When Performance Trumps Gender Bias:
Joint Versus Separate Evaluation

IRIS BOHNET

HARVARD KENNEDY SCHOOL, CAMBRIDGE, MASSACHUSETTS, IRIS_BOHNET@HARVARD.EDU

ALEXANDRA VAN GEEN

ERASMUS SCHOOL OF ECONOMICS, ROTTERDAM, THE NETHERLANDS, VANGEEN@ESE.EUR.NL

MAX BAZERMAN

HARVARD BUSINESS SCHOOL, CAMBRIDGE, MASSACHUSETTS, MBAZERMAN@HBS.EDU

Gender bias in the evaluation of job candidates has been demonstrated in business, government and academia, yet little is known about how to overcome it. Blind evaluation procedures have been proven to significantly increase the likelihood that women musicians are chosen for orchestras and are employed by a few companies. We examine a new intervention to overcome gender bias in hiring, promotion, and job assignments: an "evaluation nudge" in which people are evaluated jointly rather than separately regarding their future performance. Evaluators are more likely to base their decisions on individual performance in joint than in separate evaluation and on group stereotypes in separate than in joint evaluation, making joint evaluation the profit-maximizing evaluation procedure. Our work is inspired by findings in behavioral decision research suggesting that people make more reasoned choices when examining options jointly rather than separately and is compatible with a behavioral model of information processing.

Key words: gender; behavioral economics; decision making; performance evaluation; laboratory experiments.

1. Introduction

Gender-based discrimination in hiring, promotion, and job assignments is difficult to overcome (e.g., Neumark, Bank, and Van Nort 1996, and Riach and Rich 2002). In addition to conscious taste-based or statistical discrimination (Becker 1978), gender biases are automatically activated as soon as evaluators learn the sex of a person. Biases lead to unintentional and implicit discrimination that is not based on a rational assessment of the usefulness of sex in predicting future performance (e.g., Banaji and Greenwald 1995, Bertrand, Chugh, and Mullainathan 2005). For example, science faculty rated a male candidate who applied for a laboratory manager position as significantly more competent and hireable than an otherwise identical female candidate, and this differential evaluation was moderated by the faculty's pre-existing bias against women (Moss-Racusin et al. 2012).

Effective mechanisms to decrease the impact of such biases are blind evaluation procedures. For example, many major orchestras have musicians audition behind a curtain. These methods have proven to substantially decrease gender discrimination in the selection of musicians for orchestras (Goldin and Rouse 2000). Other attempts at overcoming gender biases include diversity training, which surprisingly seems to have had little impact (Dobbin, Kalev and Kelly 2007). Gender quotas on search and evaluation committees have had mixed results, given that stereotypes tend to affect both male and female evaluators (Bagues and Esteve-Volart 2010, Moss-Racusin et al. 2012). Quotas—e.g., for political bodies, corporate boards or senior management—are effective in increasing the fraction of members from underrepresented groups. And, with enough exposure to counter-stereotypical evidence, quotas have been shown to affect gender stereotypes (Beaman et al. 2009, Beaman et al. 2012, Dasgupta and Asgari 2004). However, in some cases, quotas had negative effects on performance (Matsa and Miller 2013).

This paper suggests a new intervention aimed at overcoming biased assessments: an "evaluation nudge," in which people are evaluated jointly rather than separately regarding their future performance.[1] We expect cognitive shortcuts, such as group stereotypes, to have less of an impact when multiple candidates are presented simultaneously and evaluated comparatively than when evaluators look at one person at a time.

Our work builds on earlier research in psychology suggesting that evaluation modes affect the quality of decisions by making evaluators switch from more intuitive decision-making in separate evaluation to more reasoned choices in joint evaluation. This often is attributed to the System 1/System 2 distinction where people are assumed to have two distinct modes of thinking that are variously activated

[1] A nudge is any aspect of choice design that is based on psychological insights into how our minds work and that alters people's behavior in a predictable way without restricting the freedom of individual choice. For nudges more generally, see Thaler and Sunstein (2008).

under certain conditions: the intuitive and automatic System 1 and the reflective and reasoned System 2 (Kahneman 2011, Stanovich and West 2000). Specifically, it has been suggested that the lack of comparison information available in separate evaluation leads people to invoke intuitively available internal referents (Kahneman and Miller 1986), focus on the attributes that can be most easily calibrated (Hsee et al 1999), and rely more on emotional desires than on reasoned analysis (Bazerman, Tenbrunsel, and Wade-Benzoni 1998) (for an overview, see Bazerman and Moore 2013).

Bazerman, Loewenstein, and White (1992) provided the original demonstration of preference reversals between joint and separate evaluation. In a two-party negotiation, they had study participants evaluate two possible negotiation outcomes—an even split of a smaller pie and a disadvantageous uneven split of a larger pie that still made both parties better off—either one at a time or jointly. When presented separately, most people preferred the equal split; when presented jointly, most preferred the money-maximizing alternative. Later studies on joint versus separate preference reversals found that brand name was more important than product features and price when people evaluated products separately rather than jointly (Nowlis and Simonson 1997); people were willing to pay more to protect animal species when evaluating separately and to invest in human health when evaluating the two causes jointly (Kahneman et al. 1993); and people were willing to pay more for a small portion of ice cream in a tiny, over-filled container when evaluating separately but for a large portion of ice cream in an under-filled huge container when evaluating the two serving options jointly (Hsee et al. 1999).

The focus of our study is to apply these insights to a new domain, the evaluation of people. In addition, we offer a new perspective on how to model a potential change in candidate assessments depending on the evaluation mode, a simple behavioral model of information processing. We assume that evaluators influenced by stereotypes start out by overweighting the importance of the characteristics of the group that the candidate belongs to. When evaluators receive more information on the candidate's individual past performance, they update their beliefs. By definition, in joint evaluation, more potentially counter-stereotypical data points are available than in separate evaluation, thus providing evaluators with more information to update their stereotypical beliefs. The difference in the amount of available information could lead evaluators to choose a lower-performing stereotypical person in separate evaluation but a higher-performing counter-stereotypical person in joint evaluation.

We employ laboratory experiments to examine whether evaluating candidates jointly rather than separately leads to individual performance playing a more important role than group stereotypes. In our experiments, we had subjects assume the role of either evaluators or candidates. Evaluators assessed the likely future performance of candidates either in separate or joint evaluation of their performance. Specifically, they were informed of candidates' past performance and their sex (plus a number of filler characteristics) and asked to decide whether given candidates were suitable for given jobs, either evaluating them separately or jointly, in one of two sex-typed tasks, a math or a verbal task.

Most studies that measure explicit gender attitudes find that females are believed to be worse at math and better at verbal tasks than males (Perie, Moran and Lutkus 2005, Price 2012). Implicit association tests (IATs) measuring people's implicit attitudes report math and verbal skills to be associated with maleness and femaleness respectively (Nosek, Banaji and Greenwald 2002, Plante, Theoret and Favreau 2009). The evidence on actual performance differences between the genders is mixed and varies by country and population, sometimes finding support for a gender gap in the expected direction, sometimes finding no gender differences, and, in recent years, finding a reversal of the gender gap in mathematics in several countries (Guiso et al. 2008). Despite the mixed evidence, we expect gendered beliefs to be sticky and these tasks to create stereotype-advantaged and stereotype-disadvantaged groups, with men being stereotype-advantaged in the math task and women in the verbal task. In addition, we expect that members of these groups will be affected by these biases even when at the individual level, conditional on the information available on the individual, gender is not informative and should not impact the evaluation.

We made a number of design choices to be able to test the impact of the evaluation mode as cleanly as possible. First, we decided to focus on cases where evaluators were faced with a dilemma, with stereotypes favoring one candidate and performance information favoring another candidate. Thus, in joint evaluation, we always studied mixed gender pairs with different performance scores. In addition, we restricted ourselves to performance levels close to the average performance level in the group with relatively small performance differences across candidates. Finally, performance was easily measurable and this information was available in our context. Clearly, in an organizational context, additional complexities come into play.[2]

In our experiment, gender stereotypes had a strong and significant impact on evaluators' candidate assessments even though gender was not correlated with task performance. Evaluators were significantly more likely to focus on group stereotypes in separate than in joint evaluation and to focus on the past performance of the individual in joint than in separate evaluation. This gender gap in separate and performance gap in joint evaluation makes joint evaluation the profit-maximizing evaluation procedure.

Our experimental findings have implications for the design of hiring and promotions procedures. Both joint and separate evaluation procedures are common for such decisions. Based on a recent survey of senior business executives in US companies with more than 1,000 employees (Penn, Schoen and

[2] In organizations, evaluators might well be confronted with various candidates of the same sex or the same performance levels where the basis of their decision is impossible to pin down. Also, performance likely is harder to measure in the field than in the lab and a candidate's gender may be more or less salient. And we expect (and hope) performance to trump gender bias in more extreme situations where large performance differences exist.

Berland 2012), in 30 percent of all promotion decisions, only one candidate was considered. For hiring decisions, we rely on the literature on sequential vs. non-sequential searches, building on Stigler (1961). In sequential search, a firm screens each applicant upon arrival and offers the job to the first applicant whose productivity exceeds a certain threshold. In non-sequential searches, a firm pools a number of applicants, screens them and offers the job to the best person in the pool. The former search strategy resembles separate and the latter joint evaluation. Recruitment strategies vary with firm and job characteristics but overall, about half of the hiring procedures studied seem to correspond to sequential (separate evaluation) and half to non-sequential (joint evaluation) searches (van Ommeren and Russo 2009, and Oyer and Schaefer 2010). Unfortunately, neither the promotion nor the hiring literature has examined the gender impacts of the different hiring and promotion strategies.

Organizations may seek to overcome biases in hiring, job assignment and promotion because they want to maximize economic returns. They may worry about the inaccuracy of stereotypes in predicting future productivity, or they may hold gender equality as a goal in itself. Introducing joint rather than separate evaluation procedures may enable them to nudge evaluators toward taking individual performance information into account rather than gender stereotypes.

Our paper is organized as follows: Part 2 offers a conceptual framework, Part 3 describes the experimental design, Part 4 reports our experimental results and Part 5 concludes.

2. Conceptual Framework

Our evaluation nudge builds on the observation in behavioral decision research that people make more reasoned decisions in joint than in separate evaluation modes. Various potential psychological mechanisms have been proposed to account for this phenomenon (summarized by Bazerman and Moore 2013). We suggest that in addition to providing new reference points, making goods and people more easily evaluable or focusing evaluators' attention on what they should be doing instead of what they want to do, joint evaluation also provides evaluators with more data than separate evaluation. Thus, evaluators have more information available to update their (possibly biased) beliefs in joint than in separate evaluation. A Bayesian-like model of information processing may illustrate this. We assume that evaluators are informed of candidate(s)' individual past performance in a given task, their sex and the average past performance of the pool of candidates. Based on the information received, evaluators have to decide whether to "hire" the candidate (s) presented to them for future performance in the task or go back to the pool and be allocated a candidate at random. Evaluators are paid based on their candidates' future performance and thus, have an incentive to select who they believe to be most productive, based on the candidate's future expected performance. Evaluators either evaluate one candidate at a time (separate evaluation) or two candidates at a time (joint evaluation). In both conditions, evaluators hire one candidate only, either by selecting one of the candidates presented or by going back to the pool and being allocated a random candidate.

A "behavioral" Bayesian model of information processing that allows evaluators to take irrelevant group characteristics into account, can explain an increase in the likelihood that evaluators choose higher-performing candidates in joint as compared to separate evaluation. Evaluating more than one person at a time implies having more data points available on the candidate's relative performance to update prior biased beliefs. If the new information is counter-stereotypical, it could theoretically shift beliefs enough for the evaluator to choose a counter-stereotypical person for a given job in joint but not in separate evaluation. We provide the formal proof for this result in the online appendix and derive the following empirically testable hypothesis:

Gender gap in separate and performance gap in joint evaluation: Candidates are more likely to be selected for future performance based on their gender when evaluated separately and based on their past performance when evaluated jointly.

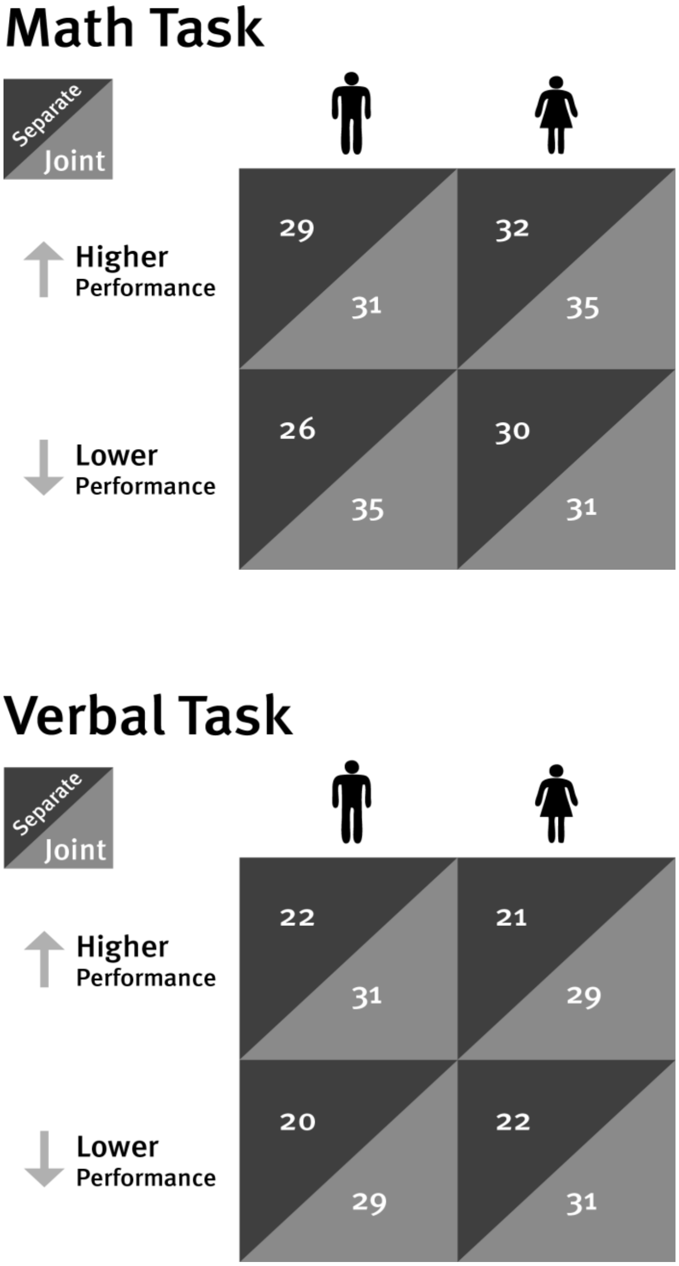
To test whether choices of evaluators are indeed based on biased expectations of future performance rather than on a preference for men for stereotypically male and women for stereotypically female tasks (taste-based discrimination), we present current-round performance information and ask evaluators whether they want to be paid based on the presented candidate's performance in the current round or be allocated a random person from the pool. We focus on the condition in which we expect most discrimination to take place, separate evaluation, and the candidates we expect to be most discriminated against, namely the higher-performing candidates from stereotype-disadvantaged groups. If there is no taste-based discrimination, they should be equally likely to be chosen as the higher-performing candidates from the stereotype-advantaged group. We do not expect taste-based discrimination in our context.

3. Experimental Design

Our experiment was conducted in the Harvard Decision Science Laboratory. We had 180 subjects participate as "candidates" in a math or a verbal task. 328 subjects assumed the role of "evaluators," selecting one of the candidates for future performance in the task. All were American college students. We employed equal numbers of male and female evaluators. All our participants were identified by code numbers and remained anonymous to each other and to the experimenter.

We employed a 2x2x2x2 experimental main (between-subject) design in which the key treatment condition of interest was the evaluation mode, joint or separate. In addition, we varied the individual candidates' past performance levels and their gender. Finally, candidates participated in either a math or a verbal task, with men being the stereotype-advantaged group in the math task and women the stereotype-advantaged group in the verbal task. Figure 1 provides an overview of our design and indicates the number of evaluators in each cell.

Figure 1: Main experimental design: Number of evaluators per treatment cell



The experiment was programmed and conducted in two stages, using Z-Tree software (Fischbacher 2007, sample instructions are included in Appendix B). In stage 1, candidates participated in either a verbal or a math task and were paid based on their performance. In stage 2, evaluators were informed of candidates' past performance and their gender and then were asked to select a candidate for future performance in the same task.

In stage 1, the candidates participating in the verbal task engaged in a word-search puzzle. They were given a list of 20 words and were instructed to mark as many of the words as they could find in three minutes in a matrix containing letters (Bohnet and Saidi 2012). Most letters appeared in random order, but some formed words, and participants could search horizontally, vertically, and diagonally. On average, the 100 candidates participating in this task found 10 words (SD=3.81) in the first round and 12 words (SD=4.56) in the second round.

The math task involved correctly adding as many sets of five two-digit numbers as possible (Niederle and Vesterlund 2007, Niederle et al. 2013). On average, the 80 candidates who participated in this task solved 10 problems correctly (SD=3.09) in the first round and 10 problems (SD=3.35) in the second round. After completing their task, participants filled out a short demographic questionnaire (most importantly for us, indicating their gender). Candidates then were paid based on their performance and were not informed of Stage 2 of the experiment.

In stage 2, evaluators in both the verbal and the math tasks were asked to choose a candidate, knowing that they would be paid based on that candidate's Round 2-performance. They could either choose the candidate presented to them, or go back to the pool and accept a randomly selected person. They had the candidate's Round 1-performance and his or her gender available as a basis for their decision, and were informed that on average, evaluators in the pool had provided 10 correct answers (as was the case for both tasks). The candidates presented to the evaluators were either average or slightly below-average performers, having provided either 10 or 9 correct answers in the first round. We chose first-round performance scores at and below the mean performance level of the pool to make sure that our results were not driven exclusively by evaluators' risk (or loss) aversion.

In the separate-evaluation condition, evaluators were presented with either a male or a female candidate who was either an average- or below-average performer. We randomly selected four candidates of the required gender-performance combinations from our pool: Male-10, Female-10, Male-9, and Female-9, with identical filler characteristics. In the joint-evaluation condition, evaluators were presented with a male and a female candidate simultaneously, drawing from the same candidates used in the separate-evaluation condition. The candidates differed on both gender and past performance, leading to two possible combinations: Male-10/Female-9 and Male-9/Female-10. We did not include same-sex pairs to create a dilemma for evaluators where the stereotype pointed them in one and the individual performance in the other direction. For example, in the math task, we expect that in a Male-10/Male-9 pair, Male-10 would clearly dominate Male-9 while in a Male-9/Female-10 pair, evaluators would be torn. We also did not include mixed-sex/same-performance level pairs although arguably, say a Male-

10/Female-10 pair would have provided us with interesting information on the power of stereotypes when performance was not an issue. As all previous joint-separate studies and our model of information processing assume a conflict between the attributes, we did not include this condition. We acknowledge, however, that given that gender is the only variable that differs in this condition, with everything else being identical, gender is likely more salient than in separate evaluation and even than in our existing joint evaluation condition with mixed-sex and mixed-performance level pairs. Gender salience may either lead to an increase in stereotypical choices or to reactance and a decrease in stereotypical choices, thus, truly making this an empirical question beyond the scope of this paper.

To make the gender-attribute less salient, without creating any additional demographic variation, we took advantage of the demographic similarity of our candidates and provided evaluators with truthful filler information on their candidates' characteristics. In addition to learning a person's sex and past performance, evaluators were also informed that he or she was a student, American, and from the greater Boston area. Despite these efforts, we cannot exclude the possibility that a person's sex was more salient than in an evaluation context outside of the lab. At the same time, presenting rather precise performance indicators compared to most performance measures in the field and using fewer possible criteria than typical in practice provides a conservative test for the impact of gender stereotypes. Heuristics likely play a more important role in situations where performance cannot be objectively measured (Stainback et al. 2010) and where multiple criteria for evaluation are available as they allow evaluators to focus on specific criteria only to justify their biased decisions (Norton et al. 2004). In our design, it seems difficult to justify neglecting individual performance information collected for the same task in the previous round.

After the experiment was completed, evaluators participated in an incentivized risk-attitude assessment task (Holt and Laury 2002) and completed a short questionnaire that collected basic demographic information. Evaluators were paid based on their decision, i.e. either the chosen candidate's second-round performance or the randomly allocated candidate's second-round performance. They received $1 for every correct answer that the candidate had provided. Evaluator earnings varied between $17.80 and $34.75, which included a $10 show-up fee, experimental earnings, and the payment for the risk-attitude assessment task.

In addition to our main experiment, we ran a small control experiment in which we informed evaluators about candidates' present rather than past performance with an additional 110 subjects. Specifically, evaluators were informed of a candidate's second-round performance and then had to decide whether or not to select this candidate and be paid based on the candidate's performance in the second round or go back to the pool and accept a randomly allocated candidate. This experiment was designed to distinguish belief-based from taste-based discrimination. While in our main experiment, both motives could lead to gender-biased decisions, in the control experiment, only taste-based discrimination was possible. We replicated the separate-evaluation conditions, in which we expected gender to be most prevalent, and used average performers, the group we were most concerned about being discriminated against. For separate evaluation, 23 evaluators participated in the male math condition, 27 in the female math condition, 33 in the male verbal condition, and 27 in the female verbal condition. Other than giving evaluators information about candidates' present rather than past performance, the control study was run identically to our main experiment.

After participants had made their decisions, learned their outcomes, and given us their demographic information, they presented their code number and were given a sealed envelope containing their earnings.

4. Results

We first present candidates' performance in the two tasks, then examine what role gender and individual performance played in the two different evaluation modes, and finally examine alternative explanations.

4. 1. Candidates' performance

We first examine whether or not having gender-stereotypical beliefs was accurate in our context. There were no significant gender differences in performance on either task, although directionally, the small differences we did observe accord with stereotypical assumptions.[3] Thus, ex-post, statistical discrimination was unwarranted. In addition, information on group characteristics in our experiment was always combined with individual performance information. Conditional on this performance information, stereotypes were completely irrelevant for predicting future performance.

Table I reports the regression results of individual past (first-round) performance and gender on future (second-round) performance for both tasks. Columns 1 and 3 show that first-round performance was highly correlated with second-round performance, while the gender of the candidate was irrelevant for second-round performance in both tasks. In Columns 2 and 4, we control for potential gender differences in the relationship between first- and second-round performance and include an interaction term between the two variables. For example, the strong first-round performance of a candidate from a stereotype-disadvantaged group could be due to luck and thus be less predictive of future performance than the same performance by a member of a stereotype-advantaged group (and vice versa for low

[3] In the math task, performance levels were as follows: Round 1, men: Mean=10.63, SD=3.41; women: Mean=10.33, SD= 2.78; p=0.67. Round 2, men: Mean=10.63, SD=3.57; women: Mean=9.95, SD =3.13; p=0.37. In the verbal task, performance levels were as follows: Round 1, men: Mean=9.82, SD=4.05; women: Mean=10.98, SD=3.49; p=0.13. Round 2, men: Mean=12.46, SD=4.27; women: Mean=12.08, SD=4.87; p=0.68. There are no significant differences in variance across the genders, and the distributions in performance are not significantly different according to Kolmogorov-Smirnov tests.

performance). Columns 2 and 4 suggest that first-round performance was equally predictive of future performance for both genders.[4]

Table 1- The Effect of Past Performance and Stereotypes on Future (Second-Round) Performance

	Math Task		Verbal Task	
	1	2	3	4
First-round Performance	0.849***	0.797***	0.708***	0.813**
	(0.08)	(0.15)	(0.10)	(0.15)
Male Candidate	0.420	-0.481	1.201	3.118
	(0.46)	(1.91)	(0.77)	(2.42)
First-round Performance x Male		0.086		-0.183
		(0.17)		(0.20)
Constant	1.189	1.723	4.311*	3.158
	(0.97)	(1.66)	(1.35)	(1.95)
N	80	80	100	100
R^2	0.6217	0.6232	0.3423	0.3478

Notes: Each specification is an OLS regression. Robust standard errors in brackets. The dependent variable is the number of correctly added sequences in round 2 for the math task and the number of words found in round 2 for the word task. *** Significant at the 1 percent level. ** Significant at the 5 percent level.* Significant at the 10 percent level.

4.2. Evaluators' choices

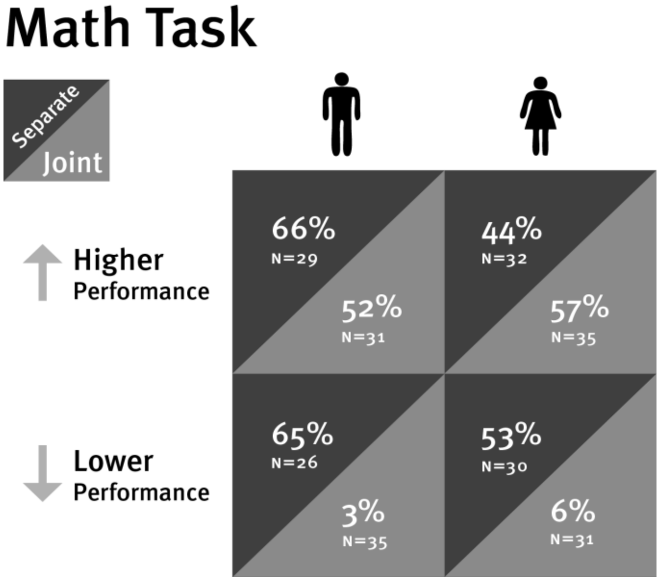
We start by aggregating across both evaluation modes and both performance levels. In the math task (N=183), the likelihood that the stereotype-disadvantaged candidate, i.e., the woman, was chosen was 0.41, and the likelihood that the stereotype-advantaged man was chosen was 0.44. In the verbal task (N=145), the likelihood that the stereotype-disadvantaged man was chosen across conditions was 0.38, while the likelihood that the stereotype-advantaged woman was chosen across conditions was 0.48. Thus, evaluators had a slight preference for men in math tasks and for women in verbal tasks, but these differences are not significant. The sex of the evaluator did not matter in the verbal task but played a more important role in the math task, with female evaluators more likely to choose a given candidate than male evaluators (also confirmed in the regression analysis of Table 3).[5]

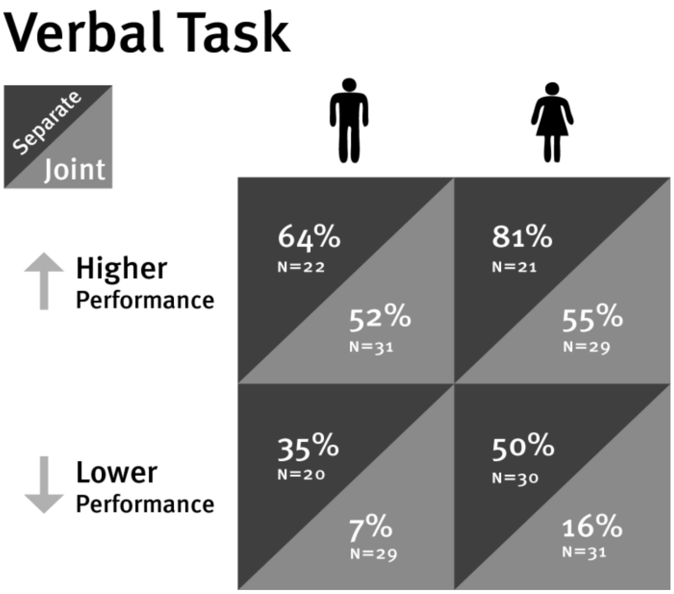
[4] In addition to controlling for the gender-specific randomness of performance across rounds, we also examined the possibility of gender-specific learning across rounds. On average and across both genders, little learning between rounds took place in the math task, while candidates in the verbal task performed significantly better in the second than in the first round, with men finding 2.64 and women 1.1 words more on average in the second than in the first round. However, the gender difference in learning was not significant, including in GLS regressions on performance in both rounds. Similar to the above results, average performance across both rounds was similarly correlated with the first-round performance of men and women in both tasks.

[5] In the math task, the likelihood that a male candidate was chosen by male evaluators was 37% and by female evaluators 50% $(X^2$ (1) = 2.12, p = .15) The likelihood that a female candidate was chosen by male evaluators was 26% and by female evaluators 51% $(X^2$ (1) = 7.57, p < .0.05). In the verbal task, the likelihood that the male candidate was chosen by male evaluators was 39% and by female evaluators 38% $(X^2$ (1) = 0.03, p = .87). The likelihood that a female candidate was chosen by male evaluators was 38% and by female evaluators 39% $(X^2$ (1) = 2.38, p = .0.12).

Looking at the two evaluation modes separately, we find that these differences were entirely driven by the stereotype-advantaged group being preferred in separate evaluation. Figure 2 shows our results for each evaluation mode, task, gender and performance level. In separate evaluation, the gender gaps in the likelihood of being selected are apparent, with the stereotype-advantaged group being favored in both the math and the verbal tasks. In joint evaluation, a performance gap emerged, with the higher-performing candidates being more likely to be selected than lower performers. Performance does not seem to matter in separate evaluation in the math task (but, in addition to gender, is relevant in the verbal task), and gender does not seem to matter in either task in joint evaluation.

Figure 2: Percentage of Candidates Selected in Separate and Joint Evaluation





Aggregating across both tasks, the following gender and performance gaps can be observed, supporting our hypothesis: Across both tasks and when evaluated separately (N=202), the likelihood that a candidate from the stereotype-advantaged group was chosen was 0.65, and the likelihood that someone from the stereotype-disadvantaged group was chosen was 0.49 (X^2 (1) = 5.45, $p < .05$). In joint evaluation (N=126), stereotypes did not matter at all: 32 percent of the evaluators chose a candidate from the advantaged group and 30 percent chose a candidate from the disadvantaged group. (The remainder of the evaluators, 38 percent, decided to go back to the pool.)[6]

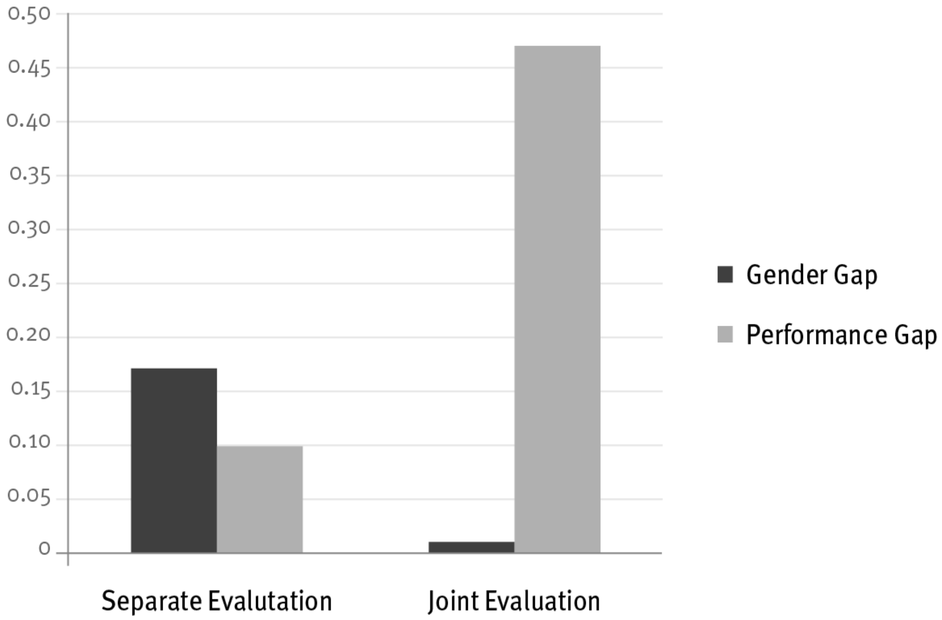
Higher-performing candidates were more likely to be chosen in joint but not in separate evaluation. Across both tasks and when evaluated jointly, the likelihood that a higher-performing candidate was chosen was 0.54, and the likelihood that a lower-performing candidate was chosen was 0.08 (X^2 (1) = 43.13, $p < .01$). In separate evaluation, performance differences hardly mattered: 62

[6] Generally, the likelihood that a given candidate was chosen was higher in separate than in joint evaluation. We attribute this to the number of options available in separate versus joint evaluation. If evaluators had chosen randomly, a given candidate would have been chosen by 50 percent of the evaluators in separate evaluation and by only 33 percent in joint evaluation. Thus, compared to random selection, the stereotype-advantaged candidates were significantly more likely to be chosen than what a random process would have predicted in separate (X^2 (1)=9.18, p<.01) but not in joint evaluation (X^2 (1)=.16, p=.69). The likelihood that stereotype-disadvantaged candidates were chosen did not differ from chance in either mechanism (for separate: X^2 (1)=.04, p=.8445; for joint: X^2 (1)=0.59, p=.4424).

percent of the evaluators chose a higher-performing candidate and 52 percent chose a lower-performing candidate (X^2 (1) =0.37, p =0.58).

Figure 3 shows the gender and performance gaps graphically. In separate evaluation, evaluators were 16 percentage points more likely to choose a candidate from the stereotype-advantaged rather than from the stereotype-disadvantaged group ($p < .05$), and in joint evaluation evaluators were 46 percentage points more likely to choose the higher- rather than the lower-performing candidate ($p < .01$). The gender gap completely disappears in joint evaluation.

Figure 3: Gender and Performance Gaps in Separate and Joint Evaluation Across Both Tasks



A regression analysis in Table 3 controlling for the relevant covariates confirms these insights. Gender only affected decisions in separate evaluation (Column 1), and performance only affected decisions in joint evaluation (Column 2). Members of the stereotype-advantaged group were significantly more likely to be chosen in the separate evaluation mode but not in the joint evaluation mode. In contrast, higher-performing candidates were only favored in joint but not in separate evaluation. Columns 3 and 4 include controls for the risk attitudes and the gender of the evaluator. Male and more risk- tolerant evaluators were less likely than female and more risk averse evaluators to select a given candidate than to go for the random option. Both of these results accord with intuition.

Table 3 -The Effect of Past Performance and Stereotypes on Candidate Selection, Marginal Effects at Mean

	(1) Separate	(2) Joint	(3) Separate plus controls	(4) Joint plus controls
First-round Performance	0.099	0.462**	0.117	0.472***
	(0.07)	(0.06)	(0.07)	(0.06)
Stereotype-Advantage	0.165**	0.009	0.164**	0.008
	(0.07)	(0.07)	(0.07)	(0.07)
Math	-0.009	-0.043	0.018	-0.040
	(0.07)	(0.05)	(0.07)	(0.05)
Risk Tolerance			-0.059***	-0.002
			(0.02)	(0.01)
Male evaluator			-0.099	-0.199***
			(0.07)	(0.05)
N	202	252	202	252
Pseudo R^2	0.0271	0.2201	0.0664	0.2579

Notes: Each specification is a Probit regression, marginal effects reported in percentage points. The dependent variable in the separate treatment is the selection of a given candidate. In the joint treatment we score two outcomes for each individual: namely, whether the employer selected the higher (1) or the lower (2) performer: This implies a total of 252 outcomes. Robust standard errors are in brackets and adjusted for clustering at the employer level. Risk tolerance is measured by the number of risky choices made in a lottery (identical to Holt and Laury (2002). *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

4.3. Alternative explanation: taste-based discrimination

We did not find any evidence for taste-based discrimination in our control experiment. Across the two tasks, the likelihood that a member of the stereotype-advantaged group was chosen was 0.46, and the likelihood that a member of the stereotype-disadvantaged group was chosen was 0.48. Specifically, instead of going back to the pool, in the math task (N=50), 35 percent of the evaluators chose the male candidate and 41 percent chose the female candidate; in the verbal task (N=60), 55 percent chose the male and 56 percent the female candidate. None of these differences are significant; women and men were just as likely to be chosen for both tasks.

5. Conclusions

This paper examines whether an “evaluation nudge,” namely evaluating candidates jointly rather than separately, can overcome gender-biased assessments of job candidates that favor men for male-typed tasks and women for female-typed tasks, even if gender is not predictive of future performance and more reliable individual performance measures are available. We employ a setting where there is a conflict between the individual performance information favoring one of the candidates and the group stereotype favoring the other candidate. Our results apply to these kinds of settings. We find that when evaluators are tasked with choosing a candidate for future performance in a math or a verbal task, a joint-evaluation mode helps them focus on individual performance, irrespective of candidates’ gender and evaluator bias:

evaluators were significantly more likely to choose the higher- rather than the lower-performing candidate in this mode. In contrast, in separate evaluation, evaluators were heavily influenced by a candidate's gender, even though gender was not predictive of future performance and individual past performance was: they were significantly more likely to choose men for the math task and women for the verbal task.

In our setting, discrimination was based on biased beliefs about future performance rather than taste. In a control treatment, we could exclude taste-based discrimination. Thus, while there might well be taste-based discrimination in organizations, our findings cannot speak to this question. Joint evaluation may affect choices by providing additional data that evaluators can use to update their stereotypical beliefs about a group to which a candidate belongs. By definition, an evaluator has more data points available in joint than in separate evaluation. If these data points provide counter-stereotypical information, they may shift an evaluator's beliefs about the group enough to make him or her choose counter-stereotypically.

Our work is in line with extensive work in behavioral decision making suggesting that people may evaluate products differently in joint than in separate evaluation. This research attributed differences in decision outcomes to a switch in judgment modes from a more intuitive mode based on heuristics in separate evaluation to a more reasoned mode when comparing alternatives in joint evaluation (Bazerman and Moore 2008, Paharia et al. 2009, and Gino et al. 2011).

Our findings have implications for organizations that want to decrease the likelihood that hiring, promotion, and job-assignment decisions will be based on irrelevant criteria triggered by stereotypes. Joint evaluation is common for many hiring decisions but rare for job assignments and for promotion decisions. Organizations concerned about discrimination in this later phase might want to review how, for example, career-relevant jobs are assigned or how promotion decisions are made. According to the Corporate Gender Gap Report 2010 (Zahidi and Ibarra 2010), in most countries, fewer than 10 percent of career-relevant jobs are held by women. In many academic fields, including economics, controlling for performance, women are less likely to be granted tenure than men (Ginther and Kahn 2004, 2009).

Organizations can move from separate-evaluation to joint-evaluation procedures to promote more accurate decision-making and maximize performance. In addition to being a profit-maximizing decision procedure, joint evaluation is also a fair mechanism, as it encourages judgments based on people's performance rather than their demographic characteristics. Companies concerned about discrimination might choose to review how job candidates are evaluated, how jobs are assigned and promotion decisions made. Our work suggests that organizations can nudge evaluators toward taking individual performance information rather than gender stereotypes into account.

Acknowledgements

We thank Pinar Dogan, Chris Muris, Farzad Saidi, Katie Shonk, Richard Zeckhauser, and the participants of seminars at Harvard University, Tilburg University, the University of California Berkeley, the MOVE conference in Barcelona, and the MBEES conference in Maastricht for many helpful comments, and Sara Steinmetz for her research assistance. Financial support from the Women and Public Policy Program and the Women's Economic Opportunity Initiative of ExxonMobil are gratefully acknowledged.

References

Bagues M, Esteve-Volart B (2010) Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment. *Review of Economic Studies* 77(4): 1301-28.

Banaji MR, Greenwald AG (1995) Implicit Gender Stereotyping in Judgments of Fame. *Journal of Personality and Social Psychology* 68(2): 181-98.

Bazerman MH, Loewenstein GF, White SB (1992) Reversals of Preference in Allocation Decisions: Judging an Alternative Versus Choosing among Alternatives. *Administrative Science Quarterly* 37(2): 220-40.

Bazerman MH, Moore DH (2013) *Judgment in Managerial Decision Making.* Hoboken, NJ: John Wiley & Sons, 8th edition.

Bazerman MH, Tenbrunse AE, & Wade-Benzoni K (1998) Negotiating with yourself and losing: Making decisions with competing internal preferences. *Academy of Management Review*, 23(2), 225-241.

Beaman L, Chattopadhyay R, Duflo E, Pande R, Topalova P (2009) Powerful Women: Female Leadership and Gender Bias. *Quarterly Journal of Economics*, 124 (4): 1497–1540.

Beaman L, Duflo E, Pande R, Topalova P (2012) Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India. *Science*, 335(6068): 582-586

Becker GS **(**1978) *The Economic Approach to Human Behavior.* Chicago, IL: University of Chicago Press.

Bertrand M, Chugh D, Mullainathan S (2005) Implicit Discrimination. *American Economic Review* 95(2): 94-98.

Bohnet, I, Saidi, F (2012). Informational Differences and Performance: Experimental Evidence. Under Review.

Dasgupta, N, Asgari, S (2004) Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Psychology* 40:642-658.

Dobbin F, Kalev A, Kelly B (2007) Diversity Management in Corporate America. *Contexts* 6(4):21-28.

Fischbacher U (2007) Z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics* 10: 171-78.

Gino F, Schweitzer ME, Mead NL, Ariely D (2011) Unable to Resist Temptation: How Self-Control Depletion Promotes Unethical Behavior. *Organizational Behavior and Human Decision Processes* 115(2): 191-203.

Ginther DK, Kahn S (2004) Women in Economics: Moving Up or Falling Off the Academic Career Ladder? *Journal of Economic Perspectives*, 18: 193-214.

Ginther DK, Kahn S (2009) Does Science Promote Women? Evidence from Academia 1973-2001. In: Science and Engineering Careers in the United States: An Analysis of Markets and Employment, eds Freeman, R-B and Goroff D-G, (Chicago: University of Chicago Press).

Goldin C, Rouse C (2000) Orchestrating Impartiality: The Impact of Blind Auditions on Female Musicians. *American Economic Review* 90(4): 715-41.

Guiso L, Monte F, Sapienza P, Zingales L (2008) Diversity: Culture, Gender, and Math. *Science* 320 (5880): 1164-5.

Holt, CA, Laury SK (2002) Risk Aversion and Incentive Effects. *American Economic Review* 92: 1644-1655.

Hsee, CK, Blount S, Loewenstein GF, Bazerman MH (1999) Preference Reversals Between Joint and Separate Evaluations of Options: A Review and Theoretical Analysis. *Psychological Bulletin* 125(5): 576-90.

Kahneman D, Ritov I, Jacowitz KE, Grant P (1993) Stated Willingness to Pay for Public Goods: A Psychological Perspective. *Psychological Science* 4(5): 310-15.

Kahneman D, Miller D (1986) Norm theory: Comparing reality to its alternatives. *Psychological* review 93(2): 136-153.

Kahneman, D (2011) *Decisions, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Matsa DA, Miller AR (2013) A Female Style in Corporate Leadership? Evidence from Quotas. *American Economic Journal: Applied Economics* 5(3): 136-69.

Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J (2012) Science faculty's subtle gender biases favor male students. *PNAS* 109 (41): 16474-16479.

Neumark D, Bank RJ, Van Nort KD (1996) Sex Discrimination in Restaurant Hiring: an Audit Study. *The Quarterly Journal of Economics* 113(3): 915-41.

Niederle M, Vesterlund L (2007) Do Women Shy Away from Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics* 122(3): 1067-101.

Niederle M, Segal C, Vesterlund L (2013) How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness. *Management Science*, 2013, 59 (1) 1-16.

Norton MI, Vandello JA, Darley JM (2004) Casuistry and social category bias. *Journal of Personality and Social Psychology* 87(6): 817-31.

Nosek B, Banaji M, Greenwald AG (2002) Math= Male, Me= Female, Therefore Math (not equal to) Me. *Journal of Personality and Social Psychology* 83(1): 44-59.

Nowlis SM, Simonson I (1997) Attribute-Task Compatibility as a Determinant of Consumer Preference Reversals. *Journal of Marketing Research* 34(2): 205-18.

Oyer, P, Schaefer S (2011) Personnel Economics: Hiring and Incentives, in Handbook of Labor Economics, vol. 4B, eds Card, D. and Ashenfelter, O. (Elsevier) pp. 1769-1823.

Paharia N, Kassam KS, Greene, JD, Bazerman MH. **(**2009) Dirty Work, Clean Hands: The Moral Psychology of Indirect Agency. *Organizational Behavior and Human Decision Processes* 109(2): 134-41.

Penn, Schoen & Berland Associates, inc. (2012) *The Capstone Project.* http://msb.georgetown.edu/document/1242764748554/Favoritism+Research+-+McDonough+School+of+Business.pdf. (Accessed July 16, 2012).

Perie M, Moran R, Lutkus AD (2005) NAEP 2004 Trends in Academic Progress: Three Decades of Student Performance in Reading and Mathematics. Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics.

Plante I, Theoret M, Favreau OE **(**2009) Student Gender Stereotypes: Contrasting the Perceived Maleness and Femaleness of Mathematics and Language. *Educational Psychology* 29(4): 385-405.

Price, Curtis R. 2012. Gender, Competition and Managerial Decisions. *Management Science* 58(1): 114-122

Riach PA, Rich J (2002) Field Experiments of Discrimination in the Market Place. *The Economic Journal* 112: 480-518.

Stainback K, Tomaskovic-Devey D, Skaggs S (2010) Organizational approaches to inequality: Inertia, relative power, and environments. *Annual Review of Sociology* 36: 225-47.

Stanovich KE, West RF (2000) Individual Differences in Reasoning: Implications for the Rationality Debate? *Behavioral and Brain Sciences* 23: 645-65

Stigler GJ (1961) The Economics of Information. *The Journal of Political Economy* 69(3): 213-225.

Thaler RH, Sunstein CR (2008) *Nudge: Improving Decisions about Health, Wealth, and Happiness.* New Haven, CT: Yale University Press.

Van Ommeren J, Russo G (2009). Firm Recruitment Behaviour: Sequential or Non-Sequential Search? IZA Discussion Paper, 4008: 1-36.

Zahidi, S, Ibarra H (2010) *The Corporate Gender Gap Report 2010*. World Economic Forum, Geneva,

Switzerland

Appendix A: Information-based Model

We show that under certain conditions, evaluators will choose the higher-performing candidate in joint but not in separate evaluation. In joint evaluation, they may observe more counter-stereotypical candidates (e.g., women performing strongly and men poorly in math) than in separate evaluation.

The evaluator is either in a joint or separate evaluation condition and has to choose a candidate for future performance. The evaluator is informed of selected candidate(s)' past performance and gender (type), as well as the total size and average performance of the candidate pool. Each candidate *i* has a type $g \in (m,f)$, and type *m* is believed to have higher expected performance. We define x_{ig} as the observed past performance of candidate *i* with type *g*, and let y_{ig} denote this candidate's (unknown) future performance. The total size of the candidate pool is known to be $N=2j$ with $x_{im} \in (x_1...x_j)$, and $x_{if} \in (x_{j+1}...x_{2j})$, reflecting a 50/50 distribution of types. The average previous performance across types, i.e $\bar{x} = \frac{1}{N}\sum_{i=1}^{2J} x_i = \frac{1}{2}\sum_{g=f}^{m} \bar{x}_g$, is known. We denote $\bar{x}_g$ as the *(unknown)* average previous performance by gender. Evaluators know $x_g \sim N(\bar{x}_g, \sigma^2)$, with known variance ($\sigma^2$).

In joint evaluation, evaluators observe the performance of two candidates (i.e., they observe $\mathbf{x} = (x_{if}\ x_{im})$). Evaluators can choose between these candidates and the random option with known expected performance of $\bar{x}$. In separate evaluation, evaluators observe $\mathbf{x} = (x_{if})$ or $\mathbf{x} = (x_{im})$, and they can choose between the candidate and the random option with (known) expected performance of $\bar{x}$.

We denote evaluators' prior expectation of $\bar{y}_g$ as $\mu_{g.}$ Evaluators have no taste for discrimination but hold stereotypical beliefs. Specifically, $\mu_m > \mu_{f,}$ and because of the 50/50 distribution and known $\bar{x}$, we can write $\mu_m = \bar{x} + h$ and $\mu_f = \bar{x} - h$with $h > 0$ (assuming no learning of the task for candidates over time). Evaluators' prior beliefs about each type's average future productivity look as follows: $\theta_g \sim N(\mu_g, v^2)$, assuming equal, positive, and known variances across the genders. Because of symmetry, we have $(|\theta_g - \bar{x}|) \sim N(\mathrm{h}, \nu^2)$.

.

We assume that evaluators are risk-neutral expected-value maximizers and that the expected future performance (y_{ig}) is a linear combination of the observed previous performance of the candidate and the (updated) belief about the candidate's performance based on gender; i.e., $E(y_{ig} \mid \mathbf{x}) = \alpha x_{ig} + (1-\alpha)\mu_g(.|\mathbf{x})$ with $0 < \alpha < 1$.

After observing **x**, evaluators update beliefs on $\bar{y}_g$ according to Bayes rule, to the posterior distribution of $\mu(. \mid \boldsymbol{x})$ over $\theta_{g.}$ An evaluator in **separate** evaluation confronted with a **below-average male**

candidate (i.e. with $x_m = b$) will update beliefs about mean performance for males in the pool to:
$$\mu_m(.|\mathbf{x}) = (\bar{x} + (h \mid x_{i\ m})) = b + \frac{\sigma^2}{\sigma^2+\nu^2}(h + (\bar{x} - b))\,)$$

If faced with **an average female candidate** (i.e. with $x_f = \mu$), an evaluator will update beliefs about the mean performance of females in the pool to: $\mu_f(.|\mathbf{x}) = (\bar{x} - (h \mid x_{i\ f})) = \bar{x} - \frac{\sigma^2}{\sigma^2+\nu^2}h$

In **joint** evaluation, evaluators have two data points available; they use both the male and the female candidates' past performance to update their prior of h. In the counter-stereotypical situation where an evaluator is confronted with a lower-performing male ($x_m = b$) and a higher-performing female candidate ($x_f = \mu$), this results in updated beliefs of mean performance for **males** in the pool of:
$$\mu_m(.|\mathbf{x}) = (\bar{x} + (h \mid x_{i\ m} x_{i\ f})) = \frac{(\bar{x}+h)\sigma^2 + \nu^2(b+\bar{x})}{\sigma^2 + 2\nu^2}$$

It results in an updated mean for **females** in the pool of:
$$\mu_f(.|\mathbf{x}) = (\bar{x} - (h \mid x_{i\ m} x_{i\ f})) = \frac{(\bar{x}-h)\sigma^2 + \nu^2((2\bar{x}-b)+\bar{x})}{\sigma^2 + 2\nu^2}$$

The evaluator will compare the updated expected future performance E(y_{ig} | $\boldsymbol{x}$) of the candidates with the expected value of the random option ($\bar{x}$) and choose the option with the highest expected performance. To have a situation where evaluators prefer a high-performing female in a joint evaluation setting over a low-performing male and the random option, but prefer the random option over the high-performing female in the separate evaluation setting, the following conditions need to simultaneously hold:

(1) The expected future performance of a higher-performing female candidate dominates the random option,

(2) The expected future performance of a higher-performing female candidate dominates the lower-performing male option in joint evaluation, and

(3) The expected future performance of a higher-performing female candidate is lower than the expected value of the random option in the separate treatment.

If $\frac{h\sigma^2}{\nu^2} < (\bar{x} - b)$ these conditions hold:

(1) $(1-\alpha) * \bar{x} + \alpha * \frac{(\bar{x}-h)\sigma^2 + \nu^2((2\bar{x}-b)+\bar{x})}{\sigma^2 + 2\nu^2} > \bar{x}$.

(2) $(1-\alpha) * \bar{x} + \alpha * \frac{(\bar{x}-h)\sigma^2 + \nu^2((2\bar{x}-b)+\bar{x})}{\sigma^2 + 2\nu^2} > (1-\alpha) * b + \alpha * \frac{(\bar{x}+h)\sigma^2 + \nu^2(b+\bar{x})}{\sigma^2 + 2\nu^2}$

(3) $\bar{x} > (1-\alpha) * \bar{x} + \alpha * (\bar{x} - \frac{h\sigma^2}{\sigma^2+\nu^2})$

Thus, whenever there is sufficient variance of the expected difference between male and female performance, there is enough counter-stereotypical evidence, and evaluators are not too biased, the joint-separate reversal can be observed. For the parameters used in the experiment, the condition reduces to $h\sigma^2 < \nu^2$.

To illustrate the model, we provide a numerical, simple example based on the specific parameters used in our experiment here: Imagine that on average, an equal number of male and female candidates solved 10 math problems correctly in the past. A (biased) evaluator having to choose someone for future performance in the math task has prior beliefs that a typical woman solves 9 math problems and a typical man 10 problems correctly. Thus, with the same number of men and women in the pool, our evaluators' expected value of the pool is 9.5.

Now, evaluators participate in the separate evaluation condition where they receive a counter-stereotypical signal, namely that the specific male candidate presented only solved 9 problems correctly. They continue to put weight on average group performance even though given information on an individual's past performance, this is completely uninformative. Evaluators adjust their priors for typical male performance downwards, maybe, to an expected value of 9.4. Given that their prior for female performance is 9 and that we have the same number of male and female candidates in the pool, the evaluators' expected value for the whole pool is now adjusted from 9.5 to 9.2. The evaluators still prefer the male candidate presented to them with an expected value of 9.4 to a random draw from the pool with an expected value of 9.2 and thus, choose the man. Similarly, when the counter-stereotypical signal is from a female candidate having solved 10 problems correctly, the evaluators also adjust their priors about the average performance of female candidates, maybe, to 9.6. Based on the same logic as above, this yields an expected value of 9.8 for the whole pool, leading the evaluators to prefer going back to the pool rather than choosing the woman (9.8>9.6).

When evaluators participate in joint evaluation, they receive two signals. Imagine that both are counter-stereotypical, namely that the male candidate performed worse (9 problems solved) and the female candidate better than expected (10 problems solved). This may shift priors enough for the evaluators to prefer the counter-stereotypical woman to the stereotypical man. Using the same logic as above, evaluators adjust their priors about expected male performance downwards to 9.4 and their priors about expected female performance upwards to 9.6, as before, yielding an expected value of the pool of 9.5. Now, the expected value of female candidates exceeds the expected value of going back to the pool and the evaluators choose the higher-performing woman (9.6>9.5).

Appendix B: Experimental Instructions

Instructions stage 1

All treatments were programmed in Z-tree. (Fischbacher 200)) We include as an example our instructions for the math task (inspired by Niederle and Vesterlund 2007), instructions for the word task were similar and are available upon request.

Welcome!

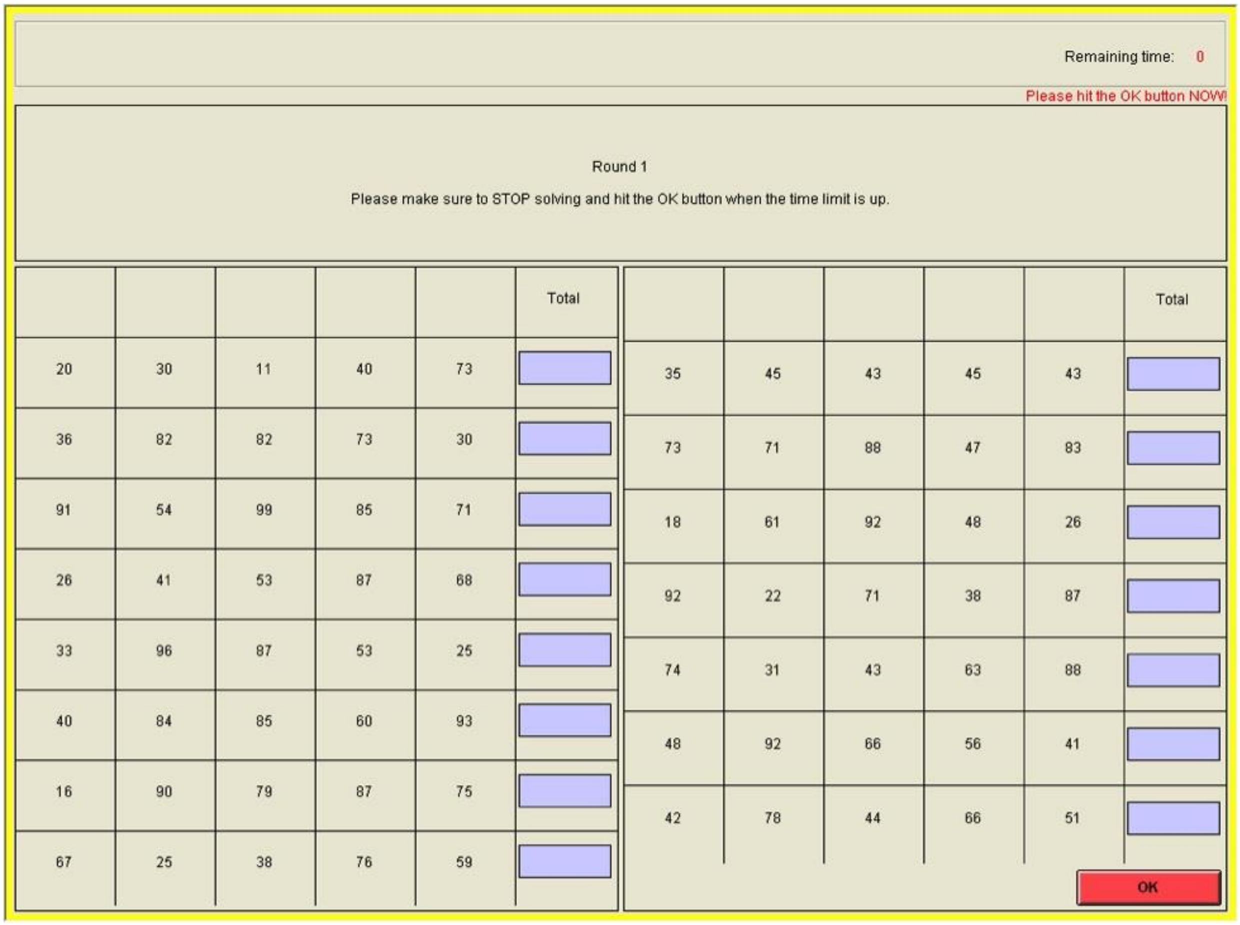
In this experiment you are asked to correctly solve as many math problems as possible. In each problem, you are asked to sum up five two-digit numbers.

For each correct answer you will receive 25 cents. There will be three rounds; each round consists of 15 problems. You have five minutes available for each round.

Before we begin with the experiment there will be a practice round where you can get used to the task.

At the end of the experiment, you will receive an overview of the number of correct answers and of your total payoff.

An example of this task is given in the figure below.



Remaining time: 0

Please hit the OK button NOW!

Round 1

Please make sure to STOP solving and hit the OK button when the time limit is up.

					Total
20	30	11	40	73	
36	82	82	73	30	
91	54	99	85	71	
26	41	53	87	68	
33	96	87	53	25	
40	84	85	60	93	
16	90	79	87	75	
67	25	38	76	59	

					Total
35	45	43	45	43	
73	71	88	47	83	
18	61	92	48	26	
92	22	71	38	87	
74	31	43	63	88	
48	92	66	56	41	
42	78	44	66	51	

After performing the task, participants filled out a questionnaire collecting demographic information.

Instructions Stage 2:

Below are the instructions for the joint treatment with the math task and a higher performing female candidate and a low performing male. We also report the instructions for the separate treatment with the math task with high performing male**.** Instructions for the other treatments were similar and are available upon request.

Joint Treatment Math Task

Welcome!

You are participating in a study in which you will earn some money. The amount will depend on a choice that you will have to make below. At the end of the study, your earnings (1 point = $ 1) will be added to a show-up fee, and you will be paid in cash.

Your Choice.--Another group of study participants has participated in two rounds of a task before this session. You will receive information on two of the participants, person A and person B and on how well they performed in Round 1. You will then have to decide whether you want to be paid according to the Round 2 performance of person A, person B or of a randomly selected person from the pool of participants.

Information on Task.---In a previous study, participants were shown rows of five two-digit numbers. Participants had to add up the numbers of each row. Participants were asked and incentivized to add up as many rows as possible as possible. They had five minutes available for each round of the task. While the task was otherwise identical, they saw different sequences containing different numbers in Rounds 1 and 2.

Their point score was calculated as follows: For every correctly added sequence they received one point. Sequences that were not correctly added received no points.

To have a better understanding of the task, please click on this button to see a sample task.

You will see the matrix for 30 seconds and not for 5 minutes.

(SAMPLE TASK)

Information on Average Round-1 Performance of all Study Participants

On average participants scored 10 points in Round 1.

Information on Persons.---You will be paid according to the Round 2-performance of one of the two study participants described below, Person A or Person B, or a study participant drawn at random from all the people who participated in the study. We had 40 male and 40 female students participate, recruited by the Harvard Decision Science Laboratory.

Person A	Person B
Student	Student
American	American
Female	Male
Caucasian	Caucasian
Performance indicator: In Round 1 the person scored 10 points in five minutes.	Performance indicator: In Round 1, the person scored 9 points in five minutes.

Procedure to Determine your Earnings.---Once you have decided whether you want to be paid based on the performance of person A, person B or a randomly selected person and have completed a short questionnaire, we will inform you of their point score and your payoffs.

If you chose to be paid according to the performance of one of the persons described above, you will receive $1 x that person’s point score for Round 2.

If you chose to be paid according to the performance of a random person, you will receive $1 x the random person’s point score for Round 2.

For example if your chosen person scores 2 points in round 2, you will receive $2.

If you have any questions, please press the help button now. Once we have addressed all questions, we will move to the main question of this study.

Main question: Do you want to be paid based on the Round 2-performance of one of the persons described above, or do you want to be paid based on the Round 2-performance of a person drawn at random from all the people who participated in the study? (Please check one box)

NOTE: THE AVERAGE SCORE IN ROUND 1 WAS 10 POINTS

Person A	Person B	Random Draw
Student	Student	
American	American	
Female	Male	
Caucasian	Caucasian	
Performance indicator: In Round 1 the person scored 10 points in five minutes.	Performance indicator: In Round 1, the person scored 9 points in five minutes.	

Note: after the main question of the experiment participants were notified of the score of the randomly selected candidate, the score of person A, and the score of person B.

Additional Decision.---Please choose Option A or Option B in all ten paired lottery-choice decisions below (select your preferred option in each row). One of the pairs will be chosen at random, the lottery will be conducted and you will be paid according to the outcome of your preferred choice.

For example, if PAIR 1 (first row) is randomly chosen, and your preferred option is A, we will conduct a lottery where the chance of winning $2 is 1/10 (1 blue ball in an urn containing 10 balls) and the chance of winning $1.6 is 9/10 (9 green balls in the urn). If the blue ball is picked, you will receive $2. If the green ball is picked, you will receive $1.6.

Option A	Option B	Select A	Select B
1/10 of $2.00, 9/10 of $1.60	1/10 of $3.85, 9/10 of $0.80		
2/10 of $2.00, 8/10 of $1.60	2/10 of $3.85, 8/10 of $0.80		
3/10 of $2.00, 7/10 of $1.60	3/10 of $3.85, 7/10 of $0.80		
4/10 of $2.00, 6/10 of $1.60	4/10 of $3.85, 6/10 of $0.80		
5/10 of $2.00, 5/10 of $1.60	5/10 of $3.85, 5/10 of $0.80		
6/10 of $2.00, 4/10 of $1.60	6/10 of $3.85, 4/10 of $0.80		
7/10 of $2.00, 3/10 of $1.60	7/10 of $3.85, 3/10 of $0.80		
8/10 of $2.00, 2/10 of $1.60	8/10 of $3.85, 9210 of $0.80		
9/10 of $2.00, 1/10 of $1.60	9/10 of $3.85, 1/10 of $0.80		
10/10 of $2.00, 0/10 of $1.60	10/10 of $3.85, 0/10 of $0.80		

Separate Treatment Math Task

Welcome!

You are participating in a study in which you will earn some money. The amount will depend on a choice that you will have to make below. At the end of the study, your earnings (1 point = $ 1) will be added to a show-up fee, and you will be paid in cash.

Your choice: Another group of study participants has participated in two rounds of a task before this session. You will receive information on one of the participants, person A, and on how well that person performed in Round 1. You will then have to decide whether for Round 2, you want to be paid according to how well person A performed in Round 2, or whether you want to be paid according to the Round 2-performance of a randomly chosen person. You now receive information on the task, the average performance of all study participants in Round 1, as well as on one person's characteristics and performance in Round 1.

Information on task: In a previous study, participants were shown rows of five two digit numbers. Participants had to add up the numbers of each row. Participants were asked and incentivized to add up as many rows as possible. They had five minutes available for each round of the task. While the task was otherwise identical, they saw different sequences containing different numbers in Rounds 1 and 2.

Their point score was calculated as follows: For every correctly added sequence, 1 point was added to their score. Sequences that were not correctly added received no points.

To have a better understanding of the task, please click on this button to see a sample task. You will see the matrix for 30 seconds and not for 5 minutes.

SAMPLE TASK

Information on average Round 1-performance of all study participants: On average participants scored 10 points in Round 1.

Information on Participants: You will be paid according to the performance of person A or of a study participant drawn at random from all the people who participated in the study. We had 40 male and 40 female students participate, recruited by the Harvard Decision Science Laboratory.

Person A
Student
American
Boston Area
Male
Caucasian
Performance indicator: In Round 1, the person scored 10 points in five minutes.

Procedure to determine your earnings:

Once you have decided whether you want to be paid based on the performance of person A or a random person and have completed a short questionnaire, we will inform everyone of their point score and your payoffs.

If you chose to be paid according to the performance of the person described above, you will receive $1 x that person's point score for Round 2.

If you chose to be paid according to the performance of a random person, you will receive $1 x the random person's point score for Round 2.

If you have any questions, please press the help button now. Once we have addressed all questions, we will move to the main question of this study:

Main question: Do you want to be paid based on the Round 2-performance of the person described above, or do you want to be paid based on the Round 2-performance of a person drawn at random from all the people who participated in the study? (Please check one box)

NOTE: THE AVERAGE SCORE IN ROUND 1 WAS 10 POINTS

Person A	Random Draw
Student	
American	
Female	
Caucasian	
Performance indicator: In Round 1 the person	

scored 10 points in five minutes.

Note: after the main question of the experiment participants were notified of the score of the randomly selected candidate, and the score of person A.

Additional Decision.---Please choose Option A or Option B in all ten paired lottery-choice decisions below (select your preferred option in each row). One of the pairs will be chosen at random, the lottery will be conducted and you will be paid according to the outcome of your preferred choice.

For example, if PAIR 1 (first row) is randomly chosen, and your preferred option is A, we will conduct a lottery where the chance of winning $2 is 1/10 (1 blue ball in an urn containing 10 balls) and the chance of winning $1.6 is 9/10 (9 green balls in the urn). If the blue ball is picked, you will receive $2. If the green ball is picked, you will receive $1.6.

Option A	Option B	Select A	Select B
1/10 of $2.00, 9/10 of $1.60	1/10 of $3.85, 9/10 of $0.80		
2/10 of $2.00, 8/10 of $1.60	2/10 of $3.85, 8/10 of $0.80		
3/10 of $2.00, 7/10 of $1.60	3/10 of $3.85, 7/10 of $0.80		
4/10 of $2.00, 6/10 of $1.60	4/10 of $3.85, 6/10 of $0.80		
5/10 of $2.00, 5/10 of $1.60	5/10 of $3.85, 5/10 of $0.80		
6/10 of $2.00, 4/10 of $1.60	6/10 of $3.85, 4/10 of $0.80		
7/10 of $2.00, 3/10 of $1.60	7/10 of $3.85, 3/10 of $0.80		
8/10 of $2.00, 2/10 of $1.60	8/10 of $3.85, 9210 of $0.80		
9/10 of $2.00, 1/10 of $1.60	9/10 of $3.85, 1/10 of $0.80		
10/10 of $2.00, 0/10 of $1.60	10/10 of $3.85, 0/10 of $0.80		