Updating, Self-Confidence and Discrimination[1]

Konstanze Albrecht[a], Emma von Essen[b], Juliane Parys[c] and Nora Szech[d]

Draft: February 2012

Abstract

Our experiment shows that subjects incorporate irrelevant group information when evaluating others: Individuals from on average badly performing groups receive low evaluations, though it is known that they themselves perform well. This group-bias occurs in a gendered setup, where females form the on average worse performing group, as well as in a non-gendered setup.

The type of discrimination we identify is neither taste-based nor statistical: It is due to conservatism in updating beliefs, which is even more pronounced in females. Furthermore, self-confident males overvalue male performers. Using our data to simulate a job promotion ladder, females get driven out quickly.

JEL classification: D81, C91, J16

Keywords: updating, conservatism, Bayes' Rule, gender, discrimination, self-confidence

---

a) Department of Economics, Institute for Empirical Research in Economics, University of Bonn, Lennéstr. 43, 53113 Bonn, Germany, and Center for Economics and Neuroscience, Nachtigallenweg 86, 53127, Bonn, Germany, email: konstanze.albrecht@uni-bonn.de

b) Department of Economics, Stockholm University, SE-10691, Stockholm, Sweden, email: emma.vonessen@ne.su.se

c) Department of Economics, University of Bonn, Kaiserstr. 1, 53113 Bonn, Germany, email: jparys@uni-bonn.de

d) Department of Economics, Chair of Economic Theory II, University of Bonn, Lennéstr. 37, 53113 Bonn, Germany, email: nszech@uni-bonn.de

[1]We thank Stefano DellaVigna, Anna Dreber, Armin Falk, Thomas Gall, Magnus Johannesson, Ulrike Malmendier, Benny Moldovanu, Astri Muren, Frank Rosar, Hannah Schildberg-Hörisch, Matthew Rabin, Patrick Schmitz and Matthias Wibral for helpful comments. We further thank Patrizia Oedyniec, Anne Roesler, Benedikt Vogt, Jana Willrodt, Parwaneh Zekri and Ulf Zoelitz for their assistance. Financial support from the German Research Foundation, the Swedish council for working life and social research (FAS), as well as the Ministry of Innovation, Science, Research, and Technology of NRW is gratefully acknowledged.

1 Introduction

Consider two individuals who perform equally well. Is their performance evaluated the same way? In a laboratory experiment, we find that this is not the case if one individual belongs to a less favorable group. We observe this group-bias in evaluations in a gendered setup, where females form the on average worse performing group, as well as in a non-gendered setup: Discrimination occurs even though the evaluator knows that group belonging is completely irrelevant. Hence, the discrimination we identify has nothing to do with statistical discrimination. Also, there is no interaction involved between evaluator and performer and thus no room for taste-based discrimination. The discrimination we find is rather due to conservatism in updating beliefs. In female evaluators, this conservatism is even more pronounced. In the gendered setup, we furthermore find self-confident male evaluators to overvalue male performers.

In our study the information on group belonging is given as the prior (or base-rate information), and the information on the individual as the new information. We hence analyze a simple updating problem. In the economic and psychological literature on updating, there is broad evidence that people do not update according to Bayes' Rule: Instead, they are either too conservative by putting too much weight on the prior[2], or neglect the prior when they should not.[3] Typical updating problems are mathematically complex: They require reweighting the prior with the new information. In contrast, in our experiment Bayes' Rule just demands to give up the prior completely. Despite the simplicity of the problem subjects do not follow Bayes' Rule. Furthermore, females put more weight on the irrelevant prior than males.[4]

We frame the updating problem in two different ways: In one set of treatments we use neutral labels to distinguish the performers' groups; concretely, we call the groups K and L. In a second set of treatments we use gender as the defining property of the two groups: Subjects evaluate either a man or an equally well performing woman. Evaluators in both setups update more conservatively than rationality would prescribe. Conservatism in updating may consequently be a potential source of discrimination observed in the labor market, pertaining not only to gender but to any group with less favorable characteristics.

In the economic literature on labor market discrimination, the focus so far has been on two possible rationalizations of discriminatory behavior: Taste-based discrimination (Becker 1957) assumes that individuals have preferences against interacting with individuals of certain groups and therefore

[2] see e.g. Lyon and Slovic 1976; Bar-Hillel 1980; Falk, Huffman, and Sunde 2006; Möbius et al. 2011
[3] see e.g. Tversky and Kahneman 1971; Kahneman and Tversky 1972; Grether 1992; El-Gamal and Grether 1995
[4] Such gender differences in updating are in line with findings from much more demanding updating tasks (Charness and Levin 2005; Falk, Huffman, and Sunde 2006; Dohmen et al. 2009; Möbius et al. 2011).

discriminate them. In contrast, according to the theory of statistical discrimination (Phelps 1972; Arrow 1973), discriminatory behavior arises due to informational frictions. Neither of these theories can explain our findings.[5] The discrimination we observe is non-rational and occurs due to a lack of basic statistical ability.

Closest to our study is Reuben et al. (2010) who also try to identify updating problems as a source of discriminatory behavior. They find that in a competitive real effort task, women are less often chosen as leaders than men even though there are no gender differences in previous performance. However, although prior performance is known in their experiment, uncertainty about the future performance leaves room for statistical discrimination. Furthermore, the chosen leader receives money for being chosen, which may provoke taste-based discrimination. In our study, we eliminate any possibility of such informational frictions and taste-driven behavior: Future performance is irrelevant and no direct or indirect interaction (e.g. via payments) between evaluator and performer occurs. Hence our study neatly shows that conservatism in updating is indeed a source of discriminatory behavior.

As an ex post result, we furthermore find a relation between self-confidence and gender discrimination: Self-confident males tend to overvalue male performers. Probably, males project their positive attitude towards themselves on performers of their own gender, but not on females. In psychology, such overestimation of similarity is well-known under the name of "the false consensus effect" (e.g. Ross, Greene, and House 1977; Bauman and Geher 2002). In general, previous literature indicates that highly self-confident individuals behave differently from individuals with low levels of self-confidence (e.g. Falk, Huffman, and Sunde 2006; Niederle and Vesterlund 2007).

We additionally use the results from our study to calibrate a simple model of a job promotion ladder: In each round, employees are promoted with probabilities derived from the evaluations made in the laboratory. We demonstrate a glass ceiling effect, i.e., there are virtually no females left after few rounds of promotions. Also, the proportions of females on the different hierarchy levels fit well to observed proportions in the real world, like academia.

The remainder of the paper is organized as follows: Section 2 introduces the design of our experiment. In Section 3, we present our results. Section 4 describes our numerical simulations of the glass ceiling effect. Section 5 concludes.

[5] There is a vast empirical literature supporting the existence of discrimination in the labor market. For an overview, see Anderson, Fryer and Holt (2006) and Blau and DeVaro (2007). For example, employers might prefer to rely on group averages rather than bearing the costs of an interview (Anderson, Fryer, and Holt 2006).

2 Design of the Experiment

The study consists of two separate stages. In the first stage, subjects perform a series of mental rotation tasks (MRTs) and are assigned to one of two groups. Second stage subjects then evaluate the performance of a randomly assigned first stage subject from one of the two groups, who differ in performance averages. The selection and assignment procedures thereby render prior information on the group performance of one of the two groups irrelevant.[6]

First Stage (Pre-Study)

In the first stage of the study, 91 subjects, called *performers*, participate (50 female, age range: 17 to 49 years, mean age: 23.12 years). 24 mental rotation tasks are presented to each subject. Each task consists of five pictures of three-dimensional objects, one being the original object, and four being rotated or mirrored versions of the original object (adapted from Peters et al. 1995; Vandenberg and Kuse 1978). Subjects indicate which two of the four objects were rotated, but not mirrored. They have two times three minutes to solve as many as possible of twelve such tasks in each of the two three minute periods. Afterwards, we ask subjects to estimate their own performance and the average male and female performance in the task. Subjects are assured that all data is treated anonymously. Each subject is paid a flat amount of EUR 2.00 for participation. An example of the task is provided in Figure 1.

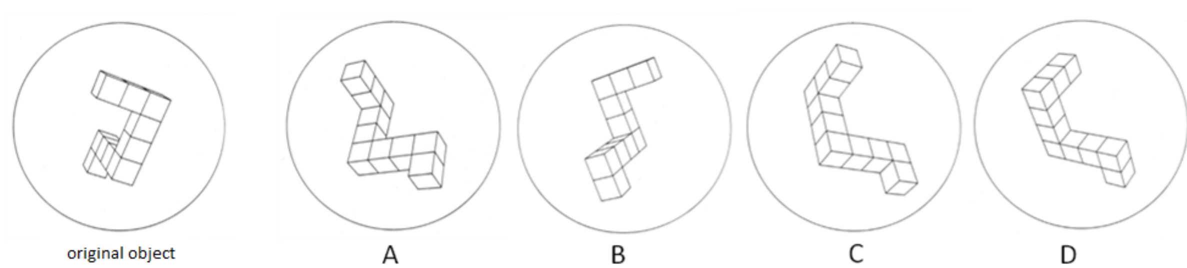


Figure 1. Example of a mental rotation task presented to the subject. The leftmost object is the original object. Subjects have to indicate which two of the four objects (A-D) are rotated but not mirrored versions of the original object. In the example, the correct solutions are B and D.

[6]The study was conducted in the Bonn Econ Lab in Bonn, Germany. Subjects were recruited via ORSEE (Greiner 2003) and mainly students at Bonn University. Fischbacher’s (2007) software zTree was used to present the tasks to the subjects.

Second Stage (Main Study)

In the second stage of the study, 308 subjects, called *evaluators*, participate. Out of these, 305 get all control questions right and are therefore included in later analyses (153 females, age range: 19-63 years, mean age: 24.70 years). No subject participating in this part of the study participated in the first stage.

All second stage subjects are informed about the first stage of the study and that they may earn money depending on their own decisions and on the performance of a randomly assigned first stage subject.

We randomly allocate second stage subjects into four treatments: Two neutral treatments, **Neutral** and **Selected-Neutral**, as well as two gendered treatments, **Man** and **Selected-Woman**.

Neutral Treatments

In both neutral treatments, subjects are informed via pie diagrams how well subjects from two groups, called group K and group L, performed in the first stage: One pie diagram shows that 43% of the subjects from group K were top, whereas 57% performed in a mediocre way. Another pie diagram shows that 14% of the subjects from group L were top, whereas 86% were mediocre (Figure 2).

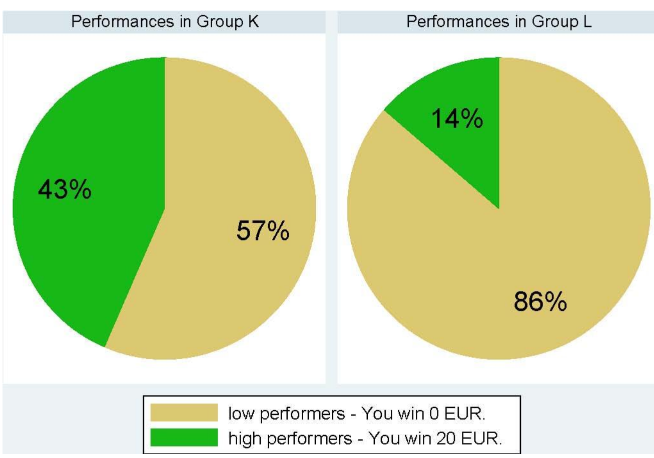


Figure 2. Pie diagrams presented to subjects

Subjects are further informed that "top" means subjects solved more than 13 tasks correctly and "mediocre" means subjects solved at least 9 and at most 13 tasks correctly. Second stage subjects do not perform MRTs themselves, but are presented the example from Figure 1.

Selection Process in the Neutral Treatments

The selection process of the afterwards assigned first stage performer is carefully described to the subjects:

- We randomly draw a first stage subject out of group K.
- We look at the performance level of the drawn subject.
- We draw a subject from group L that performed on the same level.

This means that if the randomly drawn group K subject is top, then a group L subject is selected who is also top. If the group K subject is mediocre, a group L subject is selected who is also mediocre.

Subjects are informed that they will be randomly assigned to a performer from group K (treatment Neutral) or to a selected performer from group L (treatment Selected-Neutral). Thus the probability of facing a top performer is 43% in all treatments. We make sure by control questions that evaluators understand the selection procedure, see Appendix 1.

Gendered Treatments

The gendered treatments are equal to the neutral treatments, but consider gender as an attribute that defines groups. Hence, here group K and group L performers are labeled male and female performers instead. Accordingly, the treatments are named Man and Selected-Woman. The order of the naming of the groups is counterbalanced throughout the study. An overview of treatments is provided in Table 1.

Table 1: Treatment overview

Treatment		Description
Neutral treatments	Neutral	Evaluators face a randomly drawn performer from group K.
	Selected-Neutral	Evaluators face a performer from group L, who is selected as follows: - A performer from group K is randomly chosen. - If the performer from group K is top, then a performer from group L is selected who is also top. - If the performer from group K is mediocre in, a performer from group L is selected who is also mediocre.
Gendered treatments	Man	Evaluators face a randomly drawn male performer.
	Selected-Woman	Evaluators face a female performer, who is selected as follows: - A male performer is randomly chosen. - If the male performer is top, then a female performer is selected who is also top. - If the male performer is mediocre, a female performer is selected who is also mediocre.

Evaluation Procedure

We elicit the evaluations by letting the evaluators face a series of 50 choices between a certain outcome and a lottery, varying the certain outcome from EUR 0.40 up to EUR 20.00. The lottery outcome depends on the performance of the first stage performer assigned to the evaluator: If the performer is top, the evaluator wins the lottery (and receives EUR 20.00). If the performer is mediocre, the evaluator loses the lottery (and receives EUR 0.00). The variable we use is the decision where evaluators switch from the risky option to the safe option.

Post-Experiment Survey

After the evaluators made their choices, they are asked to answer a survey about their socio-demographic background. Furthermore, evaluators are asked to hypothetically estimate their own performance in the MRTs. At the end of the experiment, one of the 50 decisions is randomly drawn for payment.

3 Results and Discussion

We start by presenting summary statistics of the first stage, showing that there are substantially more male than female top performers in the MRTs. Then, we analyze the data from the second and main part of our experiment: We first show that subjects are generally conservative in updating, irrationally putting positive weight on the irrelevant prior. Splitting our sample by gender, we find that conservatism is more pronounced in females than in males. The results are essentially the same for the neutral frame and for the gendered frame with the exception that self-confident males are found to strongly overvalue male performers in the gendered frame.

Performance evaluations are stated in amounts of EUR. All analyses are conducted using t-tests.[7] We further apply the parametric strategy proposed by Crump, Hotz, Imbens & Mitnik (2008) to address potential multiple testing concerns.[8]

First Stage

There are 43% male top performers, and accordingly 57% male mediocre performers. Only 14% of female performers are top, whereas 86% are mediocre. There are significantly more male than female top performers ($t(43) = 2.29$, $p = .03$).

Second Stage

We define the first decision where an evaluator switches from the risky to the safe option as the switching point. We take this switching point as the evaluator's evaluation of the performer. 33 evaluators switch between the risky and the safe options multiple times and are therefore excluded from the main analyses.[9] We furthermore exclude the three of the 308 evaluators who do not answer all control questions correctly. Summary statistics of the evaluations across treatments are provided in Table A1 in Appendix 2.

---

[7] A Kolmogorov-Smirnov test cannot reject the hypothesis that evaluations are normally distributed in the sample ($p > .05$).

[8] Hereby, two null hypotheses about the average treatment effect are tested. The first hypothesis is that evaluating a performer in the Selected-Neutral treatment instead of a performer in the Neutral treatment has a zero average effect for male and for female evaluators. We hold the same hypothesis concerning the gendered treatments. The second hypothesis we test is that the average treatment effect is identical for male and female evaluators, i.e. there is no heterogeneity in the average treatment effect.

[9] As a robustness check we look at average switching points of all evaluators including multiple switchers. The results are similar for both measures. Although the proportion of female evaluators, who switch multiple times, is significantly higher than among male subjects ($t(303) = 3.15$, $p < .01$), there is no gender difference in the average switching point among subjects who switch multiple times ($t(31) = 1.41$, $p = 0.17$).

In both frames, neutral and gendered, we find that evaluators switch significantly earlier in the Selected-Neutral and Selected-Woman treatments compared to the Neutral and Man treatments (neutral treatments: $t(122) = 3.77$, $p < .01$; gendered treatments: $t(125) = 3.24$, $p < .01$); see Figure 3. Hence evaluators are too conservative in updating, i.e. they take into account the group's average performance in the Selected-Neutral and Selected-Woman treatments. All in the analyses included subjects answered the control questions correctly. Hence they perfectly know that the prior should be irrelevant, but are not able to update their beliefs accordingly. The fact that this effect is significant in both the neutral and the gender frame suggests that conservatism in updating is a general phenomenon: Conservatism in updating may be a source of discrimination pertaining not only to gender but to any group with less favorable average characteristics. Table A2 in Appendix 2 provides summary statistics.

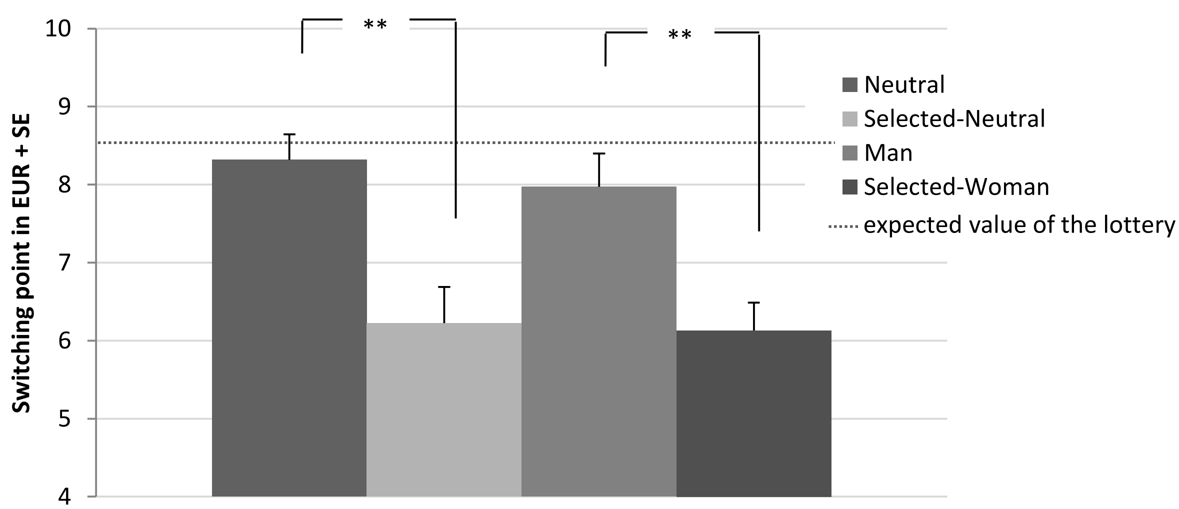


Figure 3. Evaluations in the four treatments. (SE = Standard Error; *p<.05, **p<.01)

Gender Differences in Evaluations between Treatments

Since there is a general gender difference in evaluations ($t(270) = 2.44$, $p = .02$) we further analyze evaluations separately for male and female evaluators.

We start by investigating the general evaluations in the neutral treatments, i.e. the evaluations from treatments Neutral and Selected-Neutral. Mean evaluations between these treatments display no significant differences among male evaluators ($t(62) = 1.31$, $p = .20$). In contrast, we find a highly significant difference for female evaluators ($t(58) = 4.39$, $p < .01$). On average, females evaluate performers from the better group EUR 3.29 higher than the selected performers from the worse group. Further, the difference in evaluations between the two neutral treatments is significantly higher for female compared to male evaluators (t(118) = 2.06, p = .02). Hence, females are more conservative in updating than males. An overview of differences in mean evaluations is presented in Figure 3 and Table A2 in Appendix 2.[10]

As in the neutral treatments, in the gendered treatments females evaluate the performance of a selected woman significantly lower than the performance of a randomly drawn man (t(56) = 2.77, p = .01). For male evaluators, the difference in evaluations between the gendered treatments is also significant (t(67) = 1.96, p < .05). For male subjects this difference is marginally significant in the gendered treatments, but not significantly stronger than in the neutral treatments (t(129) = 0.45, p = .33). Also, for female subjects, there is no difference in updating between the gendered and the neutral treatments (t(114) = -1.03, p = .15). An overview of differences in mean evaluations is presented in Figure 3 and Table A2 in Appendix 2.[11]

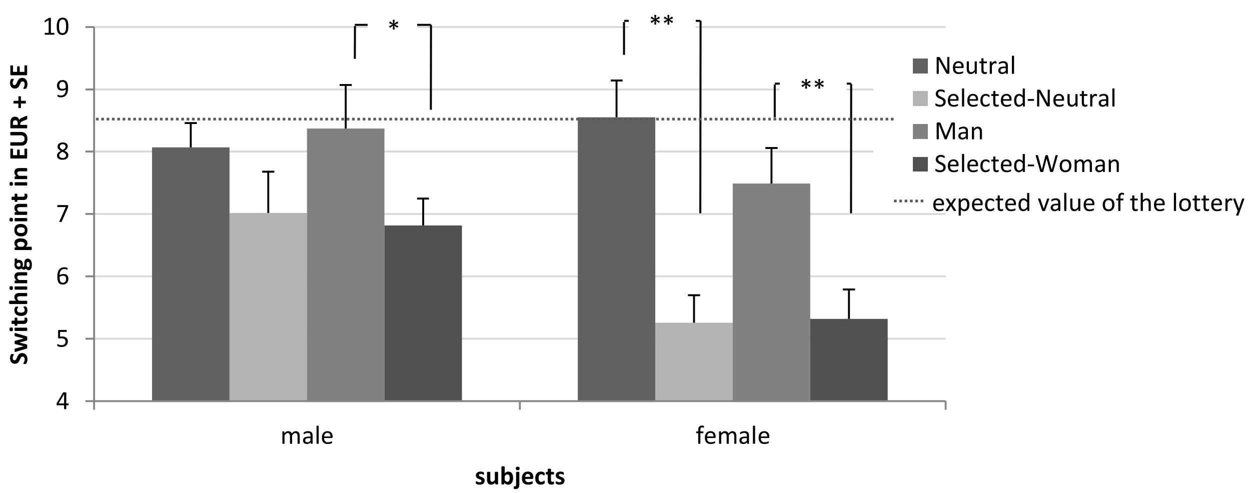


Figure 4. Evaluations by gender. (SE = Standard Error; *p<.05, **p<.01)

[10]Tables A3 and A4 in Appendix 2 provide coefficients estimated from OLS regressions. The results support the findings of the t-tests. In addition, we see that R^2 is .25 in the female regression as compared to only .06 in the male regression, which indicates that differences between treatments explain much more of the variation found in female subjects' evaluations.

[11]In the regressions in Tables A3 and A4, the dummy variable for the gendered treatments and its interactions with a neutral-treatment's dummy and a male dummy are insignificant.

Hence we find an overall indication for conservatism in updating. The gendered framing seems to increase this conservatism.

The Influence of Self-Confidence

In line with previous literature (Falk, Huffman, and Sunde 2006; Niederle and Vesterlund 2007) we find that male evaluators are more optimistic about their own (hypothetical) performance in MRTs than female evaluators ($t(303) = 5.78$, $p < .01$). We therefore investigate the influence of self-confidence on discrimination. To measure self-confidence, we take the beliefs about how many MRTs the evaluators think they would have solved themselves if they had participated in the first stage. Based on this measure, we construct two groups of evaluators: In the first group, there are evaluators whose beliefs about their own hypothetical performance are above the median belief (high self-confidence). In the second group are those with beliefs below the median (low self-confidence), within each treatment and gender. Figure 4 and Table A5 in Appendix 2 provide the performance evaluations of male and female evaluators split by level of self-confidence.[12]

[12]As a robustness check, we alternatively use the mean instead of the median for the sample split. We further create two additional measures of self-confidence: Firstly, a relative self-confidence measure is constructed performing a median-split on the difference between beliefs about the own performance and the performance of the corresponding performer. Secondly, we construct a self-confidence measure where we classify subjects as being self-confident if, according to their beliefs about their own performance, they would count themselves to the group of top performers; i.e. the belief about their own performance is to solve fourteen or more MRTs correctly. Results based on these alternative self-confidence measures lead to qualitatively similar results as the main measure described in the text, and are provided in Tables A6-A8 in Appendix 2.

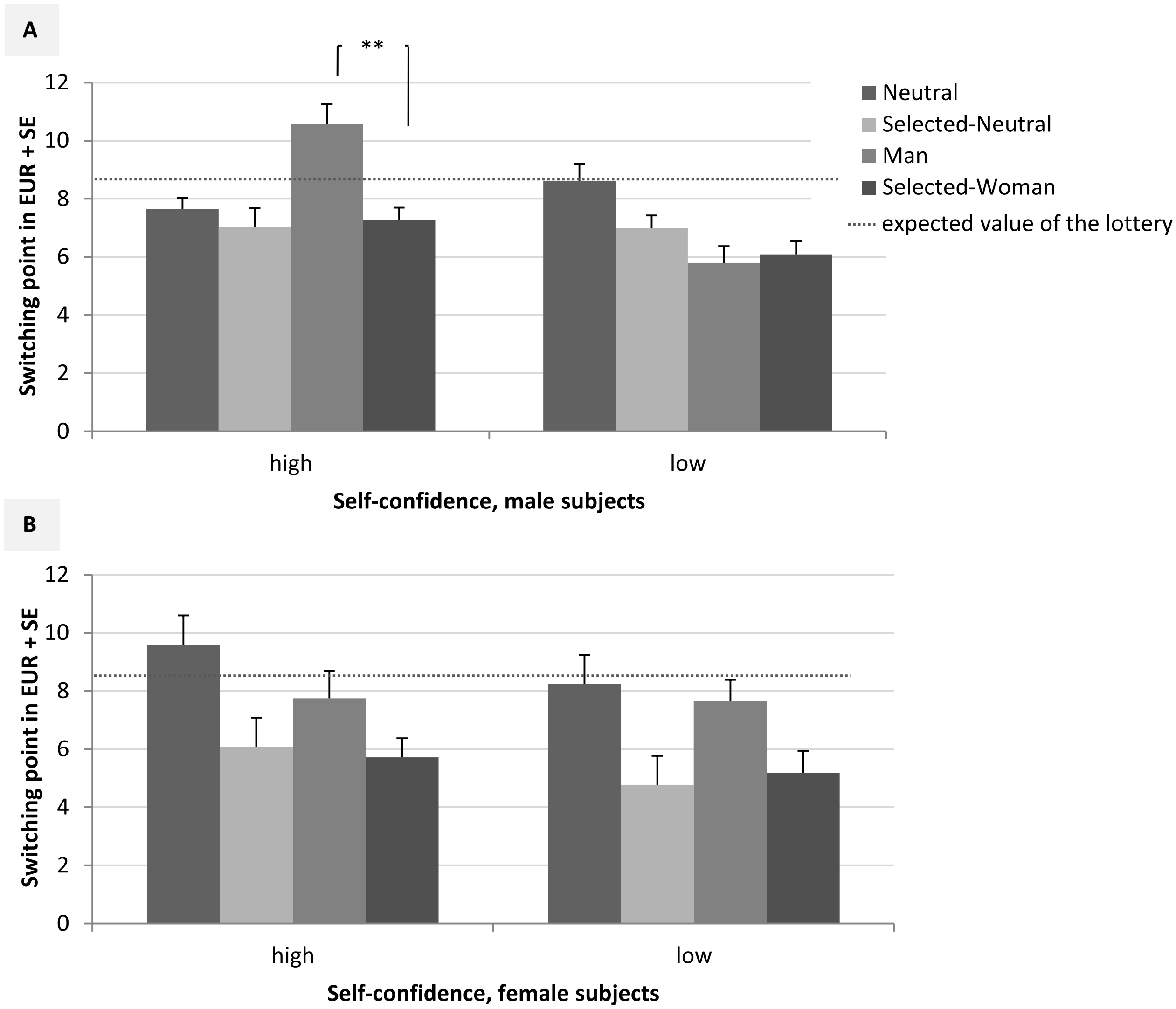


Figure 4. Evaluations by level of self-confidence of **A)** male, and **B)** female evaluators. (SE = Standard Error; *p<.05, **p<.01)

The results displayed in Figure 4 show a clear pattern indicating that self-confident male evaluators discriminate in the gendered treatments ($t(26) = 3.08$, $p < .01$), but not in the neutral treatments ($t(24) = .87$, $p = .40$). Importantly, the average evaluation of another man's performance (EUR 10.56) is well above the expected value of the lottery (EUR 8.60). Hence, self-confident male evaluators in the Man treatment lose money by switching to the save option too late. Comparing the mean evaluations between treatments, we conclude that male evaluators with high levels of self-confidence overvalue the performance of other men as opposed to undervaluing the performance of selected women. In contrast, male evaluators with low levels of self-confidence do not overvalue subjects from high performing groups (neutral: t(24) = 1.12, p = .28, gendered: t(26) = 0.24, p = .81).[13] We conclude that self-confident men are sensitive to the gender frame, although they know that there are no performance differences.

In line with psychological research on "the false consensus effect", our results may be caused by men projecting their own self-confidence on other men. The false consensus effect states that individuals overestimate the extent to which others have similar beliefs, opinions, preferences and habits as they themselves have (e.g. Ross, Greene, and House 1977; Bauman and Geher 2002). Men might consider themselves more similar to other men than to women, and accordingly project their own self-confidence only on men.

Females do not display this pattern. Self-confidence does not seem to play a role when evaluating other subjects' performance (Figure 4).[14] Self-confidence hence seems to be reflected in the behavior of men, but not of women.

4 Simulating the Glass Ceiling

In this section we investigate whether the (comparatively small) differences in male and female evaluation behavior can explain the glass ceiling phenomenon, i.e. the extremely small proportion of female employees on higher levels in most job promotion hierarchies. For this purpose we consider a simple numerical model of job promotions.

The Model

In the model there are *t* hierarchy levels in a firm with *n* employees at each hierarchy level. At each level there are male and female employees. Employees at level *s* are split randomly into *m* groups of size *g*. Each group is assigned a male evaluator with probability p_s and a female evaluator otherwise. Males in a group with a male evaluator are assigned a random evaluation drawn from the evaluations of males by male evaluators in the gendered treatments of our experiment. Females in a group with a male evaluator

[13]The estimated coefficients in Tables A9 and A10 confirm this finding. The estimated coefficient of the fourfold interaction dummy for highly self-confident male evaluators in the Man treatment is positive and highly significant. We can also conclude that self-confidence is an important omitted variable in the regression of performance evaluations by male subjects, as in column 1 of Tables A9 and A10 adding self-confidence improves the fit of the regression from an adjusted R^2 of .03 to now .19. Also, the coefficient of the dummy variables for being in a non-selected treatment becomes significant only when adding self-confidence.

[14]Accordingly, adding self-confidence only slightly increases the adjusted R^2 from .19 to .23. It does not affect the significance of any estimated coefficient.

are assigned a random evaluation drawn from the evaluations of selected females made by male evaluators in the gendered treatments of our experiment. Analogously, evaluations in groups with female evaluators are drawn from the evaluations made by female evaluators in the gendered treatments.

In each group, the group member with the highest evaluation is promoted to the next hierarchy level. The number of female employees at level *s*+1 is determined by the number of females promoted at level *s*. We consider an (approximate) steady state, i.e., we choose the proportion p_s of female evaluators at level s approximately equal to the proportion of females promoted from level s to level *s*+1. This fixed point is determined by a simple iterative algorithm.[15] We close the model by assuming that at the lowest hierarchy level there are equally many female and male employees.

Due to the asymmetry between evaluations of males and selected females in our experiment, this promotion dynamics can be thought of as a model of promotions in a job which is traditionally male-dominated. Recall also that the evaluations we collected are all on ex-ante equally-skilled subjects. Thus, we model promotions of equally-skilled and equally-sized male and female populations in a traditionally male-dominated employment field. In order to minimize the contribution of stochastic fluctuations, we average over *z* runs of these dynamics and choose a number of employees *n* of sufficient size.[16] Considering an approximate steady state is justified by the fact that it is usually reached after about three iterations of the procedure described in Footnote 14.

Results

Table 2 depicts the approximate steady state proportions of females for different values of g and for 6 hierarchy levels and thus 5 promotions.[17]

[15]We start with arbitrary proportions of female evaluators (e.g., no female evaluators) and calculate the number of promoted females. These resulting proportions are used as the new proportions of female evaluators. The procedure is iterated until proportions do not change significantly anymore.
[16]This implies that the actual size of n is irrelevant as long as n and z are sufficiently large. Notably, we could as well consider a promotion pyramid where higher hierarchies are smaller than lower ones. The advantage of equally-sized levels is that computational effort is spread equally across levels.
[17]The further parameters are n=2400 and z=400.

Table 2: Approximate steady states for 6 hierarchy levels

0.500	0.416	0.339	0.273	0.216	0.169	g=2
0.500	0.380	0.277	0.194	0.132	0.090	g=4
0.500	0.353	0.230	0.141	0.082	0.045	g=6
0.500	0.332	0.189	0.095	0.045	0.019	g=8

As could be expected from our experimental results, we see a moderate decrease in the proportion of females from one hierarchy level to the next. These decreases result in a tiny proportion of females after four or five rounds of promotions. Therefore, the discrimination driven effects we observed in our experiment are strong enough quantitatively to explain a glass ceiling effect. For smaller values of *g*, i.e. when each promoted employee is only compared to few opponents, the decrease in the proportion of females becomes smaller in each step. This could be interpreted as the situation in a relatively hierarchic company. Note however that this does not correspond to a better situation for females since each promotion carries less meaning in such a setting. In fact, if we compare, e.g. two rounds of promotions for *g*=4 to one round for *g*=8, we see an even stronger decrease.

We conclude our numerical investigation by comparing our steady state results with two extreme cases, the cases where all promotion decisions are made, respectively, by males or by females. The results for *g*=4 are shown in Table 3.

Table 3: Comparison of steady state results with two extreme cases (g=4)

0.500	0.380	0.277	0.194	0.132	0.090	steady state
0.500	0.340	0.208	0.119	0.066	0.034	all female
0.500	0.403	0.311	0.231	0.166	0.115	all male

While all three cases are qualitatively similar[18], we see that the decrease in the number of female employees when moving up the hierarchy is most pronounced when all promotion decisions are made by females. In contrast, the proportion of females falls considerably slower than in the steady state when all promotion decisions are made by males. This shows that the comparatively strong initial decrease in the proportion of females seen in Table 2 is driven significantly by the promotion decisions of females at intermediate hierarchy levels. This is in line with previous literature investigating the gender composition of evaluation committees (Blau and DeVaro 2007; Bagues and Esteve-Volart 2010). Comparing our simulations to real-life data on women in academia, we find similar patterns of declining fractions over the hierarchy levels.[19]

5 Conclusion

This study identifies problems in updating as a source of discrimination, beyond taste-based and statistical discrimination. Individuals are not able to fully give up their prior belief concerning groups with known average performances in favor of new individual-specific information. Furthermore, male subjects with high self-confidence about their own performance overestimate the performance of other men.

Our simulations explore the consequences of our findings in a hierarchical job structure with a sequence of promotion decisions. We show that the discrimination we observe adds up to a glass ceiling effect, i.e. to the virtual absence of women after few rounds of promotions. In line with empirical observations, this effect is even more pronounced when more promotion decisions are made by women.

Our experiment focuses on a benchmark case where the signal about the performance of a person is perfect. For future research it would be interesting to explore conservatism as a possible source of discrimination in other settings, notably, in the field, where signals are less perfect. Separating discrimination due to updating errors from other types of discrimination in the field would be a probably difficult but highly interesting task. Furthermore, it should be investigated how much better an individual from a less favorable group has to perform in order to receive the same evaluation as another individual with the same abilities.

[18] This qualitative similarity is also a robustness check for our fixed-point procedure.

[19] For German academia in 2004, the Center of Excellence Women and Science reports the following fractions of females: About 50% among graduates, 39% on the PhD level, 34% among junior researchers, 28% on assistant positions, 14% among full professors, with 9% among the best paid full professor positions. http://www.bosch-stiftung.de/content/language1/downloads/Kurzexpertise.pdf

References

Anderson, L., R. Fryer, and C. Holt. 2006. Discrimination: experimental evidence from psychology and economics. In *Handbook on the economics of discrimination*, ed. W. Rogers, 97. Cheltenham, UK: Elgar.

Arrow, K. 1973. The theory of discrimination. In *Discrimination in Labor Markets*, ed. O. Ashenfelter and A. Rees, 3–33. Princeton: Princeton University Press.

Bagues, M., and B. Esteve-Volart. 2010. "Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment." *Review of Economic Studies* 77 (4): 1301–1328.

Bar-Hillel, M. 1980. "The base-rate fallacy in probability judgments." *Acta Psychologica* 44 (3): 211–233.

Bauman, K., and G. Geher. 2002. "We think you agree: The detrimental impact of the false consensus effect on behavior." *Current Psychology* 21 (4): 293–318.

Becker, G. 1957. *The economics of discrimination*. Chicago: University of Chicago Press.

Bertrand, M., and S. Mullainathan. 2004. "Are Emily and Brendan more employable than Latoya and Tyrone? Evidence on racial discrimination in the labor market from a large randomized experiment." *American Economic Review* 94 (4): 991–1013.

Blau, F., and J. DeVaro. 2007. "New evidence on gender differences in promotion rates: An empirical analysis of a sample of new hires." *Industrial Relations: A Journal of Economy and Society* 46 (3): 511–550.

Charness, G., and D. Levin. 2005. "When optimal choices feel wrong: A laboratory study of Bayesian updating, complexity, and affect." *American Economic Review* 95 (4): 1300–1309.

Crump, R., V. Hotz, G. Imbens, and O. Mitnik. 2008. "Nonparametric tests for treatment effect heterogeneity." *Review of Economics and Statistics* 15 (3): 389–405.

Dohmen, T., A. Falk, D. Huffman, F. Marklein, and U. Sunde. 2009. "Biased probability judgment: Evidence of incidence and relationship to economic outcomes from a representative sample." *Journal of Economic Behavior & Organization* 72 (3): 903–915.

El-Gamal, M., and D. Grether. 1995. “Are people Bayesian? Uncovering behavioral strategies.” *Journal of the American Statistical Association* 90 (432): 1137–1145.

Falk, A., D. Huffman, and U. Sunde. 2006. “Self-confidence and search.” *Institute for the Study of Labor (IZA); University of St. Gallen* Working Paper.

Fischbacher, U. 2007. “z-Tree: Zurich toolbox for ready-made economic experiments.” *Experimental Economics* 10 (2): 171–178.

Greiner, B. 2003. An online recruitment system for economic experiments. In *Forschung und wissenschaftliches Rechnen*, ed. K. Kremer and V. Macho, 79-93. GWDG Bericht 63. Göttingen: Ges. für Wiss. Datenverarbeitung.

Grether, D. 1992. “Testing Bayes rule and the representativeness heuristic: Some experimental evidence.” *Journal of Economic Behavior & Organization* 17 (1): 31–57.

Kahneman, D., and A. Tversky. 1972. “Subjective probability: A judgment of representativeness.” *Cognitive Psychology* 3 (3): 430–454.

Lyon, D., and P. Slovic. 1976. “Dominance of accuracy information and neglect of base rates in probability estimation.” *Acta Psychologica* 40 (4): 287–298.

Möbius, M., M. Niederle, P. Niehaus, and T. Rosenblat. 2011. *Managing Self-Confidence: Theory and Experimental Evidence*. National Bureau of Economic Research.

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

Peters, M., B. Laeng, K. Latham, M. Jackson, R. Zaiyouna, and C. Richardson. 1995. “A redrawn Vandenberg and Kuse mental rotations test-different versions and factors that affect performance.” *Brain and Cognition* 28 (1): 39–58.

Phelps, E. 1972. “The statistical theory of racism and sexism.” *American Economic Review* 62 (4): 659–661.

Reuben, E., P. Sapienza, and L. Zingales. 2010. “The glass ceiling in experimental markets” Working Paper.

Ross, L., D. Greene, and P. House. 1977. “The ‘false consensus effect’: An egocentric bias in social perception and attribution processes.” *Journal of Experimental Social Psychology* 13 (3): 279-301.

Tversky, A., and D. Kahneman. 1971. “Belief in the law of small numbers.” *Psychological Bulletin* 76 (2): 105-110.

Vandenberg, S., and A. Kuse. 1978. “Mental rotation, a group test of three-dimensional spatial visualization.” *Perceptual and Motor Skills* 47 (2): 599-604.

Appendix 1

INSTRUCTIONS[20]

SAMPLE INSTRUCTIONS FOR GENDER TREATMENT ("SELECTED-WOMAN"):

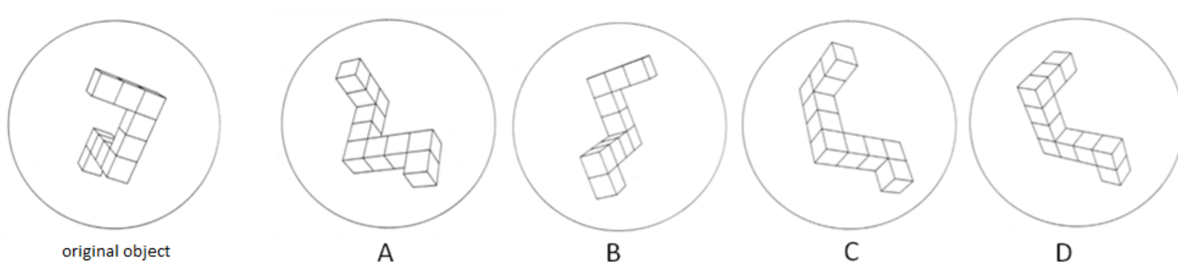
Welcome to our study. Please read the following instructions carefully.

For participating in this study you will receive 4 EUR for sure. Depending on you and another participant's performance you can earn money in addition to these 4 EUR. In this study you are anonymous and all data that you provide will be treated confidentially. If you have any questions after reading the following instructions, please raise your hand and we will come to answer your question. Please do not talk to other participants during the study – we would have to exclude you from this study then.

The study consists of two stages: **You are a second stage participant**. During the study, you will be randomly assigned to a participant who participated in the first stage.

Stage 1

This stage has already been completed by other participants. Those participants solved a number of mental rotation tasks. Here is an example for this task:



In this stage subjects had to distinguish between the two figures among A, B, C, and D that can be transferred into the original object on the left side by rotation (in our example figures B and D). The two other figures (in our example figures A and C) that cannot be transferred into the original object by rotation only, but had to be mirrored. Subjects were supposed to cross out the two figures that were rotations only. If they crossed out both correct figures, the task was solved correctly. Subjects were given 24 of these tasks, and 6 minutes to solve as many of them as possible.

Stage 2

In this stage **<u>you</u>** are the sole decision maker. You have the opportunity to earn money depending on the stage 1 performance of a first stage participant you have been randomly matched with. The payment is for you only, the first stage participant was paid for his participation already.

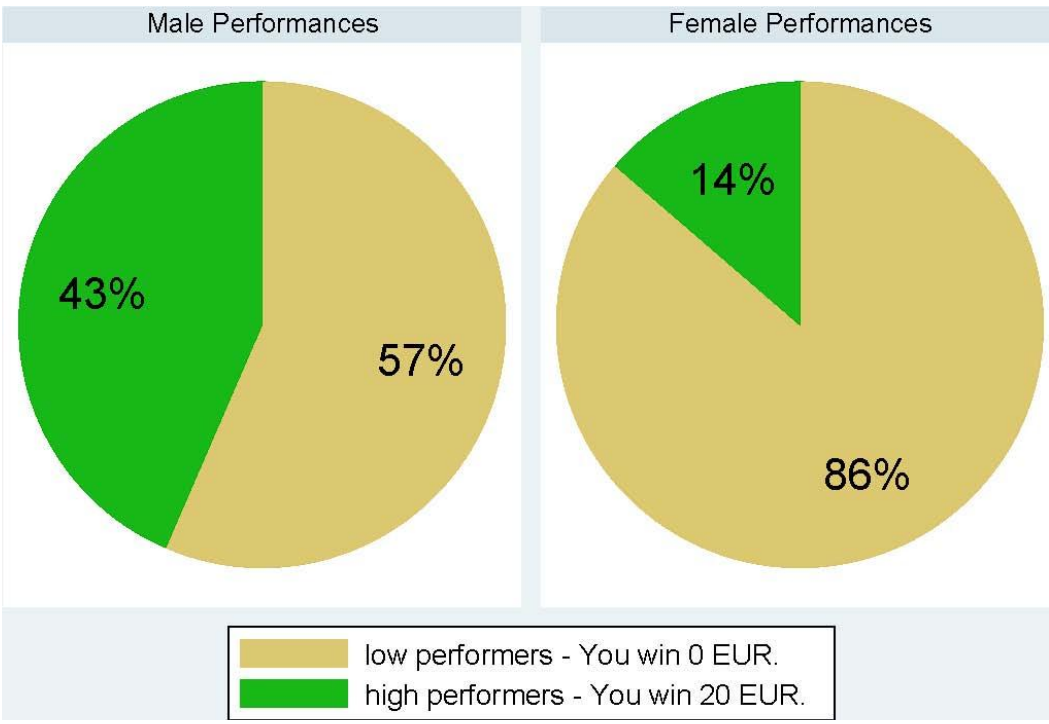
[20] The following are an English version of instructions for the treatments "Selected-Woman" and "Neutral". The original German versions are available from the authors upon request.

In the following, we only regarded participants of the first stage who solved a minimum number of tasks correctly. participants were divided into two groups, males and females.

- 43% of the male participants are top perfomers. 57% are mediocre performers.
- 14% of the female participants are top perfomers. 86% are mediocre performers.

Top performers are participants who solved 14 or more tasks correctly, whereas mediocre performers are participants who solved at least 9 tasks correctly, but not more than 13 tasks.

Below we present the distributions of the two groups in the form of a diagram.



Please answer the following control questions:

1. Imagine that there were 100 male participants in Stage 1. What is the number of male top performers?_____ participants are top performers.

2. Imagine that there were 100 female participants in Stage 1. What is the number of female top performers?_____ participants are top performers.

3. Imagine that there were 100 male participants in Stage 1. What is the number of male mediocre performers?_____ participants are mediocre performers.

4. Imagine that there were 100 female participants in Stage 1. What is the number of female mediocre performers?_____ participants are mediocre performers.

The selection of the female and male participants:

Please read the following paragraph carefully. It is important that you understand the selection process.

The **selection** of the first stage participants was as follows:

We randomly select one male participant. We call him participant M henceforth. Then we will select a female participant F as follows:

- If participant M was a top performer, we select a female participant F who also was a top performer.
- If participant M was a mediocre performer, we choose a female participant F who also was a mediocre performer.

You will get matched either to the male particioant M or the female participant F.

Later, you will choose between a fixed reward and a lottery:

- You receive 20 EUR if the participant you are matched with is a top performer
- You receive 0 EURif the participant is a mediocre performer

Please answer the following control questions:

Imagine that one male and one female participant are selected as it is described above. Please indicate by putting an X which alternative you think is correct in the following two situations.

1. If the male participant participant is a top performer, then the female participant is a

___ top performer ____ mediocre performer ___ could be either or

2. If the male participant is a mediocre performer, then the female participant is a

___ top performer ____ mediocre performer ___ could be either or

Please insert the correct answer in the following two situations.

1. If the person you got matched to is a top performer, you receive _____ EUR.

2. If the person you got matched to is a mediocre performer, you receive _____ EUR .

Decision

We now ask you to make a decision for each of the following options between getting a fixed amount of money(from 0.40 EUR going up to 20 EUR), and playing the aforementioned lottery.

At the end, one of your decisions will be randomly drawn and determine your final payoff.

Mark your answers by putting an X at the alternative you choose for each of the questions 1 to 50.

1. When a decision will be drawn in which you chose the fixed reward, you will receive this reward.
2. When a decision will be drawn in which you chose the lottery, you will receive 0 or 20 EUR, depending on the performance of your matched first stage participant.
3. If you do not put an X in the decision that was drawn, you will receive 0 EUR.

According to selection process described above, you have been matched with a female participant.

Making these decisions we ask you to take your time to think about your decisions and to take them seriously. **Also, remember that you will be paid according to one of these decisions, which is randomly drawn after the study ends.**

1) Which alternative do you choose:

☐ 0.40 EUR for sure ☐ 0 or 20 EUR depending on the performance of your female participant

2) Which alternative do you choose:

☐ 0.80 EUR for sure ☐ 0 or 20 EUR depending on the performance of your female participant

3) Which alternative do you choose:

☐ 1.20 EUR for sure ☐ 0 or 20 EUR depending on the performance of your female participant

…..

50) Which alternative do you choose:

☐ 20.00 EUR for sure ☐ 0 or 20 EUR depending on the performance of your female participant

SAMPLE INSTRUCTIONS FOR NEUTRAL TREATMENT ("NEUTRAL"):

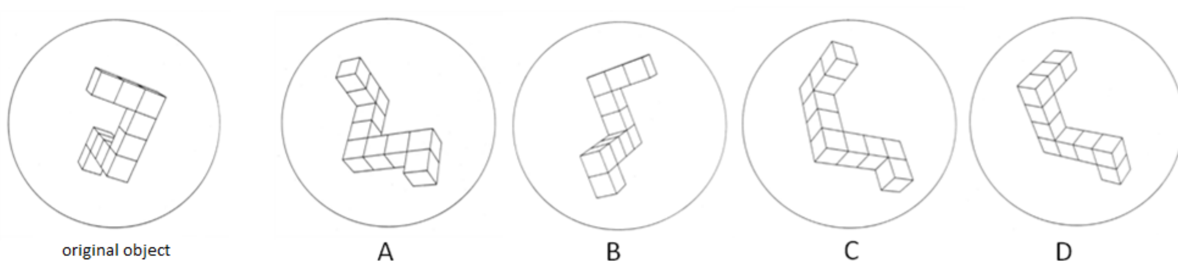
Welcome to our study. Please read the following instructions carefully.

For participating in this study you will receive 4 EUR for sure. Depending on you and another participant's performance you can earn money in addition to these 4 EUR. In this study you are anonymous and all data that you provide will be treated confidentially. If you have any questions after reading the following instructions, please raise your hand and we will come to answer your question. Please do not talk to other participants during the study – we would have to exclude you from this study then.

The study consists of two stages: **You are a second stage participant**. During the study, you will be randomly assigned to a participant who participated in the first stage.

Stage 1

This stage has already been completed by other participants. Those participants solved a number of mental rotation tasks. Here is an example for this task:



In this stage subjects had to distinguish between the two figures among A, B, C, and D that can be transferred into the original object on the left side by rotation (in our example figures B and D). The two other figures (in our example figures A and C) that cannot be transferred into the original object by rotation only, but had to be mirrored. Subjects were supposed to cross out the two figures that were rotations only. If they crossed out both correct figures, the task was solved correctly. Subjects were given 24 of these tasks, and 6 minutes to solve as many of them as possible.

Stage 2

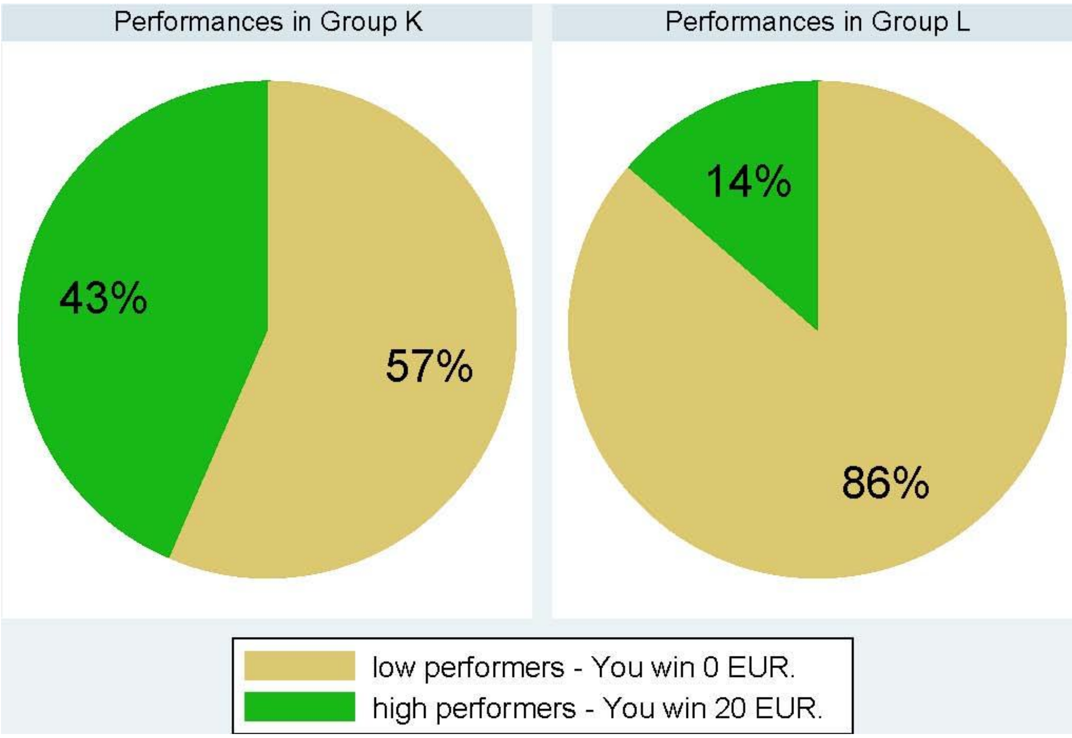
In this stage **<u>you</u>** are the sole decision maker. You have the opportunity to earn money depending on the stage 1 performance of a first stage participant you have been randomly matched with. The payment is for you only, the first stage participant was paid for his participation already.

In the following, we only regarded participants of the first stage who solved a minimum number of tasks correctly. participants were divided into two groups, K and L.

- 43% participants of group K are top perfomers. 57% are mediocre performers.
- 14% participants of group L are top perfomers. 86% are mediocre performers.

Top performers are participants who solved 14 or more tasks correctly, whereas mediocre performers are participants who solved at least 9 tasks correctly, but not more than 13 tasks

Below we present the distributions of the two groups in the form of a diagram.



Please answer the following control questions:

1. Imagine that there were 100 participants in group K in Stage 1. What is the number of group K top performers?_____ participants are top performers.

2. Imagine that there were 100 participants in group L in Stage 1. What is the number of group L top performers?_____ participants are top performers.

3. Imagine that there were 100 participants in group K in Stage 1. What is the number of group K mediocre performers?_____ participants are mediocre performers.

4. Imagine that there were 100 participants in group L in Stage 1. What is the number of group L mediocre performers?_____ participants are mediocre performers.

The selection of the participants from group K and L:

Please read the following paragraph carefully. It is important that you understand the selection process.

The **selection** of the first stage participants was as follows:

We randomly select one participant from group K. We call him participant k henceforth. Then we will select a participant L as follows:

- If participant K was a top performer, we select a participant L who also was a top performer.
- If participant K was a mediocre performer, we choose a participant L who also was a mediocre performer.

You will get matched either to the male participant K or participant L.

Later, you will choose between a fixed reward and a lottery:

- You receive 20 EUR if the participant you are matched with is a top performer
- You receive 0 EUR if the participant is a mediocre performer

Please answer the following control questions:

Imagine that one participant K and one participant L are selected as it is described above. Please indicate by putting an X which alternative you think is correct in the following two situations.

1. If participant K is a top performer, then participant L is a

___ top performer ____ mediocre performer ___ could be either or

2. If participant K is a mediocre performer, then participant L is a

___ top performer ____ mediocre performer ___ could be either or

Please insert the correct answer in the following two situations.

1. If the person you got matched to is a top performer, you receive _____ EUR.

2. If the person you got matched to is a mediocre performer, you receive _____ EUR.

Decision

We now ask you to make a decision for each of the following options between getting a fixed amount of money (from 0.40 EUR going up to 20 EUR), and playing the aforementioned lottery.

At the end, one of your decisions will be randomly drawn and determine your final payoff.

Mark your answers by putting an X at the alternative you choose for each of the questions 1 to 50.

4. When a decision will be drawn in which you chose the fixed reward, you will receive this reward.

5. When a decision will be drawn in which you chose the lottery, you will receive 0 or 20 EUR, depending on the performance of your matched first stage participant.
6. If you do not put an X in the decision that was drawn, you will receive 0 EUR.

According to selection process described above, you have been matched with a participant from group K.

Making these decisions we ask you to take your time to think about your decisions and to take them seriously. **Also, remember that you will be paid according to one of these decisions, which is randomly drawn after the study ends.**

1) Which alternative do you choose:

☐ 0.40 EUR for sure ☐ 0 or 20 EUR depending on the performance of your participant

2) Which alternative do you choose:

☐ 0.80 EUR for sure ☐ 0 or 20 EUR depending on the performance of your participant

3) Which alternative do you choose:

☐ 1.20 EUR for sure ☐ 0 or 20 EUR depending on the performance of your participant

…..

50) Which alternative do you choose:

☐ 20.00 EUR for sure ☐ 0 or 20 EUR depending on the performance of your participant

Appendix 2

TABLES

Table A1: Descriptive Statistics for Evaluators

Subjects	All		Male		Female		Gender
	Mean	Median	Mean	Median	Mean	Median	difference
Variable	(Std.dev.)	Obs.	(Std.dev.)	Obs.	(Std.dev.)	Obs.	in means
First switching point	6.84	6.8	7.28	7.60	6.35	5.60	0.93**
in EUR [a)]	(3.18)	272	(3.14)	144	(3.16)	128	[0.02]
Av. switiching point	6.98	6.8	7.39	7.60	6.57	6.00	0.82**
in EUR	(3.18)	305	(3.17)	152	(3.14)	153	[0.02]
Multiple switcher	0.11	0	0.05	0	0.16	0	-0.11**
(dummy)	(0.31)	305	(0.22)	152	(0.37)	153	[0.00]
Belief:	14.00	14.0	15.36	16	12.64	12	2.72**
Own score	(4.32)	305	(4.17)	152	(4.05)	153	[0.00]
Belief:	13.17	14.0	13.36	14	12.99	14	0.37
Participant's score	(3.11)	305	(3.25)	152	(2.97)	153	[0.30]
Diff. in beliefs:	0.83	0.0	2.01	2	-0.35	0	2.36**
Own - participant	(4.24)	305	(3.80)	152	(4.34)	153	[0.00]
Belief:	14.16	14.0	13.92	13	14.41	14	-0.49
Av.male score	(2.92)	305	(2.99)	152	(2.83)	153	[0.15]
Belief:	10.87	10.0	10.97	11	10.77	10	0.20
Av.female score	(2.87)	305	(2.99)	152	(2.74)	153	[0.55]
Task usefulness	6.52	7.0	6.64	7	6.41	7	0.23
(1 = low to 10 = high)	(2.33)	305	(2.35)	152	(2.31)	153	[0.37]
Age	24.70	24.0	25.36	25	24.04	24	1.32**
	(4.59)	305	(5.27)	152	(3.69)	153	[0.01]

** (*): Difference is significant on the 5 (10) percent level (two-sided t test).

a: When considering the first switching point, subjects with more than one switching point are excluded in all tables.

Table A2: Performance Evaluations

Treatment		First switching point in EUR			Average switching point in EUR		
	Subjects	**All**	**Male**	**Female**	**All**	**Male**	**Female**
Man	Mean	7.97	8.37	7.49	7.86	8.30	7.42
	(Std.err.)	(0.46)	(0.70)	(0.57)	(0.41)	(0.67)	(0.49)
	Median	7.6	8.0	6.8	7.6	8.0	7.2
	Obs.	42	23	19	48	24	24
Selected woman	Mean	6.13	6.82	5.32	6.32	7.11	5.47
	(Std.err.)	(0.33)	(0.43)	(0.47)	(0.32)	(0.46)	(0.43)
	Median	5.6	7.2	4.8	5.6	7.6	4.8
	Obs.	85	46	39	94	49	45
Difference in means [a)]		1.84^{**}[.00]	1.55^{*}[.05]	2.17^{**}[.01]	1.53^{**}[.01]	1.19 [.14]	1.95^{**}[.01]
Zero ATE [b)]			$\chi^2(2) = 12.16$ [.00]**			$\chi^2(2) = 11.16$ [.00]**	
Constant ATE [b)]			$\chi^2(1) = 0.32$ [.57]			$\chi^2(1) = 0.53$ [.46]	
Neutral	Mean	8.32	8.07	8.55	8.29	8.04	8.51
	(Std.err.)	(0.36)	(0.39)	(0.59)	(0.34)	(0.40)	(0.53)
	Median	8.2	7.8	8.4	8.2	7.8	8.4
	Obs.	62	30	32	70	32	38
Neutral selected	Mean	6.23	7.02	5.26	6.51	7.07	5.93
	(Std.err.)	(0.43)	(0.66)	(0.44)	(0.41)	(0.64)	(0.49)
	Median	5.6	6.8	5.0	6.0	6.8	5.6
	Obs.	62	34	28	69	35	34
Difference in means [a)]		2.09^{**}[.00]	1.05 [.20]	3.29^{**}[.00]	1.78^{**}[.00]	0.97 [.22]	2.58^{**}[.00]
Zero ATE [b)]			$\chi^2(2) = 22.07$ [.00]**			$\chi^2(2) = 14.46$ [.00]**	
Constant ATE [b)]			$\chi^2(1) = 4.48$ [.03]**			$\chi^2(1) = 2.39$ [.12]	

** (*): Difference is significant on the 5 (10) percent level. p-values in brackets.
a: Two-sided t test.
b: Tests for treatment effect heterogeneity as in Crump, Hotz, Imbens, and Mitnik (2008). The first (second) is testing whether facing a selected person has a zero (an identical) average effect for male and female subjects.

Table A3: Regression Analysis of Performance Evaluations by Gender

	Switching Point		
Subjects	**Male**	**Female**	**All**
Non-select TM * Gender TM * Male			1.61 (1.52)
Non-select TM * Male			-2.24** (1.02)
Gender TM * Male			-.26 (.99)
Non-select TM * Gender TM	.33 (1.10)	-1.04 (1.03)	-1.12 (1.07)
Male			1.75** (.74)
Non-select TM	1.16** (.71)	3.24** (.72)	3.30** (.71)
Gender TM	-.11 (.76)	.05 (.61)	.07 (.65)
Age	.07 (.12)	-.16** (.07)	.007 (.09)
Constant	5.13* (2.97)	9.02** (1.73)	5.09** (2.26)
Obs.	133	118	251
R squared	.06	.25	.14

Estimated coefficients from OLS regressions with bootstrapped standard errors in parentheses (1000 replications). Non-select TM is a dummy variable equal to one for the treatments *Man* and *Neutral*, and zero otherwise. Gender TM is a dummy that equals one for the gendered treatments *Man* and *Selected-Woman*.

Table A4: Regression Analysis of Performance Evaluations by Gender

	Average Switching Point		
Subjects:	**Male**	**Female**	**All**
Non-select TM * Gender TM * Male			.86 (1.47)
Non-select TM * Male			-1.62 (1.00)
Gender TM * Male			.49 (.99)
Non-select TM * Gender TM	.05 (1.03)	-.64 (.92)	-.63 (.95)
Male			1.15 (.71)
Non-select TM	1.07 (.67)	2.61** (.68)	2.58** (.70)
Gender TM	.14 (.73)	-.40 (.61)	-.46 (.66)
Age	.07 (.12)	-.16** (.06)	-.004 (.09)
Constant	5.33* (2.96)	9.80** (1.54)	6.02** (2.17)
Obs.	140	141	281
R squared	.04	.20	.11

Estimated coefficients from OLS regressions with bootstrapped standard errors in parentheses (1000 replications). Non-select TM is a dummy variable equal to one for the treatments *Man* and *Neutral*, and zero otherwise. Gender TM is a dummy that equals one for the gendered treatments *Man* and *Selected-Woman*.

Table A5: Performance Evaluations by Relative Level of Self-Confidence

	Self-confidence[c)]	First switching point in EUR: Male subjects High	Male subjects Low	Female subjects High	Female subjects Low	Average switching point in EUR: Male subjects High	Male subjects Low	Female subjects High	Female subjects Low
Man	Mean	10.56	5.80	7.75	7.65	10.56	5.91	7.38	7.71
	(Std.err.)	(0.94)	(0.68)	(0.70)	(1.17)	(0.94)	(0.61)	(0.72)	(0.88)
	Median	9.6	5.6	8.4	6.8	9.6	5.6	8.0	6.8
	Obs.	10	8	8	8	10	9	9	11
Selected-Woman	Mean	7.27	6.08	5.72	5.18	7.27	6.38	6.13	5.18
	(Std.err.)	(0.60)	(0.68)	(0.78)	(0.76)	(0.60)	(0.65)	(0.73)	(0.65)
	Median	7.8	5.6	4.8	4.4	7.8	6.4	5.6	4.2
	Obs.	18	20	13	18	18	22	15	22
Difference in means [a)]		3.29**	-0.28	2.03*	2.47*	3.29**	-0.47	1.25	2.53**
	[p-value]	[.00]	[.81]	[.09]	[.08]	[.00]	[.67]	[.27]	[.03]
Zero ATE [b)]		$\chi^2(2)$=8.69[.01]**		$\chi^2(2)$=6.95[.03]**		$\chi^2(2)$=8.93 [.01]**		$\chi^2(2)$=6.83[.03]**	
Constant ATE [b)]		$\chi^2(1)$=5.84[.02]**		$\chi^2(1)$=0.07[.80]		$\chi^2(1)$=6.90 [.01]**		$\chi^2(1)$=0.73[.39]	
Neutral	Mean	7.65	8.62	9.6	8.24	7.65	8.80	9.46	8.31
	(Std.err.)	(0.55)	(0.63)	(0.94)	(0.73)	(0.55)	(0.60)	(0.82)	(0.70)
	Median	7.6	8.4	9.6	8.4	7.6	8.4	9.6	8.4
	Obs.	15	11	12	15	15	12	14	18
Selected-Neutral	Mean	7.02	6.99	6.08	4.77	7.02	7.10	6.62	5.34
	(Std.err.)	(0.42)	(1.16)	(0.65)	(0.76)	(0.42)	(1.09)	(0.75)	(0.79)
	Median	7.6	7.6	6.4	4.2	7.6	7.6	6.8	4.6
	Obs.	11	15	10	12	11	16	13	14
Difference in means [a)]		0.63	1.63	3.52**	3.47**	0.63	1.70	2.84**	2.97**
	[p-value]	[.40]	[.28]	[.01]	[.00]	[.40]	[.22]	[.02]	[.01]
Zero ATE [b)]		$\chi^2(2)$=2.37 [.31]		$\chi^2(2)$=20.3[.00]**		$\chi^2(2)$=2.70 [.26]		$\chi^2(2)$=14.5[.00]**	
Constant ATE [b)]		$\chi^2(1)$=0.45 [.50]		$\chi^2(1)$=0.00[.98]		$\chi^2(1)$=0.56 [.46]		$\chi^2(1)$=0.01[.93]	

** (*): Difference is significant on the 5 (10) percent level. p-values in brackets.
a: Two-sided t test.
b: Tests for treatment effect heterogeneity as in Crump et al. (2008). The first (second) is testing whether facing a selected person has a zero (an identical) average effect for highly and less self-confident subjects.
c: Self-confidence is high (low) if the belief about own score is above (below) the median belief by gender.

Table A6: Performance Evaluations by Relative Level of Self-Confidence (Robustness Check)

		First switching point in EUR				Average switching point in EUR			
		Male subjects		**Female subjects**		**Male subjects**		**Female subjects**	
	Self-confi-dence [c)]	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**
Man	Mean	10.56	6.68	7.38	7.65	10.56	6.69	7.17	7.71
	(Std.err.)	(0.94)	(0.72)	(0.57)	(1.17)	(0.94)	(0.66)	(0.53)	(0.88)
	Median	9.6	6.0	7.6	6.8	9.6	6.4	7.6	6.8
	Obs.	10	13	11	8	10	14	13	11
Selected-Woman	Mean	7.38	6.08	5.45	5.18	7.7	6.38	5.74	5.18
	(Std.err.)	(0.55)	(0.68)	(0.60)	(0.76)	(0.62)	(0.65)	(0.59)	(0.65)
	Median	8.0	5.6	4.8	4.4	8.0	6.4	4.8	4.2
	Obs.	26	20	21	18	27	22	23	22
Difference in means [a)]		3.18**	0.6	1.93**	2.47*	2.86**	0.31	1.43	2.53**
	[p-value]	[.01]	[.56]	[.05]	[.08]	[.02]	[.76]	[.11]	[.03]
Zero ATE [b)]		$\chi^2(2)$=8.89[.01]**		$\chi^2(2)$=8.65 [.01]**		$\chi^2(2)$=6.59 [.04]**		$\chi^2(2)$=8.68 [.01]**	
Constant ATE [b)]		$\chi^2(1)$=3.08[.08]*		$\chi^2(1)$=0.11 [.74]		$\chi^2(1)$=3.06 [.08]*		$\chi^2(1)$=0.67 [.41]	
Neutral	Mean	7.65	8.48	8.82	8.24	7.65	8.38	8.68	8.31
	(Std.err.)	(0.55)	(0.57)	(0.91)	(0.73)	(0.55)	(0.57)	(0.79)	(0.70)
	Median	7.6	8.4	9.6	8.4	7.6	8.4	9.0	8.4
	Obs.	15	15	17	15	15	17	20	18
Selected-Neutral	Mean	7.03	7.02	6.08	4.80	7.03	7.14	6.62	5.50
	(Std.err.)	(0.73)	(1.31)	(0.65)	(0.56)	(0.73)	(1.22)	(0.75)	(0.64)
	Median	6.0	7.6	6.4	4.6	6.0	7.6	6.8	4.8
	Obs.	21	13	10	18	21	14	13	21
Difference in means [a)]		0.62	1.46	2.74**	3.44**	0.62	1.24	2.06*	2.81**
	[p-value]	[.53]	[.29]	[.04]	[.00]	[.53]	[.34]	[.08]	[.01]
Zero ATE [b)]		$\chi^2(2)$=1.54 [.46]		$\chi^2(2)$=19.9[.00]**		$\chi^2(2)$=1.31 [.52]		$\chi^2(2)$=12.3[.00]**	
Constant ATE [b)]		$\chi^2(1)$=0.25 [.62]		$\chi^2(1)$=0.23[.63]		$\chi^2(1)$=0.14 [.71]		$\chi^2(1)$=0.26[.61]	

** (*): Difference is significant on the 5 (10) percent level. p-values in brackets.
a: Two-sided t test.
b: Tests for treatment effect heterogeneity as in Crump et al. (2008). The first (second) is testing whether facing a selected person has a zero (an identical) average effect for highly and less self-confident subjects.
c: Self-confidence is high (low) if the belief about own score is above (below) the mean belief by gender.

Table A7: Performance Evaluations by Beliefs About Own Relative Performance (Robustness Check)

		First switching point in EUR				Average switching point in EUR			
		Male subjects		**Female subjects**		**Male subjects**		**Female subjects**	
	Self-confidence [c)]	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**
Man	Mean	10.28	6.89	7.69	7.32	10.28	6.89	7.5	7.33
	(Std.err.)	(1.00)	(0.76)	(0.61)	(0.97)	(1.00)	(0.71)	(0.54)	(0.83)
	Median	9.2	6.0	8.0	6.8	9.2	6.4	7.8	6.8
	Obs.	10	13	9	10	10	14	12	12
Selected-Woman	Mean	6.51	7.18	4.89	5.70	6.88	7.37	5.03	5.78
	(Std.err.)	(0.53)	(0.73)	(0.69)	(0.64)	(0.63)	(0.68)	(0.67)	(0.57)
	Median	6	8.0	3.8	4.8	6.4	8.0	4.0	4.8
	Obs.	25	21	18	21	26	23	19	26
Difference in means [a)]		3.77**	-0.29	2.80**	1.62	3.40**	-0.48	2.47**	1.55
	[p-value]	[.00]	[.79]	[.02]	[.17]	[.01]	[.64]	[.01]	[.13]
Zero ATE [b)]		$\chi^2(2)$=11.3[.00]**		$\chi^2(2)$=11.2[.00]**		$\chi^2(2)$=8.68 [.01]**		$\chi^2(2)$=10.6[.00]**	
Constant ATE [b)]		$\chi^2(1)$=6.92[.01]**		$\chi^2(1)$=0.62[.43]		$\chi^2(1)$=6.49 [.01]**		$\chi^2(1)$=0.48[.49]	
Neutral	Mean	7.40	8.65	9.67	7.11	7.20	8.78	9.41	7.39
	(Std.err.)	(0.61)	(0.48)	(0.66)	(0.93)	(0.60)	(0.47)	(0.60)	(0.86)
	Median	7.6	8.2	9.6	6.4	7.6	8.4	9.6	6.8
	Obs.	14	16	18	14	15	17	21	17
Selected-Neutral	Mean	6.02	8.02	5.06	5.46	6.02	8.07	5.65	6.21
	(Std.err.)	(0.51)	(1.19)	(0.62)	(0.64)	(0.51)	(1.12)	(0.70)	(0.70)
	Median	5.6	7.6	5.4	4.8	5.6	7.6	5.6	5.6
	Obs.	17	17	14	14	17	18	17	17
Difference in means [a)]		1.38*	0.63	4.61**	1.65	1.18	0.71	3.76**	1.18
	[p-value]	[.09]	[.64]	[.00]	[.15]	[.15]	[.57]	[.00]	[.30]
Zero ATE [b)]		$\chi^2(2)$=3.21 [.20]		$\chi^2(2)$=27.9[.00]**		$\chi^2(2)$=2.55 [.28]		$\chi^2(2)$=17.8[.00]**	
Constant ATE [b)]		$\chi^2(1)$=0.25 [.62]		$\chi^2(1)$=4.19[.04]**		$\chi^2(1)$=0.10 [.75]		$\chi^2(1)$=3.23[.07]*	

** (*): Difference is significant on the 5 (10) percent level. p-values in brackets.
a: Two-sided t test.
b: Tests for treatment effect heterogeneity as in Crump et al. (2008). The first (second) is testing whether facing a selected person has a zero (an identical) average effect for highly and less self-confident subjects.
c: Self-confidence classified as high (low) if beliefs about own minus the participant's score is above (below) the mean.

Table A8: Performance Evaluations by Absolute Self-Confidence (Robustness Check)

		First switching point in EUR				Average switching point in EUR			
		Male subjects		**Female subjects**		**Male subjects**		**Female subjects**	
	Self-confidence [c)]	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**
Man	Mean	9.30	6.23	7.75	7.31	9.30	6.30	7.38	7.44
	(Std.err.)	(0.87)	(0.62)	(0.70)	(0.87)	(0.87)	(0.54)	(0.72)	(0.66)
	Median	9.2	5.6	8.4	6.8	9.2	6.2	8.0	6.8
	Obs.	16	7	8	11	16	8	9	15
Selected-Woman	Mean	7.05	6.29	5.45	5.18	7.32	6.68	5.74	5.18
	(Std.err.)	(0.48)	(0.93)	(0.60)	(0.76)	(0.54)	(0.85)	(0.59)	(0.65)
	Median	7.6	5.8	4.8	4.4	7.6	7.8	4.8	4.2
	Obs.	32	14	21	18	33	16	23	22
Difference in means [a)]		3.29**	-0.28	2.03**	2.47*	3.29**	-0.47	1.25	2.53**
	[p-value]	[.02]	[.97]	[.04]	[.08]	[.05]	[.77]	[.12]	[.02]
Zero ATE [b)]		$\chi^2(2)$=4.96[.08]*		$\chi^2(2)$=9.71[.01]**		$\chi^2(2)$=3.77 [.15]		$\chi^2(2)$=9.06[.01]**	
Constant ATE [b)]		$\chi^2(1)$=2.40[.12]		$\chi^2(1)$=0.01[.91]		$\chi^2(1)$=2.68 [.10]*		$\chi^2(1)$=0.22[.64]	
Neutral	Mean	8.13	7.80	8.82	8.24	7.98	8.23	8.68	8.31
	(Std.err.)	(0.49)	(0.31)	(0.91)	(0.73)	(0.49)	(0.50)	(0.79)	(0.70)
	Median	7.8	7.8	9.6	8.4	7.6	8.0	9.0	8.4
	Obs.	24	6	17	15	25	7	20	18
Selected-Neutral	Mean	7.06	6.90	6.08	4.80	7.06	7.11	6.62	5.50
	(Std.err.)	(0.59)	(2.16)	(0.65)	(0.56)	(0.59)	(1.92)	(0.75)	(0.64)
	Median	6.8	5.6	6.4	4.6	6.8	7.6	6.8	4.8
	Obs.	26	8	10	18	26	9	13	21
Difference in means [a)]		0.63	1.63	3.52**	3.47**	0.63	1.70	2.84*	2.97**
	[p-value]	[.17]	[.73]	[.04]	[.00]	[.24]	[.63]	[.08]	[.01]
Zero ATE [b)]		$\chi^2(2)$=2.08 [.35]		$\chi^2(2)$=19.92[.00]**		$\chi^2(2)$=1.74 [.42]		$\chi^2(2)$=12.31[.00]**	
Constant ATE [b)]		$\chi^2(1)$=0.01 [.94]		$\chi^2(1)$=0.23[.63]		$\chi^2(1)$=0.01 [.93]		$\chi^2(1)$=0.26[.61]	

** (*): Difference is significant on the 5 (10) percent level. p-values in brackets.

a: Two-sided t test.

b: Tests for treatment effect heterogeneity as in Crump et al. (2008). The first (second) is testing whether facing a selected person has a zero (an identical) average effect for highly and less self-confident subjects.

c: Self-confidence is high (low) if the subjects believes he/she is a top performer himself/herself.

Table A9: Regression Analyses of Performance Evaluations by Relative Level of Self-Confidence for Switching Point (One-Time Switchers Only)

	Switching Point		
Subjects:	**Male**	**Female**	**All**
Non-select TM * Gender TM * Self-confident * Male			4.06**
			(1.62)
Non-select TM * Self-confident * Male			-.43
			(1.86)
Gender TM * Self-confident * Male			2.33
			(1.78)
Non-select TM * Male			-2.30*
			(1.21)
Gender TM * Male			-1.63
			(1.20)
Self-confident * Male			-1.49
			(1.45)
Non-select TM * Gender TM	-2.56*	-.64	-1.62*
	(1.40) (	1.69)	(.90)
Non-select TM * Self-confident	-1.27	.77	-.42
	(1.29)	(1.63)	(1.22)
Gender TM * Self-confident	.84	-.30	-1.11
	(1.35)	(1.44)	(1.18)
Male			2.28**
			(1.04)
Non-select TM	2.14**	3.02**	3.87**
	(.99)	(1.06)	(.95)
Gender TM	-.33	.03	.76
	(1.04)	(1.03)	(.95)
Self-confident	.41	.93	1.56
	(.98)	(1.05)	(.97)
Age	.17	-.17*	.07
	(.15)	(.10)	(.13)
Constant	2.28	8.95**	2.93
	(3.69)	(2.53)	(3.25)
Obs.	108	96	204
R squared	.25	.29	.24

Estimated coefficients from OLS regressions with bootstrapped standard errors in parentheses (1000 replications). Non-select TM is a dummy variable equal to one for the treatments *Man* and *Neutral*, and zero otherwise. Gender TM is a dummy that equals one for the gendered treatments *Man* and *Selected-Woman*. Self-confidence is a dummy equal to one if the belief about the own MRT score is above the median belief by gender.

Table A10: Regression Analysis of Performance Evaluations by Relative Level of Self-Confidence for Average Switching Point

	Average Switching Point		
Subjects:	**Male**	**Female**	**All**
Non-select TM * Gender TM * Self-confident * Male			3.95**
			(1.59)
Non-select TM * Self-confident * Male			.13
			(1.66)
Gender TM * Self-confident * Male			2.08
			(1.60)
Non-select TM * Male			-2.12**
			(1.07)
Gender TM * Male			-1.17
			(1.06)
Self-confident * Male			-1.79
			(1.31)
Non-select TM * Gender TM	-2.83**	-.10	-1.47*
	(1.32) (	1.46)	(.84)
Non-select TM * Self-confident	-1.25	.41	-.81
	(1.32)	(1.53)	(1.12)
Gender TM * Self-confident	.67	.008	-.89
	(1.36)	(1.45)	(1.06)
Male			2.02**
			(.95)
Non-select TM	2.11**	2.68**	3.51**
	(.97)	(1.04)	(.90)
Gender TM	-.16	-.37	.32
	(1.02)	(.99)	(.86)
Self-confident	.26	1.02	1.64*
	(1.00)	(1.09)	(.92)
Age	.17	-.15*	.06
	(.15)	(.08)	(.12)
Constant	2.52	9.04**	3.64
	(3.63)	(2.09)	(2.87)
Obs.	113	116	229
R squared	.24	.24	.20

Estimated coefficients from OLS regressions with bootstrapped standard errors in parentheses (1000 replications). Non-select TM is a dummy variable equal to one for the treatments *Man* and *Neutral*, and zero otherwise. Gender TM is a dummy that equals one for the gendered treatments *Man* and *Selected-Woman*. Self-confidence is a dummy equal to one if the belief about the own MRT score is above the median belief by gender.