WHEN PERFORMANCE TRUMPS GENDER BIAS:

JOINT VERSUS SEPARATE EVALUATION*

Iris Bohnet

Alexandra van Geen

Max H. Bazerman

Abstract

We examine a new intervention to overcome gender biases 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 focus on individual performance in joint than in separate evaluation and on group stereotypes in separate than in joint evaluation, making joint evaluation the money-maximizing evaluation procedure. Our findings are compatible with a behavioral model of information processing and with the System 1/System 2 distinction in behavioral decision research where people have two distinct modes of thinking that are activated under certain conditions.

JEL: C91; D03

Total Word Count: 8,236

---

*We thank Pinar Dogan, Chris Muris, Farzad Saidi, 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.

Corresponding author: Iris Bohnet, Harvard Kennedy School, 79 JFK Street, Cambridge, MA 02138; phone: 617 495 5605; iris_bohnet@harvard.edu

I. Introduction

Gender-based discrimination in hiring, promotion, and job assignments is difficult to overcome (e.g., Neumark, Bank, and Van Nort [1996]; Riach and Rich [2002]). In addition to conscious taste-based or statistical discrimination (Becker 1978), recent evidence suggests that gender biases are automatically activated as soon as evaluators learn the sex of a person. These 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]; Greenwald, McGhee, and Schwarz [1998]; Bertrand, Chugh, and Mullainathan [2005]). Attempts to decrease the role of gender biases in the evaluation of candidates have included blind evaluation procedures (Goldin and Rouse 2000) and gender quotas on hiring and promotion committees (Bagues and Esteve-Volart 2010; Zinovyeva and Bagues 2010).

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 evaluators to rely less on cognitive shortcuts, such as group stereotypes, when multiple candidates are presented simultaneously and evaluated comparatively than when they are evaluating one person at a time.

In hiring decisions, and particularly at more junior levels, it is common for candidates to be subject to "joint evaluation." Interviewing various candidates at the entry-level stage (for positions as analysts, programmers, or assistant professors, for example), organizations explicitly compare them with one another. By contrast, job assignments and promotion decisions are typically made on an individual basis, or through "separate evaluation": a manager is evaluated on whether she is ready to work on a more complex project, an attorney is assessed on whether

[1] For nudges more generally, see Thaler and Sunstein (2008).

he should be promoted to partner, or a junior faculty member is reviewed on whether she will be granted tenure.

It is at these more senior levels where the gender gap in economic participation is most pronounced (Bertrand and Hallock 2001). In Fortune 500 companies, for example, only 3.6 percent of CEOs, 14.1 percent of executive officers, and 16.1 percent of board members were female in 2011,[2] when at the same time, women made up 46.7 percent of the U.S. labor force (Bureau of Labor Statistics, 2011.) Gender differences in business-critical job assignments and promotions have been identified as contributing to the gender gap in career advancement (Ginther and Kahn 2009; Ibarra, Carter and Silva 2010; Zahidi and Ibarra 2010), in addition to supply-side factors, such as gender differences in training and time devoted to the job (Bertrand, Goldin, and Katz 2010).

We experimentally examine whether our evaluation nudge may help close gender gaps in economic participation in contexts where group characteristics are unrelated to performance. Our experimental design is intended to mimic the process of making internal job assignments, where an employer decides whether a given employee is suitable for a given job. More generally, it applies to all hiring, assignment, and promotion decisions where employers must evaluate a given candidate's suitability for a specific job, although in an organizational context, additional complexities come into play.

A change in assessments depending on the evaluation mode is compatible with the System 1/System 2 distinction made in behavioral decision research where people have two distinct modes of thinking that are variously activated under certain conditions: the intuitive and automatic System 1 and the reflective and reasoned System 2 (Stanovich and West 2000;

[2] http://www.catalyst.org/publication/132/us-women-in-business

Kahneman 2011). There is also an informational component to our evaluation nudge. In joint evaluation, more potentially counter-stereotypical data points are available at a time than in separate evaluation, thus providing evaluators with more information to update their stereotypical beliefs. This 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. While an information-based approach assumes more consciousness than an automatic switch between systems of thinking, it is also behavioral. It allows for evaluators to take into account irrelevant group characteristics that are not predictive of future behavior and then update their biased beliefs based on the new information received.

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

Our experimental findings have implications for the design of talent management and promotions procedures. Organizations may seek to overcome biases in 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. If so, they may be able to nudge evaluators toward taking individual performance information into account rather than gender stereotypes by introducing joint rather than separate evaluation procedures.

Our paper is organized as follows: Part II offers a conceptual framework, Part III describes the experimental design, Part IV reports our experimental results and Part V concludes.

II. Conceptual Framework

Evaluation procedures may affect decisions by making evaluators switch from a more intuitive mode of thinking in separate evaluation to a more reasoned approach in joint evaluation when assessing a candidate's likely future performance, and/or by providing different amounts of data that allow evaluators to update their beliefs about group characteristics to different degrees.

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).

It has been suggested that the lack of comparison information available in separate evaluation leads people to invoke intuitively available internal referents (see Kahneman and

Miller's [1986] norm theory), focus on the attributes that can be most easily calibrated (see Hsee's [1996] evaluability hypothesis), and rely more on emotional desires than on reasoned analysis (Bazerman, Tenbrunsel, and Wade-Benzoni's [1998] want/should proposition). In contrast, in joint evaluation, the availability of alternatives provides a comparison set for evaluation, makes calibration easier, and highlights the tradeoffs between what people want and what they should do.

There is also an informational component to joint evaluation. By definition, more information is available in joint than in separate evaluation. A "behavioral" Bayesian model of information processing could explain an increase in the likelihood that evaluators choose higher-performing employees in joint as compared to separate evaluation under certain conditions. Evaluating more than one person at a time implies having more data points available on the candidate's relative ability to update prior stereotypical 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 present this approach more formally in Appendix A. We take as a measure of a person's ability his/her past performance in a given task. For ease of understanding, we refer to people in this first stage as "employees" (but used neutral terminology in the experiment). In a second stage, the evaluators or "employers" are informed of the employees' past performance, their gender and the average past performance of the pool of employees. They then have to decide whether to "hire" the employee(s) presented to them for future performance in the task or go back to the pool and be allocated an employee at random. Employers are paid based on their employees' future performance and thus have an incentive to select who they believe to be most productive, based on the employee's future expected performance.

Both mechanisms, a switch in judgment modes and an informational approach, yield the same empirically testable prediction:

Employees are more likely to be selected based on their performance when evaluated jointly and more likely to be selected based on their gender when evaluated separately.

III. Experimental Design

Our experiment was conducted in the Harvard Decision Science Laboratory. A total of 654 individuals participated in the experiment. All were American college students. Among the participants, 554 played the role of employer; they participated in evaluating and selecting an employee for a job assignment. 180 participants played the role of employee (100 employees were new participants, 80 employees were drawn from a subset of employer participants who previously had participated in evaluations using a different task). We employed equal numbers of male and female employees.

All our participants were identified by code numbers and remained anonymous to each other and to the experimenter. To make gender less salient without creating any additional demographic variation, we took advantage of the demographic similarity of our employees and provided employers with truthful filler information on their employees' characteristics. In addition to learning a person's gender and past performance, employers were also informed that he or she was a student, American, and from the greater Boston area.

To examine the role of gender stereotypes, we employed two sex-typed tasks, a math and a verbal task. Most studies that measure explicit gender attitudes find that females are believed to be worse at math tasks and better at verbal tasks than males (e.g., Guimond and Roussel [2000]; Perie, Moran, and Lutkus [2005]). Implicit association tests (IATs) measuring

people's implicit attitudes report math and verbal skills to be associated with maleness and femaleness respectively (e.g., 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 (e.g., Xie and Shauman [2005]; Guiso et al. [2008]).

Figure 1 provides an overview of our experimental 2x2x2x2 main design. The key treatment condition of interest is how employers evaluated employees, separately or jointly. In addition, employees were either of high or low ability, male or female, and participated in either the math or the verbal task. We indicate the number of employers in each cell. Forty-four percent of the employers were male, 56 percent were female.

-----FIGURE 1 ABOUT HERE-----

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, employee participants participated in either a verbal or a math task and were paid based on their performance. Participants 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 2011). Most letters appeared in random order, but some formed words, and participants could search horizontally, vertically, and diagonally. On average, the 100 participants 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). On average, the 80 participants 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). Employee participants then were paid based on their performance and were not informed of Stage 2 of the experiment.

In stage 2, employers in both the verbal and the math tasks were asked to choose an employee, knowing that they would be paid based on that employee's Round 2-performance. They could either choose an employee presented to them or go back to the pool and accept a randomly selected employee. They had the person's Round 1-performance and his or her gender available as a basis for their decision (plus the filler demographic information). In addition, they were informed that on average, the employees in the pool had provided 10 correct answers (as was the case for both tasks).

The employees presented to the employers were either of average or slightly below-average ability, 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 employers' risk (or loss) aversion. We expect gender biases to play a smaller role when an employee's individual past performance clearly dominates the expected value of the pool and thus, expect our framework to be most relevant for situations where there is some ambiguity about an employee's past performance and/or where hiring or not hiring is a "close call."

In the separate-evaluation condition, employers were presented with either a male or a female employee who was either an average- or below-average performer. We randomly selected

four employees of the required gender-performance combinations from our pool: Male-10, Female-10, Male-9, and Female-9; they all had identical filler characteristics. In the joint-evaluation condition, employers were presented with a male and a female employee simultaneously, drawing from the same employees used in the separate-evaluation condition. The employees differed on both gender and past performance, leading to two possible combinations: Male-10/Female-9 and Male-9/Female-10.

After the experiment was completed, employers participated in an incentivized risk-attitude assessment task (Holt and Laury 2002) and completed a short questionnaire that collected basic demographic information. Employers were paid based on either their chosen employee's second-round performance or a randomly allocated employee's second-round performance. They received $1 for every correct answer that the employee had provided. Employer earnings varied between $17.8 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 employers were informed of an employee's second-round performance and then had to decide whether or not to select this employee and be paid based on the employee's performance in the second round or go back to the pool and accept a randomly allocated employee. This experiment was designed to distinguish belief-based from taste-based discrimination. While in our main experiment, both motives could lead to gender-based 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 employers 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 employers information about employees' 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.

IV. Results

We first focus on employee performance and 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. Thus, even accurate group stereotypes could become irrelevant, as an employee's past performance might be much more predictive of her future performance.

Table I reports the regression results of first-round performance and gender on 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 employee 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

[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.

include an interaction term between the two variables. For example, strong first-round performance of an employee 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 performance). Columns 2 and 4 suggest that first-round performance was equally predictive of future performance for both genders.[4]

-----TABLE I ABOUT HERE---

We now examine employers' choices. We aggregate across both evaluation modes and both performance levels. In the math task, the likelihood that the stereotype-disadvantaged employee, i.e., the woman, was chosen was 0.4, and the likelihood that the stereotype-advantaged man was chosen was 0.46. In the verbal task, the likelihood that the stereotype-disadvantaged man was chosen across conditions was 0.39, while the likelihood that the stereotype-advantaged woman was chosen across conditions was 0.5. Thus, employers have a slight preference for men in math tasks and for women in verbal tasks.

Looking at the two evaluation modes separately, we find that these gender differences in preferences are entirely driven by the stereotype-advantaged group being preferred in separate evaluation: Across both tasks, the likelihood that an employee from the stereotype-advantaged group was chosen was 0.66 when evaluated separately (thus 34 percent opted to go back to the pool) and 0.32 when evaluated jointly (thus 68 percent opted to go back to the pool or chose the stereotype disadvantaged person). Indeed, in joint evaluation, stereotypes did not matter at all; 32 percent of the employers chose an employee from the advantaged group and 31 percent from the

[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 employees 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.

disadvantaged group. The remainder of the employers, 37 percent, decided to go back to the pool and choose a random employee.

If employers had chosen randomly among the options available in the two evaluation modes, then, on average, 50 percent would have chosen a given employee in separate evaluation, and the other 50 percent would have gone back to the pool. In contrast, in joint evaluation, where by design employers had three options from which to choose, 33 percent of the employers would have chosen a given employee randomly on average. Thus, the stereotype-advantaged employees were significantly more likely to be chosen than what a random process would have predicted in separate but not in joint evaluation.

Figure II reports the data in more detail. It shows the percentages of average- and below-average performing male and female employees chosen in the math and the verbal tasks for separate and joint evaluation in the double-round experiments.

----FIGURE II ABOUT HERE ----

Past performance seems to be crucial in joint evaluation, while gender plays a critical role in separate evaluation. There are significant gender gaps in the likelihood of being chosen in separate but not in joint evaluation in both tasks. In contrast, there is a significant difference in the likelihood that a higher- rather than a lower-performing employee was chosen in joint but not in separate evaluation in the math task (In the verbal task, past performance matters in both evaluation modes.) These patterns lead to a preference reversal in the math sessions. In separate evaluation, 65 percent of the employers selected the lower-performing male employee (Round 1 score: 9) while only 44 percent chose the higher-performing female employee (Round 1 score: 10). In contrast, in joint evaluation, 57 percent of the employers preferred the higher-performing female employee, and only 3 percent chose the lower-performing male employee.

Generally, the likelihood that a given employee was chosen was higher in separate than in joint evaluation (with the exception of higher-performing women in the math task). We attribute this to the number of options available in separate versus joint evaluation. If employers had chosen randomly, a given employee would have been chosen by 50 percent of the employers in separate evaluation and by only 33 percent in joint evaluation.

To more precisely examine the differences in outcomes between separate and joint evaluation, we estimate a series of regressions. Each individual *i* is either in a separate or a joint treatment. We have *r* people in single and *q* people in the joint treatments. For the separate treatment, we generate the vector $\boldsymbol{y}_s$'=[c_1 .. c_r], where c_i denotes whether person *i* selected the described candidate (and thus did not go back to the pool). In the joint treatment, we pool the data and generate the vectors $\boldsymbol{y}_f$= [c_{f1}....c_{fq}] and $\boldsymbol{y}_{mi}$= [c_{m1} ... c_{mq}], where c_{fi} denotes whether person *i* selected the female candidate and c_{mi} indicates whether person *i* selected a male. Note that these are both zero whenever person *i* chose to go back to the pool. To obtain the probability that a candidate is selected, we estimated a probit regression, clustering the standard errors on *i*.[5]

Table II reports the marginal effects at the mean for the probit regressions on employee selection. We focus on separate evaluation in Column 1 and on joint evaluation in Column 2, and we include both evaluation modes in Columns 3-6. The group stereotype (*Stereotype-Advantage*) only affected the likelihood of being chosen in separate evaluation, while the employee's ability (*Performance*) only affected the likelihood of being chosen in joint evaluation. Columns 3 and 4 show that this difference is significant: Employees were significantly more likely to be selected based on their performance when evaluated jointly rather than separately (*Performance x*

[5] The pooling of the data in the joint treatment results in 2*q observations as for each individual we obtain both c_{mi} and c_{fi}. We let $\boldsymbol{Y}$'=[$\boldsymbol{y}_s$'; $\boldsymbol{y}_f$'; $\mathbf{y}_m$'] and let $\boldsymbol{X}$' = [$\boldsymbol{x}_s$'; $\boldsymbol{x}_f$'; $\boldsymbol{x}_m$'] be the matrix with the explanatory variables. Note that $\boldsymbol{x}_f$ and $\boldsymbol{x}_m$ are the same except for the indicators for the gender and the previous round performance of the candidate.

Separate in Column 3) and significantly more likely to be selected based on their gender when evaluated separately rather than jointly (*Stereo-Advant. x Separate* in Column 4). Put differently, employers were significantly more likely to select the higher-performing employee rather than the lower-performing employee in joint rather than in separate evaluation, which accords with our prediction. The size of the effect is considerable, as the marginal effect negates most of the effect of past performance.

In Column 5, we include both interactions simultaneously. While the size of the effects of both past performance and stereotypes remains very similar, only performance remains significant. Employers basically stopped choosing lower-performing employees in joint evaluation. Our results are generally robust to the inclusion of additional control variables, such as employers' own gender (and their attitudes toward risk (*Male Employer* and *Risk Tolerance* in Column 6). The more risk tolerant as well as male employers were less likely to select an employee (i.e., more likely to choose the random option).

-----TABLE II ABOUT HERE----

In our single-round experiments, we replicate the separate-evaluation condition for higher-performing employees to examine whether the focus on group characteristics in our first set of experiments was driven by stereotypical beliefs about group performance in the two tasks or by an (implicit) distaste of female employees for math tasks and male employees for verbal tasks.

We do not find any evidence for taste-based discrimination in our 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 0.43. Specifically, instead of going back to the pool, in the math task, 35 percent of the employers chose the male and 41 percent the female employee; in the verbal task, 55 percent

chose the male and 56 percent the female employee. Women and men were as likely to be chosen for both tasks. In a simple regression (not shown), we also find that employers were significantly less likely to choose a given employee rather than going back to the pool in the math than in the verbal task. (In our main, two-round experiment, the coefficient on math is also negative but not significant.) Attitudes toward risk again significantly affected people's choices, with the more risk tolerant less likely to choose an employee.

V. Discussion and Conclusions

This paper shows that a joint-evaluation mode succeeds in helping employers choose based on past performance, irrespective of an employee's gender and the implicit stereotypes the employer may hold. In our experiments, gender, the group stereotype of interest, was not predictive of future performance on math and verbal tasks, but individual past performance was. Still, employers tasked to choose an employee for future performance were influenced by the candidate's gender in separate evaluation. In contrast, in joint evaluation, gender was irrelevant; employers were significantly more likely to choose the higher- rather than the lower-performing employee.

Extensive research in behavioral decision making suggests that employers may decide differently in joint than in separate evaluation because they switch from a more intuitive evaluation 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]; Gino et al. [2011]). In addition, joint evaluation might also affect choices by providing additional data that employers can use to update their stereotypical beliefs about a group to which an employee belongs. By definition, an employer 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 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. Organizations can move from separate-evaluation to joint-evaluation procedures to promote a more reasoned approach to decision making and maximize performance. In our experiment, where not choosing higher-performing employees was costless in expectation as the highest possible performance level of a given employee corresponded to the average in the pool, efficiency losses occurred when employers chose the below-average employee instead of going back to the pool. Only about 8 percent of the employers engaging in joint evaluation, as compared to about 51 percent of the employers engaging in separate evaluation, chose the under-performing employee. 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.

Joint evaluation is common for most hiring decisions, especially at the lower levels, but it is rarely used when job assignments and promotions are being considered. Companies concerned about discrimination in these phases of employment might choose to review how, for example, career-relevant jobs are assigned and how promotion decisions are made. According to the Corporate Gender Gap Report (Zahidi and Ibarra 2010), in most countries, fewer than 10 percent of career-relevant jobs are held by women. In economics departments at American universities, controlling for performance, women are less likely to be granted tenure than men (Ginther and Kahn [2004]; for other fields, see Ginther and Kahn [2009]). Such decisions are typically made

by committees, which evaluate candidates separately. While it is not always feasible to bundle promotion decisions and explicitly compare candidates, our research suggests that, whenever possible, joint evaluation would increase both efficiency and equality.

Iris Bohnet and Alexandra van Geen are affiliated with the Harvard Kennedy School. Max Bazerman is affiliated with the Harvard Business School and the Harvard Kennedy School.

Appendix A

We adopt a simple behavioral Bayesian model to show that if the variance of the prior expected difference between male and female employee performance is sufficiently high, then an (expected-value-maximizing) employer who does not select a higher-performing counter-stereotypical employee in separate evaluation may do so in joint evaluation.

We use a male-typed task as our example and assume that an employer has information on the employees' past performance and gender and has to choose an employee for future performance based on this information. For simplicity, we assume that the employees participate in a task for two rounds.

We define x_g as a given employee's Round 1-performance. Employers observe $\boldsymbol{x}=[x_f\, x_m]$ in joint evaluation and $\boldsymbol{x}=[x_f]$ or $\boldsymbol{x}=[x_m]$ in separate evaluation. On average, the score reflects the mean past performance for each gender, but there is some noise. Round 1- performance comes from a normal distribution with known variance (σ^2).

Employers use **x** to update their priors on the expected performance (Round 2-performance) of the employee. The average past performance across the genders is known to be μ. For simplicity, we assume no learning of employees, so μ equals the expected second round performance. The female/male distribution in the pool of employees is known to be 50:50.

We assume that employers have no taste for discrimination but that they hold stereotypical beliefs. Specifically, for the math task, the prior expected male performance, μ_m, exceeds the prior expected female performance, μ_f. Assuming equal, positive and known variances across the genders, the priors look as follows: $\theta_f \sim N(\mu_f,\nu^2)$ and $\theta_m \sim N(\mu_m,\nu^2)$. Because of the known gender distribution it also holds that $\mu = 0.5*\mu_f + 0.5*\mu_m$ with $\mu_m > \mu_f$.

We define the ex ante expected deviation from the mean performance for males and females as h, such that the prior expected performance of male employees is $\mu_m = \mu + h$, with h>0 and the prior expected performance of female employees is $\mu_f = \mu - h$. Because of symmetry, we have $(|\theta_g - \mu|) \sim N(h,\nu^2)$, where $g \in [m,f]$. By updating their expectation of h incorporating the performance in round 1 using Bayes rule[6], employers can derive the posterior distribution of θ given $\boldsymbol{x}$.

An employer in separate evaluation confronted with a below-average male employee (i.e. with $x_m = b$) will update beliefs about mean performance for males in the pool to

$$\mu_m^1 = (\mu + (h \mid x_m)) = b + \frac{\sigma^2}{\sigma^2 + \nu^2}(h + (\mu - b)).$$

If faced with an average female employee (i.e. with $x_f = \mu$), an employer will update beliefs about the mean performance of females in the pool to:

$$\mu_f^1 = (\mu - (h \mid x_f)) = \mu - \frac{\sigma^2}{\sigma^2 + \nu^2} h.$$

In joint evaluation, employers have two data points available; they use both the male and the female employees' Round 1-performance to update their prior of h. In the counter-stereotypical situation where an employer is confronted with a lower-performing male (x_m=b) and a higher-performing female employee ($x_f = \mu$), this results in updated beliefs of mean performance for males in the pool of

$$\mu_m^2 = (\mu + (h \mid x_m, x_f)) = \frac{(\mu + h)\sigma^2 + \nu^2(b + \mu)}{\sigma^2 + 2\nu^2}.$$

[6] Note that this assumes that the employer takes the presence of the other (stereotype-advantaged) employee as a random draw from the (suitable) pool. Thus the employer does not believe that a poor employee is selected to accompany the good one just to make her look better as this would make the Bayesian argument irrelevant.

It results in an updated mean for females in the pool of

$$\mu_f^2 = (\mu - (h \mid x_m, x_f)) = \frac{(\mu - h)\sigma^2 + \nu^2((2\mu - b) + \mu)}{\sigma^2 + 2\nu^2}.$$

We assume that employers are risk-neutral expected-value maximizers for whom the expected Round 2-performance (y) is a linear combination of Round 1-performance and the (updated) belief about the employee's performance based on gender; i.e., $E(y_g) = \alpha x_g + (1-\alpha)\mu^t_g$ with $0 < \alpha < 1$, $g \in [m,f]$ reflecting gender and where $t \in [1,2]$ refers to the treatment (i.e. separate or joint). If $\frac{h\sigma^2}{\nu^2} < (\mu - b)$ then the following three conditions hold:

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

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

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

Thus, whenever there is sufficient variance of the expected difference between male and female performance, there is enough counter stereotypical evidence, and employers are not too biased, it holds that (1) the expected Round 2- performance of a higher-performing female employee dominates the random option and (2) the lower-performing male option in joint evaluation. Therefore, the expected-value-maximizing employer will select the female in the joint treatment. Because of (3), the expected Round 2- performance of a higher-performing female employee is lower than the expected value of the random option in the separate treatment, and the expected value maximizing employer will select the random option in the separate treatment. For the parameters used in the experiment the condition reduces to $h\sigma^2 < \nu^2$.

Appendix B: 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!

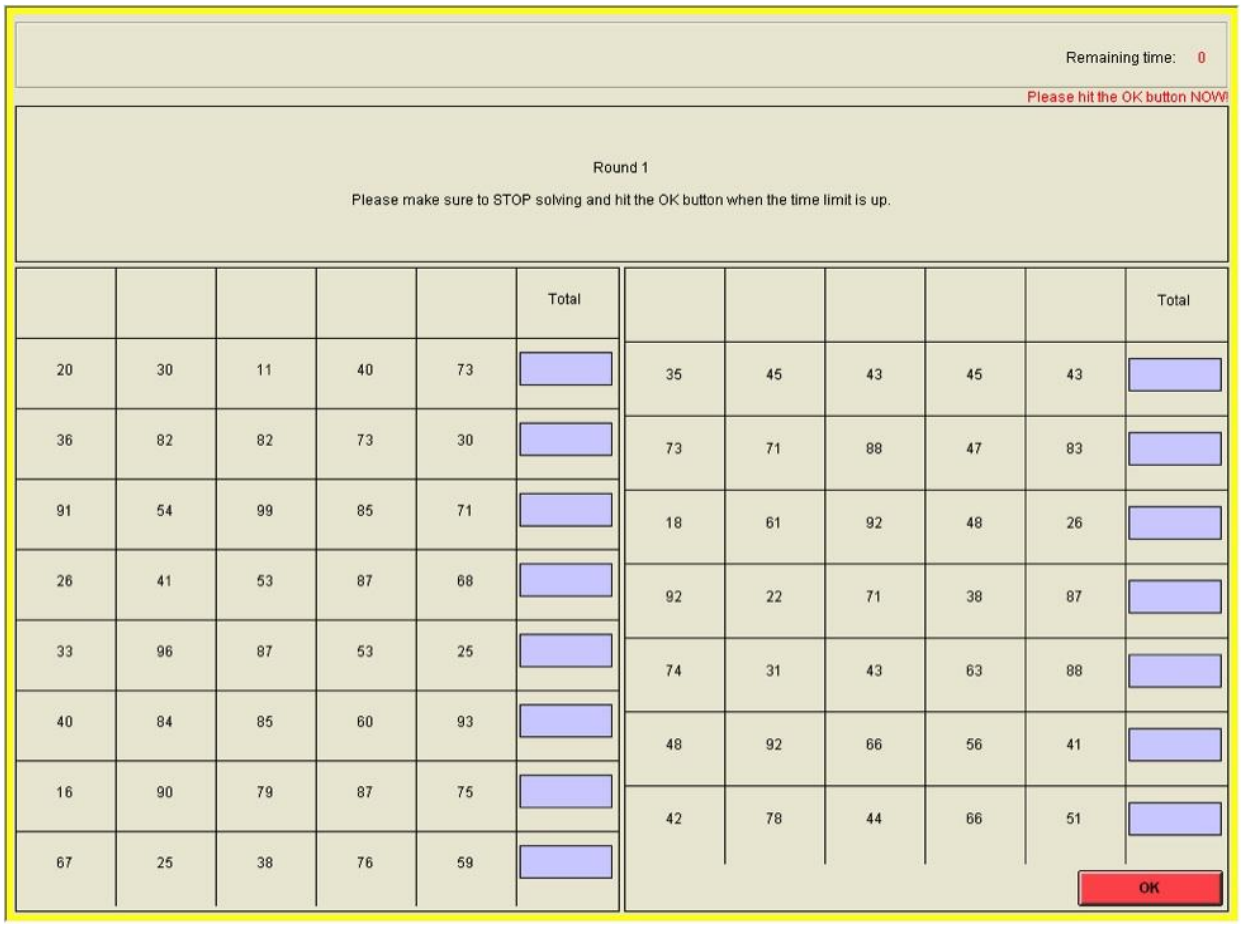
In this experiment you are asked to correctly solve as many math problems as possible. In each problem, you are asked to sum up five two-digit numbers.

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.



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

Instructions Stage 2

These were the instructions for the joint treatment with the math task and a high performing female candidate and a low performing male. Instructions for the other treatments were similar and are available upon request.

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

[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 three minutes.	Performance indicator: In Round 1, the person scored 9 points in three 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 three minutes.	Performance indicator: In Round 1, the person scored 9 points in three 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		

After performing the task, participants filled out a questionnaire collecting demographic information. Then, they were informed of their payoffs in the risk attitude question, all payoffs were summed up, and their final earnings were paid out in cash.

References

Bagues, Manuel, and Berta Esteve-Volart, "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment," *Review of Economic Studies*, 77 (2010), 1301-1328.

Banaji, Mahzarin R., and Anthony G. Greenwald, "Implicit Gender Stereotyping in Judgments of Fame," *Journal of Personality and Social Psychology*, 68 (1995), 181-198.

Bazerman, Max H., George F. Loewenstein, and Sally B. White, "Reversals of Preference in Allocation Decisions: Judging an Alternative Versus Choosing among Alternatives," *Administrative Science Quarterly*, 37 (1992), 220-240.

Bazerman, Max H., and Don A. Moore, *Judgment in Managerial Decision Making* (Hoboken, NJ: John Wiley & Sons, 2008).

Bazerman, Max H., Ann E. Tenbrunsel, and Kimberly A. Wade-Benzoni, "Negotiating with Yourself and Losing: Making Decisions with Competing Internal Preferences," *Academy of Management Review*, 23 (1998), 225-241.

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

Berger, James O., *Statistical Decision Theory and Bayesian Analysis* (New York: Springer-Verlag, 1985).

Bertrand, Marriane, Dolly Chugh, and Sendhil Mullainathan, "Implicit Discrimination," *American Economic Review*, 95 (2005), 94-98.

Bertrand, Marianne, Claudia Goldin, and Lawrence F. Katz, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics*, 2 (2010), 228-255.

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

Bohnet, Iris, and Farzad Saidi, "Informational Differences and Performance: Experimental Evidence." Working Paper, Harvard Kennedy School 2011.

Bureau of Labor Statistics, Current Population Survey, "Table 3: Employment Status of the Civilian Noninstitutional Population by Age, Sex, and Race," Annual Averages 2010, 2011.

Fischbacher, Urs, "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10 (2007), 171-178.

Gino, Francesca, Maurice E. Schweitzer, Nicole L. Mead, and Dan Ariely, "Unable to Resist Temptation: How Self-Control Depletion Promotes Unethical Behavior," *Organizational Behavior and Human Decision Processes*, 115 (2011), 191-203.

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

---, "Does Science Promote Women? Evidence from Academia 1973-2001," in *Science and Engineering Careers in the United States: An Analysis of Markets and Employment*, B. Freeman Richard, and L. Goroff Daniel, eds. (Chicago: University of Chicago Press, 2009).

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

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwarz, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology*, 74 (1998), 1464-1480.

Guimond, Serge, and Lydie Roussel, "Bragging About One's School Grades: Gender Stereotyping and Students' Perception of Their Abilities in Science, Mathematics, and Language," *Social Psychology of Education,* 4 (2001), 275-293.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales, "Diversity: Culture, Gender, and Math," *Science*, 320 (2008), 1164-1165.

Holt, Charles A., and Susan K. Laury, "Risk Aversion and Incentive Effects," *American Economic Review,* 92 (2002), 1644-1655.

Hsee, Christopher K., "The Evaluability Hypothesis: An Explanation for Preference Reversals Between Joint and Separate Evaluations of Alternatives," *Organizational Behavior and Human Decision Processes,* 67 (1996), 247-257.

Hsee, Cristopher K., Sally Blount, George F. Loewenstein, and Max H. Bazerman, "Preference Reversals Between Joint and Separate Evaluations of Options: A Review and Theoretical Analysis," *Psychological Bulletin,* 125 (1999), 576-590.

Ibarra, Herminia, Nancy M. Carter, and Christine Silva, "Why Men Still Get More Promotions Than Women," *Harvard Business Review,* 88 (2010), 80-126.

Kahneman, Daniel, *Decisions, Fast and Slow*, NY: Farrar, Strausand Giroux, 2011.

Kahneman, Daniel, and Dale T. Miller, "Norm Theory: Comparing Reality to its Alternatives," *Psychological Review,* 93 (1986), 136-153.

Kahneman, Daniel, Ilana Ritov, Karen E. Jacowitz, and Paul Grant, "Stated Willingness to Pay for Public Goods: A Psychological Perspective," *Psychological Science,* 4 (1993), 310-315.

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

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

Nosek, Brian, Mahzarin Banaji, and Anthony G. Greenwald, "Math= Male, Me= Female, Therefore Math [not equal to] Me," *Journal of Personality and Social Psychology*, 83 (2002), 44-59.

Nowlis, Stephen M., and Itamar Simonson, "Attribute-Task Compatibility as a Determinant of Consumer Preference Reversals," *Journal of Marketing Research*, 34 (1997), 205-218.

Paharia, Neeru, Kassam, Karim. S., Greene, Joshua D. & Bazerman, Max H. "Dirty Work, Clean Hands: The Moral Psychology of Indirect Agency". *Organizational Behavior and Human Decision Processes,* 109(2), (2009), 134-141.

Perie, Marianne, Rebecca Moran, and Anthony D. Lutkus, "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, 2005).

Plante, Isabelle, Manon Theoret, and Olga Eizner Favreau, "Student Gender Stereotypes: Contrasting the Perceived Maleness and Femaleness of Mathematics and Language," *Educational Psychology*, 29 (2009), 385-405.

Riach, Peter A., and Judith Rich, "Field Experiments of Discrimination in the Market Place," *The Economic Journal*, 112 (2002), 480-518.

Stanovich, Keith E., and Richard F. West, "Individual Differences in Reasoning: Implications for the Rationality Debate?," *Behavioral and Brain Sciences*, 23 (2000), 645-665.

Thaler, Richard H., and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (New Haven, CT: Yale University Press, 2008).

Xie, Yu, and Kimberlee A. Shauman, *Women in Science: Career Processes and Outcomes* (Cambridge, MA: Harvard University Press, 2005).

Zahidi, Saadia, and Herminia Ibarra, "The Corporate Gender Gap Report 2010 (Geneva, Switzerland: World Economic Forum, 2010).

Zinovyeva, Natalia, and Manuel F. Bagues, "Does Gender Matter for Academic Promotion? Evidence From a Randomized Natural Experiment," (Institute for the Study of Labor, 2010).

Table I

The Effect of Ability and Stereotypes on Employee Performance

	Math Task		Verbal Task	
	1	2	3	4
Performance	0.849***	0.797***	0.708***	0.813**
	(0.08)	(0.15)	(0.10)	(0.15)
Male Employee	0.420	-0.481	1.201	3.118
	(0.46)	(1.91)	(0.77)	(2.42)
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

Significance levels: *: .1, **: .05, *** .01. 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.

Table II

The Effect of Ability and Stereotypes on Employee Selection, MEM

	(1)	(2)	(3)	(4)	(5)	(6)
	Separate	Joint	Both	Both	Both	Both
Performance	0.099	0.462**	0.536**	0.325**	0.532**	0.545**
	(0.07)	(0.06)	(0.06)	(0.05)	(0.06)	(0.07)
Stereotype-Advantage	0.165**	0.009	0.089*	-0.02	0.012	0.017
	(0.07)	(0.07)	(0.05)	(0.07)	(0.08)	(0.08)
Math	-0.009	-0.043	-0.030	-0.029	-0.030	-0.025
	(0.07)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Separate			0.533**	0.203**	0.468**	0.475**
			(0.06)	(0.07)	(0.08)	(0.08)
Performance x Separate			-0.402**		-0.397**	-0.393**
			(0.06)		(0.07)	(0.07)
Stereo-Advant. x Separate				0.179*	0.153	0.157
				(0.11)	(0.11)	(0.11)
Risk Tolerance						-0.028**
						(0.01)
Male employer						-0.170**
						(0.05)
Decision outcomes	202	252	454	454	454	454
Pseudo R^2	0.0271	0.2201	0.1693	0.1358	0.1730	0.1999

Significance levels: *: .1, **: .05. 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 employee. 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 454 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).

Figure I

Experimental Design: Double-Round Experiments (Number of Subjects)

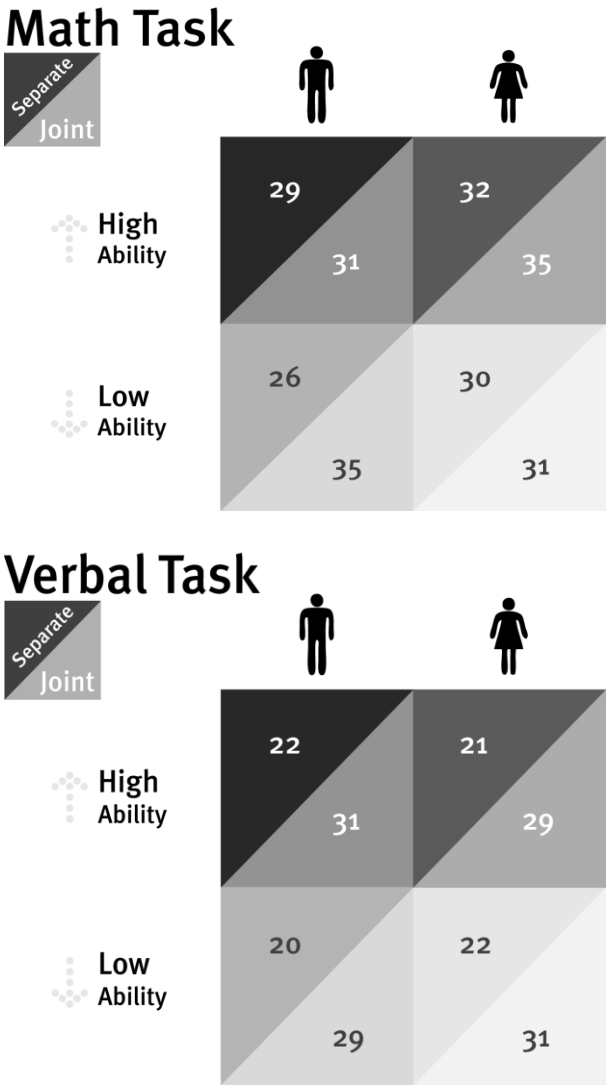


Figure II

Likelihood of Employee Selection in Separate and Joint Evaluation (Double Rounds)