The glass ceiling in experimental markets

Ernesto Reuben

Columbia University

Paola Sapienza

Northwestern University, NBER, and CEPR

Luigi Zingales

University of Chicago, NBER, and CEPR

ABSTRACT

We study an experimental market where, in spite of equal performance across genders, individuals discriminate towards women in hiring decisions. We show that discrimination is neither taste-based nor based on a correct statistical inference regarding differences in performance. Instead, it is rooted in biased beliefs about women's abilities. The gender gap increases when candidates are allowed to influence expectations by declaring their expected performance and it narrows if individuals receive accurate information of the performance of the applicants. However, even when accurate information is transmitted, the gender gap persists because individuals do not completely update their initially biased belief. Furthermore, we show, by using the Implicit Association Test, that unconscious stereotypes are partly responsible for the initial bias in beliefs and the subsequent lack of updating.

Date of this version: March 2010

Keywords: discrimination, gender gap, glass ceiling, biased beliefs, implicit associations

1. Introduction

There are large differences not only in the relative compensation, but also in the relative presence of women versus men in the highest paid jobs (e.g., Azmat, Güell, and Manning, 2004; Arulampalam, Booth, and Bryan, 2007). This underrepresentation is especially severe at the highest levels of the corporate ladder: for example, Bertrand and Hallock (2001) show that only about 2.5 percent of the five-highest paid executives in S&P 500 firms are women and Wolfers (2006) documents that over the period 1992-2004, the CEO of an S&P 1500 firm was a woman only 1.3% of the time. Moreover, the gap persists despite the narrowing gender gap in (business) education and evidence that there is no link between firm performance and the gender of top executives (Albanesi and Olivetti, 2008).

The relative differences in compensation and presence of women can be due to several reasons both form the supply side and the demand side of the labor market. On the supply side, women might be less interested in top positions because they do not like those types of jobs. The experiment of Niederle and Versterlund (2007) suggests that women like competitive environments less than men.[1] Another reason is that women may be less willing to aggressively negotiate for pay and promotion (Babcock and Laschever, 2003).

On the demand side, women may be experiencing some form of discrimination. Discrimination can be taste-based—for example, men prefer to work with other men as opposed to women (Becker, 1971)—or expectation-based (Phelps, 1972; Arrow, 1973). From a policy perspective, it makes a big difference whether expectation-based discrimination arises from an unbiased statistical inference regarding differences in performance (on average women run slower than men) or from a biased perception of women’s abilities (women are believed to be worse drivers than men even though there is no statistical evidence supporting this belief).[2] In this paper, we study whether biased perceptions alone can be responsible for discrimination toward women and what mechanisms exacerbate or mitigate such phenomenon.

[1] Some of the other papers supporting this literature are Gneezy, Niederle, and Rustichini (2003), Gneezy and Rustichini (2004), Günther et al. (2008), Dreber, von Essen, Ranehill (2009), and Gneezy, Leonard, and List (forthcoming). For a survey see Booth (2009).

[2] In addition, as argued by Arrow (1973) and Lundberg and Startz (1983), even if there are no differences in innate abilities, biased beliefs can lead to real differences in performance because they cause underinvestment in human capital by the disadvantaged group. Moreover, this effect is not necessarily corrected by market forces (Milgrom and Oster, 1987) or affirmative action (Coate and Loury, 1993).

The effects of discrimination on market outcomes are very difficult to disentangle when relying on naturally occurring data. Even if one could perfectly control for differences in ability (which we generally cannot), it is very difficult to separate between preferences on the supply side (i.e., women desiring the job) and on the demand side (the people recruiting them). There is evidence that at least part of the difference in the presence of women is not supply driven. In an ingenious paper, Goldin and Rouse (2000) exploit the increase in the use of blind auditions for selecting orchestra musicians. They show that blind auditions account for 25 percent of the increase in the percentage of female orchestra musicians during the studied period. However, their setting does not allow them to disentangle between different forms of discrimination. A lower fraction of women promoted in face-to-face auditions may reflect an expectation bias of the recruiter, but it may also be the result of statistical discrimination based not on the sound quality, but on their expected overall performance as members of an orchestra. If women are more likely to miss rehearsal for medical reasons (Ichino and Moretti, 2009) or more likely to drop out of their job (as in Bertrand, Goldin, Katz, 2009), then the lower fraction of women promoted in the face-to-face auditions does not necessarily reflect a bias of the recruiters. Finally, the under-selection of women could be the result of a dislike of women musicians by other musicians or by the audience (taste-based discrimination as in Akerlof, 1985).[3]

We resort to a laboratory experiment to study discrimination that is driven solely by a biased perception of women's abilities.[4] In other words, we design an experiment in which, first, supply side considerations do not apply (job candidates are chosen randomly and cannot opt out), and second, we can rule out other types of discrimination. To rule out other forms of expectation-based discrimination, we select a short-term task with easily measured outcomes in which there is no gender difference in performance and there are no long-term effects of hiring someone. To eliminate the possibility that choices are biased because of a distaste for interacting with a particular gender (taste-based discrimination), we ensured that the decisions involved do not require any subsequent interaction between the chooser and the chosen individual. By shutting these channels of discrimination, we are

[3] The same limitations are present in the field experiments carried out so far, such as studies that send altered résumés or confederates to interviews and then measure callback rates (Newmark, Bank, and Van Nort, 1996; Bertrand and Mullainathan, 2004; Bravo, Sanhueza, and Urzua, 2008), and studies measuring voting behavior for choosing participants in TV game shows (Levitt, 2004; Antonovics, Arcidiacono, and Walsh, 2005). See Altonji and Blank (1999) for an overview of the literature on discrimination and Heckman (1998) for a discussion of the difficulties of detecting discrimination in the field.

[4] To our knowledge this is the first experiment designed to investigate this type of discrimination. For a survey of experiments studying discrimination see Anderson, Fryer, and Holt (2005).

less likely to find any evidence in favor of it. Thus if such evidence does occur, we can be sure that it comes from biased performance expectations.

An important part of our experimental design is that we elicit directly the participants' expectations of the performance of the job candidates. This not only allows us to test whether expectations are indeed biased and are the driving force behind any observed discrimination, but it also lets us investigate whether there is also a bias in the way participants update their expectations as they receive more information concerning the performance of job candidates. Lastly, to better understand the source of expectation biases, we investigate whether associations captured with the Implicit Association Test (Greenwald, McGhee, and Schwarz, 1998) correlate with biases in the participants' initial beliefs and their updating process.

2. The experiment

The experiment is divided into *five parts*, one of which is randomly selected for payment at the end of the experiment. A total of 143 Northwestern undergraduate students (63 men and 80 women) participated in 14 sessions of 10 to 12 people each. Participants received a $12 show up fee and on average earned around $20.

Upon arrival at the laboratory, subjects are given instructions for part 1. This part consists of performing sums of four two-digit numbers for four minutes. The subjects' earnings in this part equal $0.50 per correctly-answered sum. We chose this task because it has been demonstrated that men and women perform equally well in it (e.g., see Niederle and Versterlund, 2007). After reading the instructions for part 1, subjects perform the adding task. Subjects are informed of their own performance (the number of sums answered correctly). They are not informed of the performance of other subjects. Thereafter, subjects receive the instructions for part 2, which we describe below.

At the beginning of part 2, two subjects are selected at random by drawing cards from a deck.[5] We refer to these subjects as *candidates*. If part 2 is chosen for payment, these two candidates take part in a *contest*. The contest consists of adding sums once again for four minutes. The candidate that answers the most sums correctly wins (in case of a tie, a winner is chosen at random). After they are selected, candidates are asked to stand up for a moment so that everyone in the room can see who they are. Thereafter, candidates simultaneously indicate how many sums they expect to answer correctly if they get to take part in the contest.

[5] To disguise the purpose of the experiment, we do not put any restriction on the gender composition of the pair.

The rest of the subjects in the session make six decisions, which are divided into three information conditions or treatments (two decisions per treatment). The first decision of each treatment consists of guessing the number of sums each candidate will answer correctly (i.e., their expected performance). The second decision of each treatment consists of choosing the candidate who will win the contest. Choosing a candidate increases that candidate's expected earnings. Note that, by eliciting separately the subjects' expectations from their candidate choice, we are able to observe whether subjects have significant taste-based motivations for choosing a candidate.[6]

In the first treatment, subjects do not have any information other than the physical appearance of the two candidates (*no information* treatment). In the second treatment, subjects are informed of the candidates' claim concerning their expected future performance (*cheat talk* treatment). In the third treatment, subjects are informed of the actual performance of each candidate in the first part (*past information* treatment). After all six decisions are made, part 2 finishes and part 3 begins. Parts 3, 4, and 5 are identical to part 2 except that they are done with a different pair of randomly selected candidates (draws are done without replacement).

If it is chosen for payment, earnings in parts 2 through 5 are calculated as follows. In order to avoid hedging between decisions, earnings are determined by randomly selecting one of the two decisions and one of the three treatments. For subjects who are not candidates, if the first decision is selected, they earn between $0 and $8 depending on the accuracy of their guesses concerning the candidates' performance,[7] and if the second decision is selected they earn $8 if they chose the winning candidate and $4 if they chose the losing candidate. Candidates earn $1 per subject that chooses them in the selected treatment. In addition, the candidate that wins the contest receives $18 more.

All instructions are common knowledge. Therefore, candidates know that their claimed performance will be communicated to the other subjects, and the other subjects know that by choosing a candidate they increase the candidate's earnings. In Appendix A, we provide the instructions given to the subjects and a detailed description of the procedures used to run the experiment.

[6] More specifically, unless a preference for choosing men/women is perfectly correlated with the relative expected performance of men and women, we can see whether a significant amount of taste-based discrimination is present our experiment. That is, we can observe whether there are a significant number of cases in which subjects expect one candidate to win and are willing to sacrifice their earnings by choosing the losing candidate (which increases the losing candidate's earnings).

[7] The payment is designed to incentivize risk-neutral individuals to reveal the mean of their distribution. Specifically, for each guess, we take the absolute difference between the subject's guess and the candidate's actual performance. Subjects earn $4.00 if this difference is 0, $3.89 if it is 1, $3.56 if it is 2, $3.00 if it is 3, $2.22 if it is 4, $1.22 if it is 5, and $0.00 if it is 6 or more.

As a final step in our study, we asked all subjects to return to the lab two months after the experiment to complete an Implicit Association Test (IAT) between gender and scientific abilities (Greenwald, McGhee, and Schwartz, 1998).[8] Specifically, we asked subjects to associate male and female pictures with words related to science/math or to liberal arts (for details concerning the IAT see Appendix A). Subjects were informed they could take part in a follow-up experiment and earn $10 more. Subjects did not find out they had to do a gender IAT until they arrived to take the test. Out of 143 subjects, 102 of them returned to take IAT.

3. Results

In total, we have 14 sessions and 56 pairs of candidates (each subject observed either 3 or 4 pairs). Of these pairs, 25 turned out to be pairs composed of a man and a woman. Unless it is otherwise noted, we focus on the data from the mixed-gender pairs. Additional figures and tables are available in Appendix B.

3.1 Performance in the adding task

Like Niederle and Vesterlund (2007), we find that there is no gender difference in performance. In part 1, the average number of sums answered correctly is 11.52 for men and 11.76 for women. We cannot reject the null hypothesis that the distributions of men and women significantly differ with a Mann-Whitney U (MW) test (p = 0.585). The standard deviation in the performance of men is slightly higher (4.66 vs. 3.78), but the difference is not statistically significant (variance ratio test, p = 0.313).[9]

In the contest, the average number of sums answered correctly is 13.38 for men and 13.20 for women.[10] We still do not find a significant difference in performance or in the distributions' standard deviation (MW test, p = 0.853; variance ratio test, p = 0.633). Moreover, even though Wilcoxon signed-rank (WSR) tests indicate that the improvement in performance is significantly different from zero for

[8] The IAT was not done before because we wanted to ensure that no subject found out we were conducting a gender study until all the data was collected.

[9] These averages correspond to performance in mixed-gender pairs. The mean performance of all men is 11.59 sums and that of all women is 11.28 sums. We do not find a significant difference between these two distributions (MW test, p = 0.593) or a significant difference between the standard deviations (variance ratio test, p = 0.936). Figure B1 in Appendix B shows the similarity between the distributions of the men's and women's performance.

[10] These averages and the subsequent tests correspond to the performance of the 16 men and 10 women who participated in a contest. Recall that only one pair of candidates per session took part in the contest (i.e., the pair from the part that was chosen for payment). Unfortunately, we only have contest data for four mixed-gender pairs. The mean performance of men in these pairs is 13.25 sums and for women it is 14.75 sums. There is no significant difference between the two (MW test, p = 0.663).

both genders (p = 0.018 for men, p = 0.018 for women), we do not find a significant difference between the men's improvement and the women's improvement (MW test, p = 0.489).[11]

The subjects' relative performance in the adding task in part 1 is highly predictive of their performance in the contest. For example, the Spearman's rank correlation coefficient between the two is ρ = 0.81 (p < 0.001). Consequently, the relative performance of candidates in part 1 is an excellent predictor of who wins the contest: it predicts the winner of contest 92% of the time.[12] Moreover, this predictive power is equally strong for men and women (for men it predicts the winner 94% of the time and for women 90%). Given this high predictive power, we use the candidates' past performance to determine who the ideal candidate to choose is.

3.2 Initial choices and beliefs

To analyze the subjects' beliefs and choices, we take sessions to be independent observations and make within-session comparisons of session means. Moreover, we report the p-values of one-sided tests for the hypothesis that men are expected to perform better (or are chosen more often) than women.[13]

As seen in Table 1A, when subjects have no information other than the physical appearance of the two candidates, they choose a female candidate only 34% of the time. This is significantly less often than 50%, which is what one would expect if subjects randomize between choosing a man and a woman (WSR test, p = 0.005). In comparison, if subjects knew the past performance of candidates and based their choice solely on this information, they would choose women 47% of the time (see Table 1B).[14] This

[11] Unlike Gneezy, Niederle, and Rustichini (2003) and Gneezy and Rustichini (2004) but in line with Günther et al. (2008) and Dreber, von Essen, Ranehill (2009), we do not find that the relative performance of women worsens when competing in a tournament. Unfortunately, we have too few observations to properly test whether women perform worse when they compete against a man. However, since women improve more on average and perform better than men in the mix-gender pairs (differences are not statistically significant; see footnote 10) our results suggest that in this task competition is not detrimental to the performance of women.

[12] Of the 13 pairs of candidates that played the contest, one pair had candidates who had the same performance in part 1. Of the remaining 12 pairs, in 11 of them the candidate with the best past performance won the contest, and in one of them, there was a tie. In the pair with equal past performances one of the candidates outperformed the other during the contest. Finally, in the 4 pairs with mixed genders, the candidates' past performance perfectly predicted the winner of the contest.

[13] Specifically, we test either H_0: expected performance of men ≤ expected performance of women and H_a: expected performance of men > expected performance of women, or H_0: fraction of men chosen ≤ 0.5, H_a: fraction of men chosen > 0.5.

[14] One could argue that the right comparison is the probability that a randomly chosen woman wins against a randomly chosen man. If we calculate this probability based on the performance of all subjects in part 1, we get that women have a 47% chance of winning. If we calculate it based on the 50 subjects who were candidates in mixed-gender pairs, we get that women have a 54% chance of winning. Using these probabilities does not affect the significance of the reported results.

Table 1 – Descriptive statistics

Note: Panel A presents the mean fraction of women chosen. The ideal full-information choice corresponds to the fraction of women that would be chosen if only the candidates' performance in part 1 is taken into account. Panel B presents the expected performance of the candidates depending on their gender and the gender of the subject doing the guessing. Only mix-gender pairs of candidates are considered. We take each session as an independent observation (i.e., we first take the mean within sessions and then the mean across sessions).

Panel A: Fraction of women being chosen

	Gender of the choosing subject →	All	Male	Female
	Ideal full-information choice	0.47		
	No information	0.34	0.39	0.33
Treatments:	Cheap talk	0.27	0.28	0.27
	Past information	0.39	0.44	0.37

Panel B: Expected performance

	Gender of the choosing subject →	All		Male		Female	
	Gender of the candidate →	Male	Female	Male	Female	Male	Female
	No information	12.38	11.49	12.23	11.50	12.38	11.41
Treatments:	Cheap talk	13.63	11.77	13.71	11.95	13.72	11.78
	Past information	12.47	12.14	12.66	12.54	12.32	12.07

bias in favor of men produces a gender gap in the expected earnings of candidates: women's expected earnings are 19% lower than men's. Note that the propensity to choose men more often is present for both genders and is even slightly stronger for female subjects than it is for males (females choose women only 33% of the time while males choose women 39% of the time).

The subjects' choices are in line with their expectations. A look at Table 1B shows that on average subjects expect men to perform significantly better: 12.38 sums versus 11.49 sums for women (WSR test, p = 0.003). Moreover, we not observe evidence that motivations other an expected performance play a significant role in the subjects' candidate choice. Of all the cases where subjects expect one candidate to outperform the other, there are only 3 cases out of 162 (1.85%) where the candidate with the lower expected performance is chosen.[15] As with their candidate choice, both men and women have on average a biased expectation of the candidate's relative performance: female

[15] The candidate with the lower expected performance is a man in two cases and a woman in one.

subjects expect men to outperform women by 0.97 sums while male subjects expect men to outperform women by 0.73 sums. Moreover, the subjects' mean difference between the expected performance of women and the expected performance of men is not significantly correlated with the subjects' own performance (Spearman's ρ = 0.02, p = 0.884 for men, and ρ = –0.02, p = 0.861 for women).[16] In other words, even high-performing women do not expect other women to outperform men.

Since the biased expectations concerning the relative performance of women is not justified by actual gender differences in performance, either in this sample nor in the vast literature on gender differences in arithmetic (see, Hyde, Fennema, and Lamon, 1990), we are interested in studying how it is related to prejudices against women. In other words, we check whether there is a relationship between the subjects' expectations and their score in an Implicit Association Test (IAT) between gender and scientific/mathematical traits (see Appendix A for details).

The mean IAT score is 0.42, which reveals that, on average, subjects have more difficulties associating women with science/math than associating men with science/math—positive numbers indicate women are associated less with science/math whereas negative numbers indicate men are associated less with science/math. The IAT score is positive and significantly different from zero for both men (mean score of 0.45; WSR test, p < 0.001) and women (mean score 0.40; WSR test, p < 0.001).[17] The distribution of IAT scores and other descriptive statistics are available in Appendix B.

In order to test the relationship between the subjects' expectations and their IAT score, we calculate the Spearman's rank correlation coefficient between the latter and the mean difference between the expected performance of women and the expected performance of men. For the whole sample, there is a negative correlation that is not statistically significant (ρ = –0.09, p = 0.438). However, if we restrict the sample to only male subjects, the correlation between the subjects' expectations and their IAT score is strong and statistically significant (ρ = –0.43, p = 0.019). By contrast, there is no correlation for female subjects (ρ = 0.02, p = 0.892). This can also be seen in Figure 1 where we plot the IAT score against the difference in expectations between the performance women and men. The prediction lines are calculated with OLS regressions, which are available in Appendix B. Interestingly, in the case of men, the expected difference between the performance of women and men for unbiased

[16] These correlation coefficients are calculated using subject means. The correlation coefficients using session means are ρ = 0.08 (p = 0.810) for men and –0.15 (p = 0.649) for women.

[17] As with the subjects' expectations, neither the IAT score of men nor of women is significantly correlated with the subjects' own performance (Spearman's ρ = 0.15, p = 0.358 for men, and ρ = –0.08, p = 0.528 for women). That is, even high-performing women associate science/math more with men than with women.

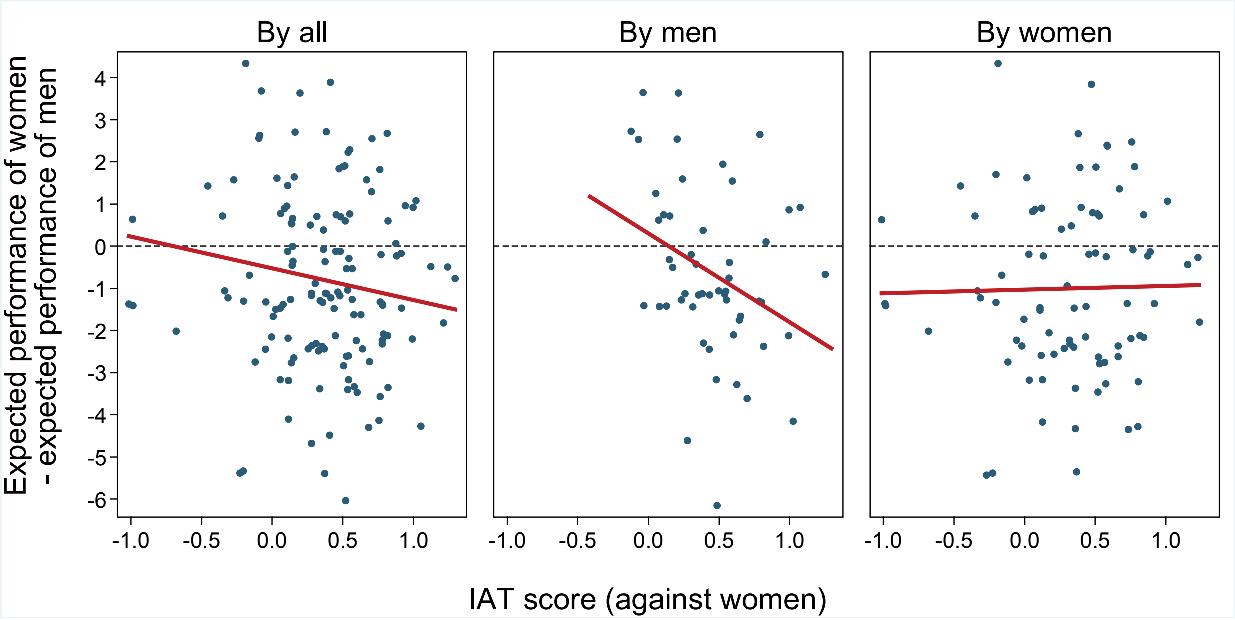


Figure 1 – IAT score and beliefs of the relative performance of men and women

Note: The lines are calculated from a regression with the difference between the expected performance of women and the expected performance of men as the dependent variable. The independent variables are the IAT score, gender, and interaction terms. OLS estimates with subject random effects and session fixed effects (see Appendix B).

individuals (i.e., with an IAT score of zero) is predicted to not be significantly different from zero (Wald test, p = 0.692).

3.3 Successive choices and updated beliefs

In real life people do not rely only on their priors but try to integrate them with extra information. We study the effect of additional information with our two other treatments.

We start with the cheap talk treatment. In spite of being highly correlated with their actual performance in the part 1 (Spearman's ρ = 0.59, p < 0.001), revealing the candidates' self-reported claims results in less women being chosen. As seen in Table 1A, women are now chosen only 27% of the time, which is significantly less than the 34% in the no-information treatment (WSR test, p = 0.017). This deterioration seems to be due to subjects treating the claims of men and women similarly, which ignores two facts. First, the claims of female candidates are significantly lower than those of male candidates (12.33 sums versus 14.82 sums; MW test, p = 0.028). Second, women's claims are more informative than those of men's. The Spearman's rank correlation coefficients between claims and actual performance are ρ = 0.66 for women and ρ = 0.56 for men.

In the past-information treatment, even with precise information of the candidates' relative performance, women candidates are chosen only 39% of the time (see Table 1A). This is significantly less than the 47% of women that would be chosen if subjects base their decision solely on the candidates' past performance (WSR test, $p = 0.038$).[18] Hence, we find a persistent bias against choosing women even after subjects have the opportunity to update their initial beliefs.

Next we take a closer look at how subjects update their beliefs concerning the relative performance of women. In particular, we use the IAT score to analyze whether the updating of beliefs is affected by implicit prejudices against women. To have a measure of the degree to which subjects update their expectations we calculate the following variable:

$$\varphi_{ik} = (\sigma_{ik} - \mu_i)/(s_{ik} - \mu_i)$$

where μ_i is subject *i*'s expected difference in performance between a woman and a man in the no-information treatment (i.e., *i*'s prior belief), s_{ik} is the information observed by *i* in treatment *k* concerning the difference in performance (i.e., either the difference in claimed performance or the difference in past performance), and σ_{ik} is *i*'s expected difference in performance after observing s_{ik} (i.e., *i*'s posterior belief). Note that, on the one hand, if φ_{ik} equals one, it indicates that *i*'s updated belief equals the information observed in treatment *k*, which is consistent with *i* treating his prior belief as being completely uninformative. On the other hand, if φ_{ik} equals zero, it indicates that *i* did not update his belief at all, which is consistent with *i* treating the observed information in treatment *k* as completely uninformative. The mean value of φ_{ik} is 0.60 in the cheap talk treatment and 0.77 in the past performance treatment.[19] We provide the complete distribution of φ_{ik} in Appendix B.

To evaluate the effect of the IAT score on the updating process we ran a regression with φ_{ik} as the dependent variable. As independent variables we use: a dummy variable indicating who the better-performing candidate is according to the new information (i.e., it equals one if the woman's claimed/past performance is higher than the man's and zero if the opposite is true),[20] two interaction variables that equal the subject's IAT score when the better-performing candidate is the woman (or the man) and zero otherwise, and a dummy variable indicating whether the observed information

[18] The hypothesis being tested is H_0: fraction of women picked≥ fraction of women predicted to win based on their past performance, H_a: fraction of women picked < fraction of women predicted to win based on their past performance.

[19] In the cheap talk treatment, around 18% of subjects do not update their expectation ($\varphi_{ik} = 0$) and 25% of them update as if their prior belief was completely uninformative ($\varphi_{ik} = 1$). In the past information treatment the respective numbers are 12% for $\varphi_{ik} = 0$ and 33% for $\varphi_{ik} = 1$.

[20] We excluded observations where the woman and the man have the same performance according to the new information because it is not clear whether this information confirms/contradicts the subjects' implicit associations.

Table 2 – Degree to which expectations are updated and the IAT score

Note: The table presents regressions with φ_{ik} as the dependent variable. We use OLS estimates with robust standard errors and subject random effects.

	All		*Men*		*Women*	
	coef.	std. err.	coef.	std. err.	coef.	std. err.
Constant	0.537***	(0.060)	0.546***	(0.115)	0.571***	(0.066)
Woman is better	0.048	(0.073)	–0.170	(0.145)	0.141*	(0.079)
Woman is better × IAT score	–0.071	(0.069)	0.205	(0.210)	–0.136**	(0.065)
Man is better × IAT score	0.230**	(0.104)	0.141	(0.162)	0.242**	(0.114)
Cheap talk	–0.156***	(0.055)	–0.281***	(0.088)	–0.044	(0.072)
Number of observations	214		84		130	
Number of subjects	74		29		45	
R^2	0.050		0.126		0.058	

corresponds to the cheap-talk (=1) or the past-information (=0) treatment. We ran a regression for all subjects and then a separate regression for each gender (see Table 2).[21]

Subjects with a high IAT score update significantly more in cases where they observe information that conforms with their implicit association (the man is the better-performing candidate) compared to subjects with a low IAT score (p = 0.028) and compared to cases where they observe information that contradicts their implicit association (the woman is the better-performing candidate, p = 0.010). As seen in Table 2, this relationship between the IAT score and updating is driven by women and not men. As one would expect, subjects consider the candidates' past performance to be more informative than the candidates' claimed performance.

4. Conclusions

In this paper, we provide evidence of discrimination against women that is neither taste-based, nor driven by a statistical inference that is justified by actual differences in performance, and instead, it is rooted in a biased belief about women's abilities. In spite of equal performance, only about one woman

[21] We use OLS estimates with robust standard errors and subject random effects. Moreover, we excluded observations with negative values for φ_{ik} since these subjects seem to be updating irrationally (less than 3% of all observations correspond to φ_{ik} < 0).

is chosen for every two men.[22] Furthermore, the bias against choosing women persists, albeit reduced, when very accurate information about the performance of women is provided.

We also observe the paradoxical result that more information in the form of self-reports leads to a worse outcome (from the perspective of gender equality). The reason for this result is that men seem to boast more about their future performance thereby garbling the quality of information, and yet, participants do not account fully for this when making their choice. Thus, participants handicap women when they should not, and they do not handicap men when they should.

Among men, the degree to which initial beliefs are biased against women is strongly correlated with stereotypes (implicit associations) measured by the IAT. The same is not true for women. Nevertheless, among women these implicit associations seem to impair the degree to which they update their beliefs when they observe new information. There is a lively discussion on how to interpret IAT scores and to what extent they explain behavior (Greenwald, et al., 2009). However, as argued by Bertrand, Chugh, and Mullainathan (2005), there is compelling evidence that the IAT captures unconscious processing of information that is distinct from conscious reasoning. That is, probably some of our participants were discriminating against women without realizing it, which is a very different form of discrimination to the ones that are normally modeled in economics. Importantly, discrimination driven by implicit associations would require different (less coercive) policies to try to remedy it (see, Bertrand, Chugh, and Mullainathan, 2005).

References

Albanesi, Stefania, and Claudia Olivetti (2008). Gender and Dynamic Agency: Theory and Evidence on the Compensation of Top Executives. Working Paper.

Altonji, Joseph G., and Rebecca M. Blank (1999). Race and Gender in the Labor Market. In O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics, Volume 3*, pp. 3143–3259. Amsterdam, North-Holland: Elsevier.

Akerlof, George (1985). Discriminatory, Status Based Wages Among Tradition-Oriented, Stochastically Trading Coconut Producers. *Journal of Political Economy* 92: 265-276.

[22] While the gender gap in choosing women might appear small in this study compared to some of the gender gaps observed in the field, it is sufficient to generate the very low frequency of women CEOs reported in the introduction if we assume that only immediate subordinates are considered for promotion to the next level of a hierarchy. A 34%-66% bias in each promotion leads to a 3.5% presence of women at the top of a five-layer organization.

Albanesi, Stefania, and Claudia Olivetti (2008). Gender and Dynamic Agency: Theory and Evidence on the Compensation of Top Executives. Working paper.

Anderson, Lisa R., Roland G. Fryer, and Charles A. Holt (2005) Discrimination: Evidence from Psychology and Economics Experiments. In William Rogers (ed.), *Handbook on Economics of Discrimination*.

Antonovics, Kate, Peter Arcidiacono and Randall Walsh (2005). Games and Discrimination: Lessons from The Weakest Link. *Journal of Human Resources* 40(4): 918-947.

Arrow, Kenneth J. (1973). The Theory of Discrimination. In Orley Ashenfelter and Albert Rees (eds.), *Discrimination in Labor Markets*, pp. 3-33. Princeton NJ: Princeton University Press.

Arulampalam, W., Alison L. Booth, and M.L. Bryan (2007). Is there a glass ceiling over Europe? Exploring the gender pay gap across the wages distribution. *Industrial and Labor Relations Review* 60(2): 163-186.

Azmat, G., M. Güell, and A. Manning (2006). Gender gaps in unemployment rates in OECD countries. *Journal of Labor Economics* 24: 1-37.

Babcock, Linda, and Sara Laschever (2003). *Women Don't Ask: Negotiation and the Gender Divide*. Princeton, NJ: Princeton University Press.

Becker, Gary S. (1971). *The Economics of Discrimination*. Chicago, IL: The University of Chicago Press.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan (2005). Implicit Discrimination. *American Economic Review Papers and Proceedings* 95(2): 94-98.

Bertrand Marianne, Lawrence Katz, and Claudia Goldin (2009). Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors. Working Paper.

Bertrand, Marianne, and Kevin Hallock (2001). The Gender Gap in Top Corporate Jobs. *Industrial and Labor Relations Review* 55(October): 3-21.

Bertrand, Marianne, and Sendhil Mullainathan (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94(4): 991-1013.

Booth, Alison L. (2009). Gender and Competition. Discussion Paper No. 4300, IZA.

Bravo, David, Claudia Sanhueza, and Sergio Urzua (2008). An Experimental Study of Labor Market Discrimination: Gender, Social Class and Neighborhood in Chile. RES Working Paper No. 3242, Inter-American Development Bank.

Coate, Stephen, and Glenn C. Loury (1993). Will Affirmative Action Policies Eliminate Negative Stereotypes? *American Economic Review* 83(5): 1220–1240.

Dreber, Anna, Emma von Essen, and Eva Ranehill (2009). Outrunning the Gender Gap—Boys and Girls Compete Equally. SSE/EFI Working Paper No. 709, Stockholm School of Economics.

Gneezy, Uri, and Aldo Rustichini (2004). Gender and Competition at a Young Age. *American Economic Review Papers and Proceedings* 94(May): 377-381.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini (2003). Performance in Competitive Environments: Gender differences. *Quarterly Journal of Economics* 118(August): 1049-1074.

Gneezy, Uri, Kenneth Leonard, and John A. List (forthcoming). Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society. *Econometrica*.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74(6): 1464-1480.

Greenwald, Anthony G., Poehlman, T.A., Uhlmann, E., 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: 17-41.

Goldin, Claudia, and Cecilia Rouse (2000). Orchestrating Impartiality: The Impact of “Blind” Auditions on Female Musicians. *American Economic Review* 90(4): 715-741.

Günther, Christina, Neslihan Arslan Ekinci, Christiane Schwieren, and Martin Strobel (2008). Women can't jump? An experiment on competitive attitudes and stereotype threat. Working Paper, University of Heidelberg.

Heckman, James (1998). Detecting Discrimination. *Journal of Economic Perspectives* 12(2): 101-116.

Hyde, Janet Shibley, Elizabeth Fennema, and Susan J. Lamon (1990). Gender Differences in Mathematics Performance: A Meta-Analysis. *Psychological Bulletin* 107(2): 139-155.

Ichino, Andrea, and Enrico Moretti (2009). Biological Gender Differences, Absenteeism and the Earning Gap. *American Economic Journal: Applied Economics* 1(1):.

Levitt, Steven (2004). Testing Theories of Discrimination, Evidence from ‘The Weakest Link.’ *Journal of Law and Economics* 47: 431-452.

Lundberg, Shelly J., and Richard Startz (1983). Private Discrimination and Social Intervention in Competitive Labor Markets. *American Economic Review* 73(3): 340-347.

Phelps, Edmund (1972). The Statistical Theory of Racism and Sexism. *American Economic Review* 62: 659-661.

Milgrom, Paul, and Sharon Oster (1987). Job Discrimination, Market Forces, and the Invincibility Hypothesis. *Quarterly Journal of Economics* 102(August): 453-476.

Newmark, David, Roy J. Bank and Kyle D. Van Nort (1996). Sex Discrimination in Restaurant Hiring: An Audit Study. *Quarterly Journal of Economics* 111(3): 915-941.

Niederle, Muriel, and Lise Vesterlund (2007). Do Women Shy Away from Competition? Do Men Compete too Much? *Quarterly Journal of Economics* 122(August): 1067-1101.

Wolfers, Justin (2006). Diagnosing Discrimination: Stock Returns and CEO Gender. *Journal of the European Economic Association* 4(May): 531-41.

Appendix A – Experimental procedures and instructions

This section is divided as follows. In subsection A.1, we describe in detail the procedure followed to run the experiment. In subsection A.2, we reproduce the instructions used in the first part of the experiment and in subsection A.3, the instructions used for parts 2 through 5. Finally, in subsection A.4, we describe the IAT and the procedure used to give it to the subjects.

A.1 Experimental procedures

The computerized experiment was conducted in 2008 in the laboratory of the Kellogg School of Management in Northwestern University. Subjects were recruited through the recruitment website and the experiment was programmed with z-Tree (Fischbacher, 2007). The experiment lasted around 1 hour. In total, 206 subjects participated in the experiment. None of the subjects had previously participated in a similar experiment. Average earnings, including a $12 show-up fee, were approximately $20.

Upon arrival to the lab, subjects were asked to take a seat in the laboratory where they could read and sign the study's consent form. Once subjects had consented, they were given the instructions for part 1 of the experiment (see subsection A.1). Subjects were told that the experiment consisted of five parts and that they would be paid their earnings from one randomly-selected part. Thereafter the subjects performed part 1.

Once part 1 was complete and subjects were informed of the number of sums they answered correctly, they received the instructions for the four remaining parts (see subsection A.3). After reading these instructions, we asked them to answer a series of questions to ensure they correctly understood them. Once everyone finished answering the control questions, we started with part 2.

At the beginning of the part 2, we asked subjects to take a card from a deck. The deck had once card per subject and was shuffled in front of the subjects after showing them that the deck had two aces. Subjects that drew an ace were assigned to be candidates in that part. Candidates were asked to stand up holding a sign indicating whether they were Candidate A or Candidate B. All other subjects were asked to look at the candidates before making their decisions. After everyone saw them, the candidates could sit down in order to enter their claims. During part 2, subjects who were not candidates made two decisions per treatment. Each pair of decisions was made simultaneously on the screen (see subsection A.3). Subjects never received feedback concerning the choices of others. Once everyone finished entering their decisions we continued to the next part.

Parts 3, 4, and 5 are repetitions of part 2. The only difference is that when selecting the candidates, only subjects who had never been a candidate in a previous part could draw a card from the deck. In other words, a subject could be a candidate at most once.

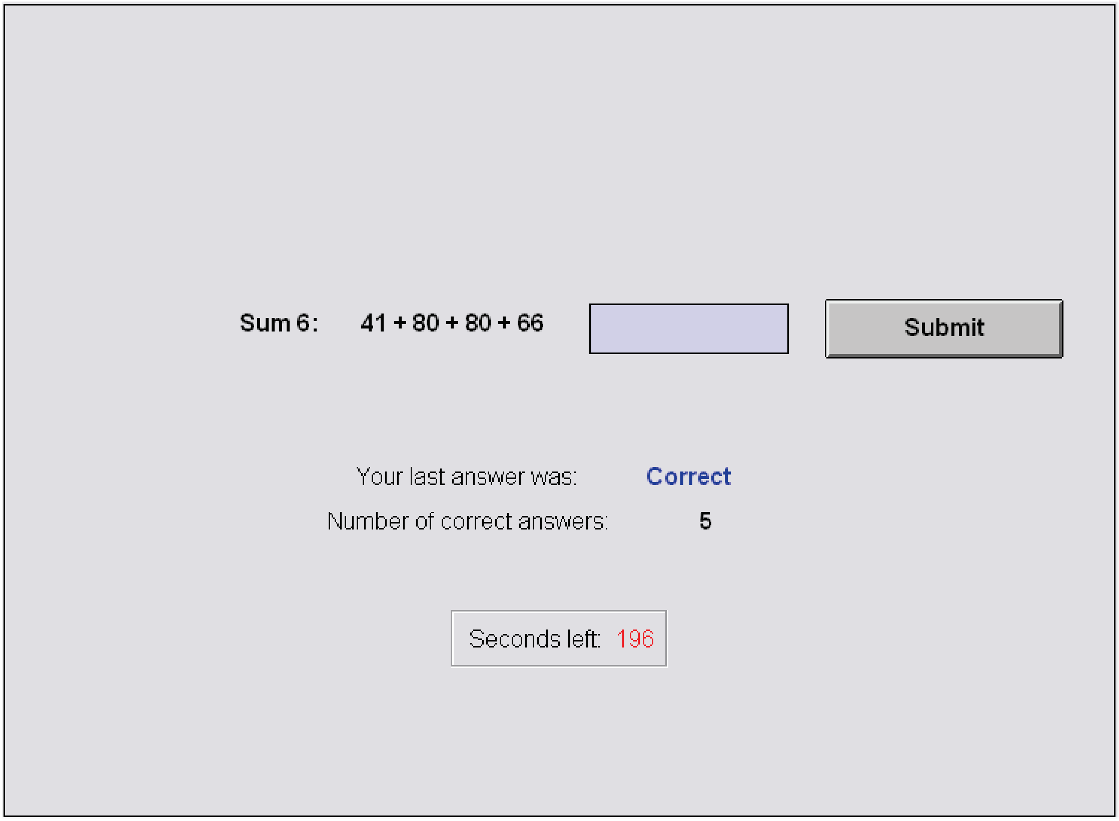
Once part 5 had finished, we used the deck of cards to randomly select a part to be paid (one of the subjects in the session drew one out of five cards). If the part to be paid was not part 1, we used the deck of cards once again to determine the decision to be paid (another subject in the session drew one out of six cards). Next, the two candidates of the part that was randomly selected for payment were asked to add sums for four minutes (other subjects could watch the candidates' progress on their screens). Thereafter, subjects were paid their earnings and dismissed.

A.2 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. If part 1 is the part randomly selected for payment, then you will get **$0.50 for every correct sum**. Your payment does 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.

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



Sum 6: 41 + 80 + 80 + 66

Your last answer was: Correct

Number of correct answers: 5

Seconds left: 196

A.3 Instructions for parts 2 to 5

Parts 2 to 5 are identical. Before each part, two participants will be selected at random. We will call them **contender A** and **contender B**. We will call the rest of you **observers**.

Specifically, everyone will draw a card from the deck on the table. Contender A will be the one who draws the **red ace** and contender B will be the one who draws the **black ace**. Each participant gets to be a contender at most once. Hence, in parts 3 to 5 those of you that were contenders in previous parts will not draw a card and will play as observers.

Contenders

As mentioned, at the end of the study, we will randomly select a part to be paid. If part 2, 3, 4, or 5 is selected, the two participants that were Contender A and contender B in the selected part will have another four minutes to answer sums. This time, however, their earnings depend on their relative performance. The contender who correctly answers the most sums will be the **winning contender**. In case of a tie, one contender will be randomly selected as the winner. The winning contender earns **$18** and the other contender earns **$0**.

Note that both contenders will face the same sequence of randomly generated sums. That is, they will face the same difficulty.

Contenders can earn additional money depending on the decisions of the observers. This is explained further down.

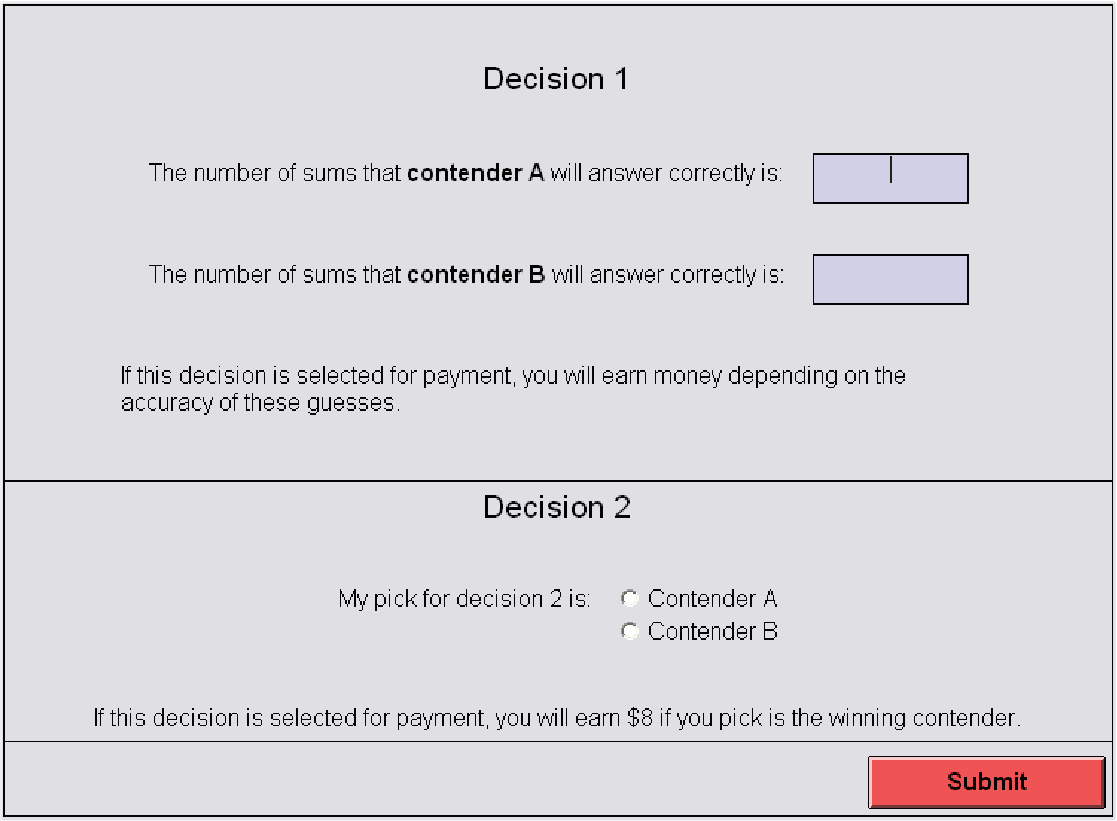
Observers

In each part, observers make **six decisions**. Decisions consist of either: (i) accurately guessing the number of sums that each contender answers correctly, or (ii) picking the winning contender.

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

Observers: Decisions 1 and 2

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



Decision 1

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

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

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

Decision 2

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

If this decision is selected for payment, you will earn $8 if you pick is the winning contender.

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. 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)	$4.00
1 sum away	$3.89
2 sums away	$3.56
3 sums away	$3.00
4 sums away	$2.22
5 sums away	$1.22
6 sums away or more	$0.00

On the bottom part of the screen, you make **decision 2**. This decision consists on **picking one of the two contenders**. If the contender you picked becomes the winning contender, you earn **$8**. If your pick does not become the winning contender, you earn **$4**.

Additional earnings for contenders

Contenders earn additional money depending on the number of observers that pick them. If observers are paid according to decision 1 or 2 in a given part, every observer that picks contender A in this screen, increases A’s earnings by **$1**. Similarly, every observer that picks contender B in this screen, increases B’s earnings by **$1**. Contenders receive the additional earnings independently of whether they win or not.

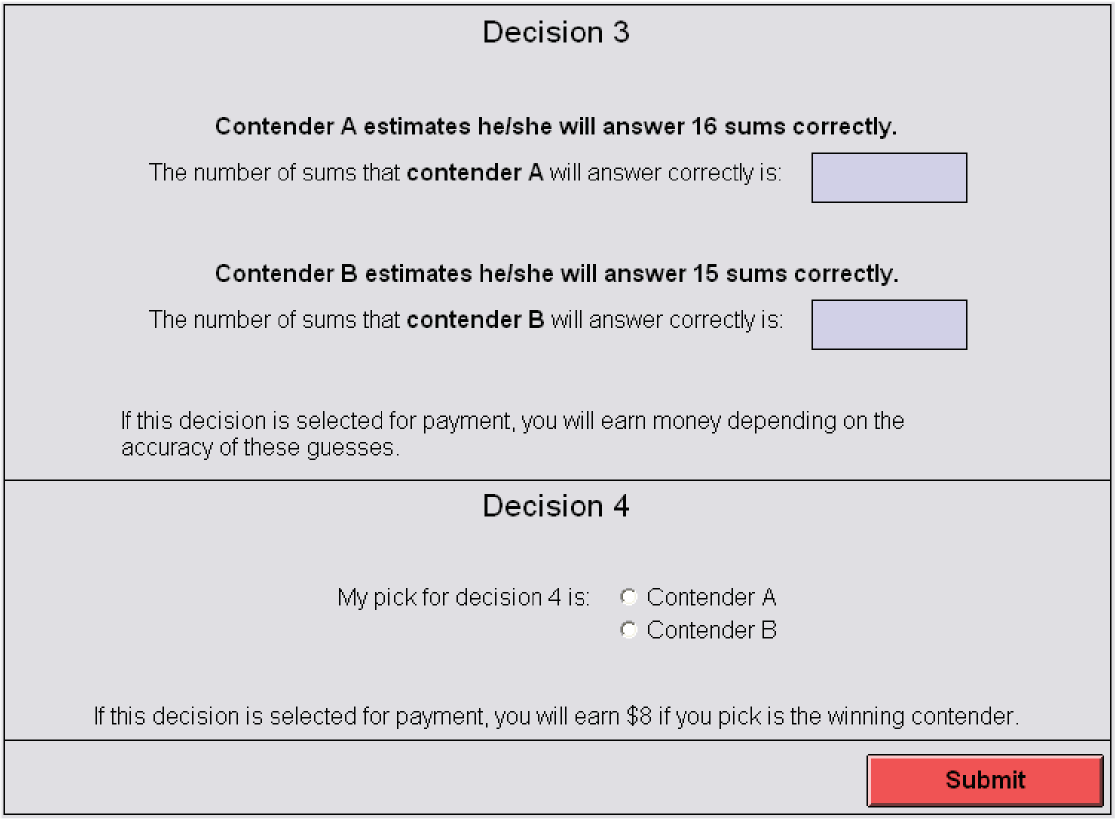
Observers: Decisions 3 and 4

Before decisions 3 and 4, **contenders** will be asked the following question:

- Indicate below the number of sums that you think you will answer correctly if you attempt the summing task once again.

Contenders can then submit any number.

If you are an observer, you will make decisions 3 and 4 on the following screen:



Decision 3

Contender A estimates he/she will answer 16 sums correctly.

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

Contender B estimates he/she will answer 15 sums correctly.

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

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

Decision 4

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

If this decision is selected for payment, you will earn $8 if you pick is the winning contender.

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. 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 abovementioned question**.

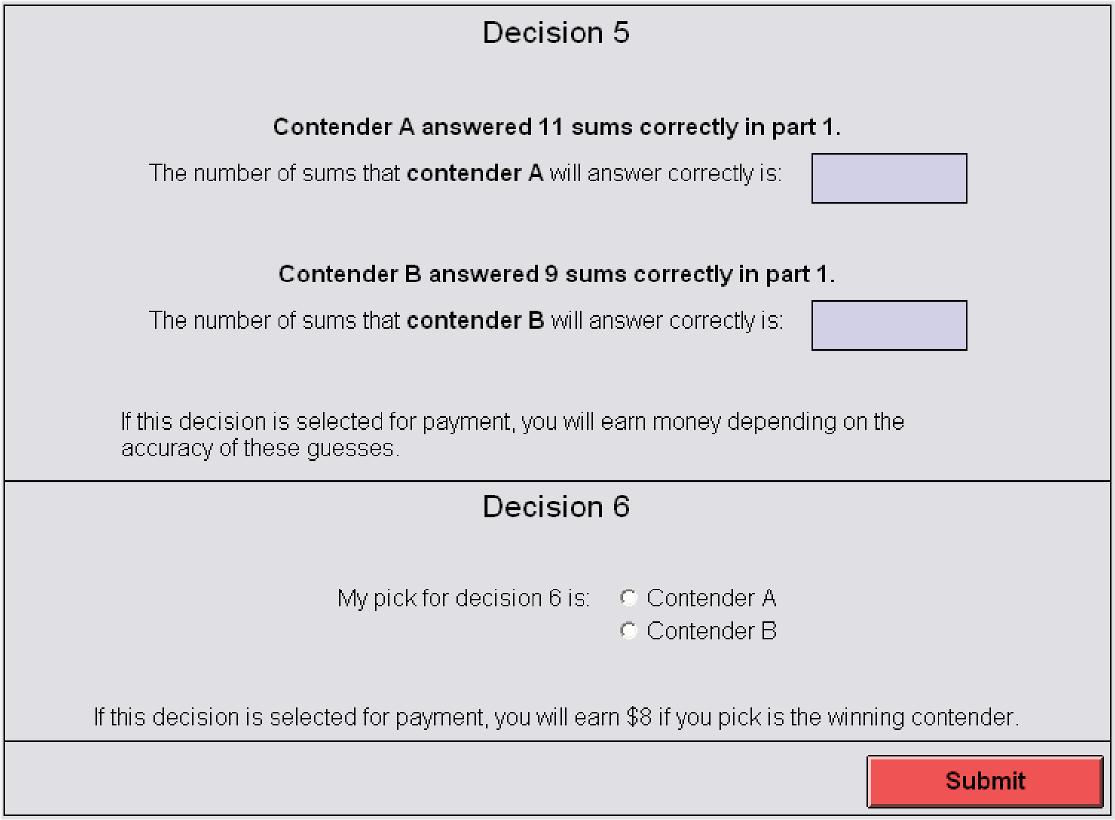
On the bottom part of the screen, you make **decision 4**. Again, you are asked to pick one of the two contenders. If the contender you picked becomes the winning contender, you earn **$8**. If your pick does not become the winning contender, you earn **$4**.

Additional earnings for contenders

As before, contenders earn additional money depending on the number of observers that pick them. If observers are paid according to decision 3 or 4 in a given part, every observer that picks contender A in this screen, increases A’s earnings by $1. Similarly, every observer that picks contender B in this screen, increases B’s earnings by $1. Contenders receive the additional earnings independently of whether they win or not.

Observers: Decisions 5 and 6

If you are an observer, you will make decisions 5 and 6 on the following screen:



Decision 5

Contender A answered 11 sums correctly in part 1.

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

Contender B answered 9 sums correctly in part 1.

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

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

Decision 6

My pick for decision 6 is: ○ Contender A ○ Contender B

If this decision is selected for payment, you will earn $8 if you pick is the winning contender.

On the top part of the screen, you make **decision 5**. You are asked once again to guess the number of sums that each contender will answer correctly. Your earnings depend on the accuracy of your guesses according to the same table as in decision 1. Note that, unlike in decision 1 and 3, you can also see the **number of sums that each contender answered correctly in part 1**.

On the bottom part of the screen, you make **decision 6**. Again, you are asked to pick one of the two contenders. If the contender you picked becomes the winning contender, you earn **$8**. If your pick does not become the winning contender, you earn **$4**.

Additional earnings for contenders

As before, contenders earn additional money depending on the number of observers that pick them. If observers are paid according to decision 5 or 6 in a given part, every observer that picks contender A in this screen, increases A’s earnings by $1. Similarly, every observer that picks contender B in this screen, increases B’s earnings by $1. Contenders receive the additional earnings independently of whether they win or not.

Example of how to calculate earnings

Suppose that you are an observer in part 2 and that this part is picked for payment. Furthermore, in decision 1 you guessed that contender A will answer 10 sums correctly and contender B will answer 14 sums correctly. In decision 2 you picked contender B.

If it turns out that contender A answered 8 sums correctly and contender B answered 11 sums correctly, then:

- If decision 1 is selected for payment, your earnings would be: $3.56 for your guess of A's performance + $3.00 for your guess of B's performance + the $12.00 show-up fee = $18.56.
- If decision 2 is selected for payment, your earnings would be: $8.00 for picking the winning contender + the $12.00 show-up fee = $20.00.

For the earnings of contenders, suppose that in addition to you, contender B was picked by 5 other observers in decision 2 and contender A was picked by 4 observers in decision 2. In this case, if decision 1 or 2 are selected for payment, Contender B's earnings would be: $18.00 for being the winning contender + $6.00 for being picked by 6 observers + the $12.00 show-up fee = $36.00, and contender A's earnings would be: $0.00 for not being the winning contender + $4.00 for being picked by 4 observers + the $12.00 show-up fee = $16.00.

Final note

Note that when they perform the sums, contenders will **not know** how many observers have picked them. This will be revealed after they finished answering sums. 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.

A.4 Implicit association test

We used the IAT introduced by Greenwald, McGhee, and Schwartz (1998). In particular, we asked subjects to associate pictures to the categories “male” or “female” and to associate words to the categories “math and science” or “liberal arts.” The precise words used for “math and science” are: physics, engineering, chemistry, biology, statistics, geometry, calculus, and algebra, and the words used for “liberal arts” are: literature, music, philosophy, writing, history, arts, civics, and humanities. Pictures are not reproduced in this document due to copyright but are available upon request. Figure A1, provides a sample screenshot of the IAT.

The IAT serves as an indirect measure of associations between different categories (in contrast to directly asking subjects to self-report them). The advantage of the IAT over self-reported measures is that in socially sensitive domains subjects might be reluctant to report biased associations. In fact, it is in this domain where the IAT has been found to have greater predictive ability than self-reported measures (Greenwald et al., 2009).

The IAT score is constructed by comparing response times in a task that requires rapid classification of images by pressing one of two keys in a keyboard. For example, subjects that are significantly 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 are classified as having an automatic association between math/science and males relative to females. To calculate the IAT score, we used the new scoring algorithm described in Greenwald, Nosek, and Banaji (2003).

In order to have subjects take the IAT, two months after their participation in the experiment, we sent them an email indicating that they could take part in a follow up to the experiment and earn $10 additional dollars. Subjects who were interested in participating in the follow up were asked to come one of our offices. In order to avoid selection problems, when asked to participate, we did not inform subjects that they would take part in a gender IAT test. Subjects took the IAT test individually over a period of two weeks. Taking the test takes approximately 15 minutes.

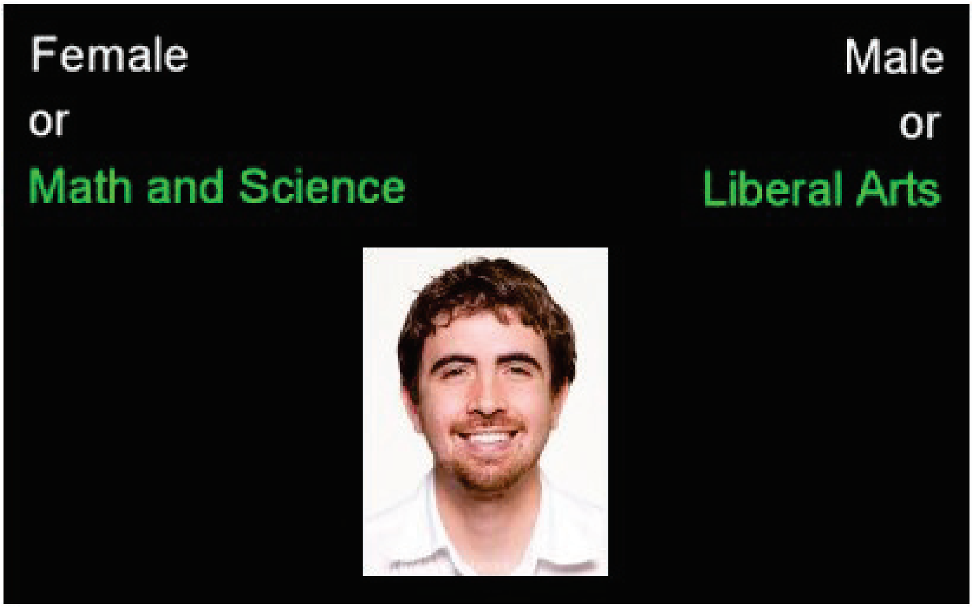


Figure A1 – Screenshot of the IAT

Appendix B – Additional data analysis

In this section we provide additional figures and tables to support the data analysis described in the main body of the paper.

Figure B1 shows the cumulative distribution of the subjects' performance in the adding task divided by gender (using the performance of all subjects). As can be seen, the distributions of men and women are practically identical. The means of the distributions are 11.59 sums for men and 11.28 sums for women. The standard deviations are 3.74 sums for men and 3.70 sums for women.

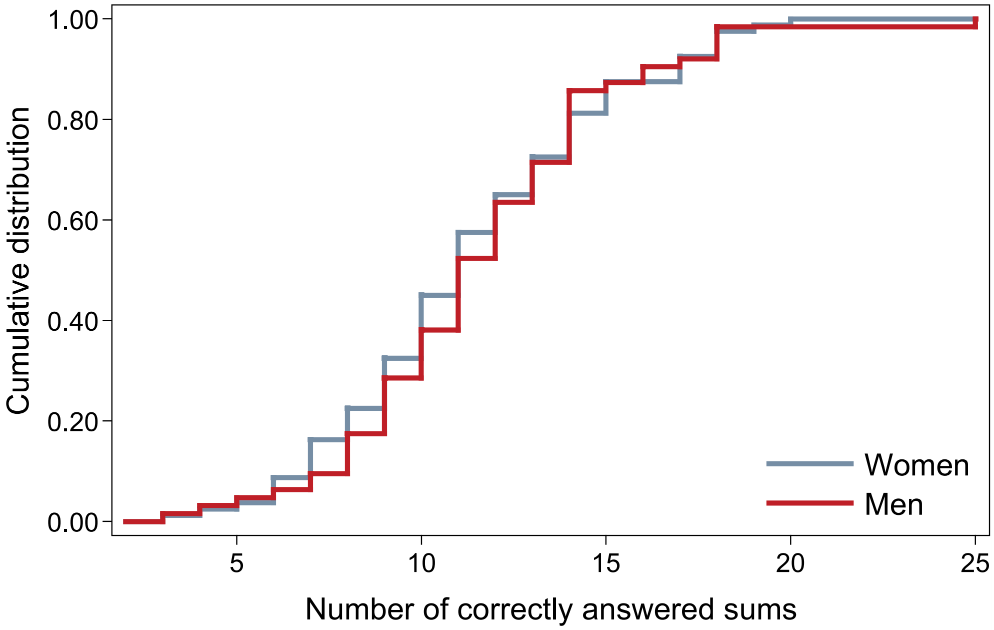


Figure B1 – Cumulative distribution of performance in the adding task by gender

Figure B2 shows the cumulative distribution of the subjects IAT score divided by gender. As can be seen, the distributions of men and women are similar. The means of the distributions are 0.45 for men and 0.40 for women. The standard deviations are 0.33 for men and 0.47 for women. We do not find a statistically significant difference in means (*t*-test, *p* = 0.589). However, we do find a significant difference in the standard deviations (variance ratio test, *p* = 0.021). A look at the figure reveals that this is due to more women than men having a negative IAT score (i.e., associating men and less with science/math than women).

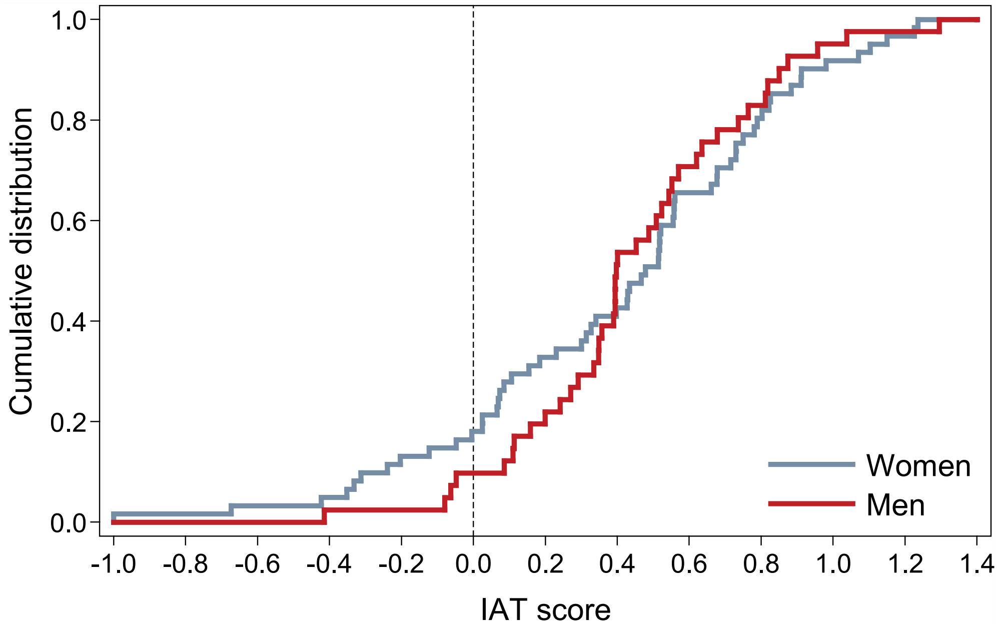


Figure B2 – Cumulative distribution of IAT scores by gender

Table B1 reports the regression used to construct Figure 1 in the main body of the paper. The dependent variable is the difference between a subject's expected performance of the female candidate and the subject's expected performance of the male candidate. We use data only from parts in which candidates were of different genders. As independent variables we use the subject's IAT score, a dummy variable indicating the (choosing) subject's gender and a slope-dummy variable interacting gender and IAT score. We use OLS estimates with robust standard errors. Moreover, we include subject random effects and session fixed effects (to control for the candidates' other characteristics). In total, we have 136 observations, for 78 subjects (between 1 and 4 observations per subject), and 14 sessions (between 3 and 27 observations per session).

Table B1 – Expected difference in performance of women compared to men and IAT score

	coefficient	standard error	*p*-value
IAT Score	-2.105	1.052	0.045
Female	-1.331	0.611	0.029
Female × IAT Score	2.194	1.177	0.062
R^2	0.237		

Figure B3 shows the distribution of the variable φ_{ik} (defined in the main body of the paper) divided by whether the new information corresponds to the difference in the candidates' claimed performance or to the difference in their actual past performance. Recall that φ_{ik} = 1 corresponds to the case where the subjects' updated belief equals their new information (i.e., their prior belief is completely uninformative), and φ_{ik} = 0 corresponds to the case where beliefs are not updates at all (i.e., the new information is completely uninformative).

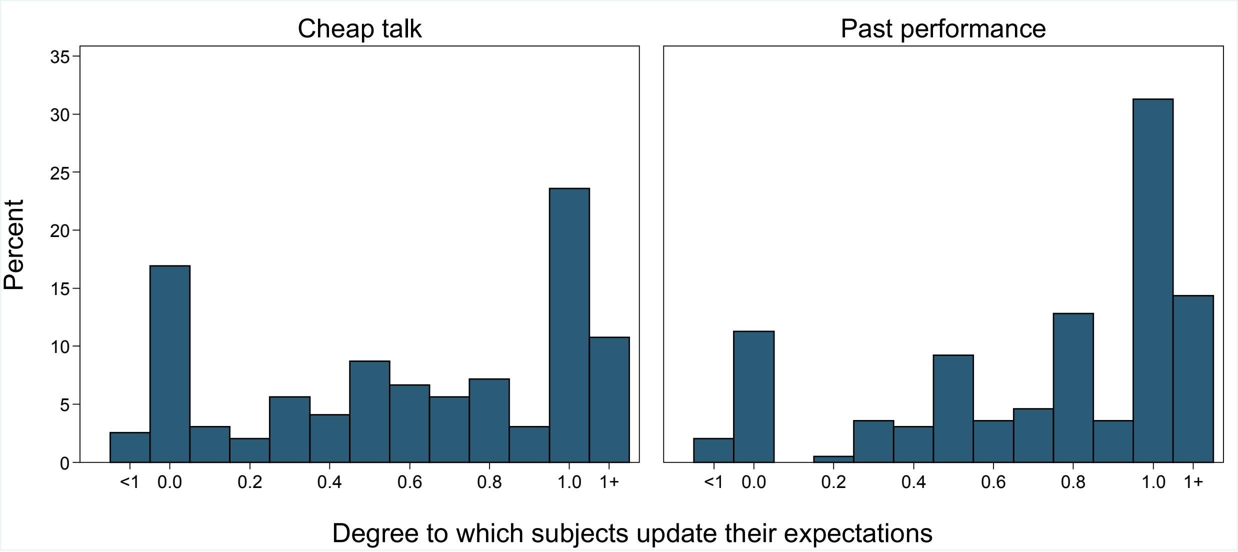


Figure B3 – Distribution for φ_{ik}, which measures the degree to which subjects update their expectations of relative the performance of men and women with the arrival of new information

Appendix References

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10: 171-178.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74(6): 1464-1480.

Greenwald, Anthony G., Nosek, B.A., and Banaji, M.R. (2003). Understanding and Using the Implicit Association Test: An Improved Scoring Algorithm. *Journal of Personality and Social Psychology* 85: 197-216.

Greenwald, Anthony G., Poehlman, T.A., Uhlmann, E., 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: 17-41.