male candidates (females were chosen 39.1% of the time), but the difference was smaller than in the No Information condition (in which females were chosen 33.9% of the time). When employers received subjective information about the candidates' past performance, the gender gap did not shrink; instead, if anything, it increased (females were chosen 32.0% of the time). As a result, suboptimal decisions were made in favor of the male candidates significantly more often than in favor of the female candidates (a lower-performing male was chosen over a higher-performing female 85.7% of the time in the Decision Then Cheap Talk condition and 82.1% of the time in the Decision Then Past Performance condition).

4. Discussion

Although there is some evidence of a gender difference in mathematics performance (Hyde et al., 2008; Guiso et al., 2008), which is shrinking over time (Hyde et al., 1990), there is no gender disparity in performance on an arithmetic task such as ours (Niederle and Vesterlund, 2007). Nevertheless, the stereotype of women's inferior performance on every mathematics-related task is pervasive (Guiso et al., 2008; Hyde and Mertz, 2009). This stereotype can lead to a decreased demand for women in STEM fields and/or a reduction in the number of women choosing to specialize in these fields. The effect of this stereotype on the hiring of women has been shown to be important in at least one field experiment (Moss-Racusin et al., 2012). However, that study was unable to rule out the possibility that the decision to hire fewer women is the rational response to the lower effective quality of women's future performance because of underinvestment by women caused by inferior career prospects (Arrow, 1973; Lundberg, 1983) or stereotype threat (Sekaquaptewa and Thompson, 2003).

For this reason, we used a laboratory experiment in which we could ensure there was no quality difference between genders, because women performed equally well on the task in question, whether or not they were hired. Despite this equality, employers in our study discriminated against female candidates to a degree that correlated with their implicit bias against women as suggested by their IAT score. Thus, stereotypes do affect the demand for women in mathematics-related tasks, regardless of quality considerations.

There is a lively discussion about how to interpret IAT scores and to what extent they explain behavior (Greenwald et al., 2009). Nevertheless, there is compelling evidence that the IAT captures implicit processing of information that is distinct from more conscious reasoning (Greenwald et al., 1998, 2009; Bertrand et al., 2005). Our findings seem to suggest that both men and women discriminate against women without realizing that they do so. This form of discrimination is very different from the forms normally modeled in economics. Importantly, discrimination driven by implicit associations requires different (less coercive) policies for remediation (Bertrand et al., 2005).

In most situations, employers do not rely only on their priors. They benefit from some information about the candidates: objective measures of past performance, self-reports, or

both. The additional advantage of the laboratory environment is that we can show that the provision of additional information interacts with this initial bias and affects the discrimination outcome. When objective information about past performance is available, it attenuates but does not eliminate the gender bias in hiring. Although the preexisting stereotype does not contaminate the information received (probably because the information is considered objective), it still affects the posterior distribution of expectations. Thus, even in the face of valuable new information, employers continue to rely at least in part on their biased priors.

The effect is very different when self-reported information becomes available. Men tend to be more self-promoting than women in these reports, but employers, particularly those demonstrating evidence of stronger implicit gender bias (higher IAT), do not fully appreciate the extent of this difference. Thus, the bias against women measured by the IAT seems to act in two ways: It penalizes women when an unfounded negative stereotype against them exists, and it does not penalize men when there is evidence (Beyer, 1990; Reuben et al., 2012) that they overpromote themselves.

Acknowledgments

We thank Alice Eagly, Adam Galinsky, Arno Riedl, Martin Strobel, and Elke Weber for helpful comments. P.S. received financial support from the Zell Center for Risk and Research at Kellogg School of Management, Northwestern University. L.Z. received financial support from the Stigler Center and the Initiative on Global Markets at the University of Chicago Booth School of Business.

References

Arrow, K.J. (1973). The theory of discrimination. In Ashenfelter, O., and Rees, A. (eds.) *Discrimination in Labor Markets*, pp 3–33. Princeton University Press: Princeton, NJ.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). New approaches to discrimination: Implicit discrimination. *American Economic Review* 95(2):94–98.

Beyer, S. (1990). Gender differences in the accuracy of self-evaluation of performance. *Journal of Personality and Social Psychology* 59(5):960–970.

Correll, S.J. (2001). Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology* 106(6):1691–1730.

Greenwald, A.G., McGhee, D.E., and Schwartz, J.L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74(6): 1464–1480.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., and Banaji, M.R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97(1):17–41.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, math, and gender. *Science* 320(5880):1164–1165.

Hill, C., Corbett, C., and St. Rose, A. (2010). Why So Few? Women in Science, Technology, Engineering, and Mathematics. American Association of University Women: Washington, DC.

Hyde, J.S., and Mertz, J.E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences* 106(22):8801–8807.

Hyde, J.S., Fennema, E., and Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin* 107(2):139–155.

Hyde, J.S., Lindberg, S.M., Linn, M.C., Ellis, A.B., and Williams, C.C. (2008). Diversity. Gender similarities characterize math performance. *Science* 321(5888):494–495.

Kiefer, A.K., and Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes: A prospective study of female college students. *Psychological Science* 18(1):13–18.

Lundberg, S.J., and Startz, R. (1983) Private discrimination and social intervention in competitive labor markets. *American Economic Review* 73(3):340–347.

Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J., and Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* 109(41): 16474–16479.

Niederle, M., and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics* 122(3):1067–1101.

Niederle, M., Segal, C., and Vesterlund, L. (2013). How costly is diversity? Affirmative action in light of gender differences. *Management Science* 59(1):1–16.

Reuben, E., Rey-Biel, P., Sapienza, P., and Zingales, L. (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behavior & Organization* 83(1):111–117.

Rudman, L.A., Greenwald, A.G., and McGhee, D.E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin* 27(9):1164–1178.

Sekaquaptewa, D., and Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology* 39(1): 68–74.

Summers, L. (2005). Remarks at NBER Conference on Diversifying the Science and Engineering Workforce. Available at www.harvard.edu/president/speeches/summers_2005/nber.php. Accessed December 20, 2013.

Zafar, B. (2013). College major choice and the gender gap. *Journal of Human Resources* 48(3):545–595.