I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17877

Gender Differences in Performance Evaluations

Katja Görlitz
Tim Sels

APRIL 2025

IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17877

Gender Differences in Performance Evaluations

Katja Görlitz
Hochschule der Bundesagentur für Arbeit and IZA

Tim Sels
UC Berkeley and FU Berlin

APRIL 2025

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Gender Differences in Performance Evaluations*

This study analyzes the gender gap in self- and peer evaluations based on a laboratory experiment. Five players performed a creativity task in a high-stakes winner-takes-all tournament. The treatment without validation informed all players that evaluations that they will conduct determine who will win. The treatment with public validation additionally informed them that they can see an objective performance measure of all players (including themselves) at the end of the experiment which is irrelevant for winning. The results show that men give themselves better selfevaluations compared to women when there is no validation. This gender difference vanishes completely when providing public validation.

JEL Classification: J16, M50

Keywords: gender, self-evaluation, peer evaluation, public validation

Corresponding authors:

Tim Sels
UC Berkeley
Haas School of Business
2220 Piedmont Ave
Berkeley, CA 94720
USA

E-mail: tim.sels@berkeley.edu

Katja Görlitz
Hochschule der Bundesagentur für Arbeit
Seckenheimer Landstraße 16
68163 Mannheim
Germany

E-mail: katja.goerlitz@hdba.de

* The authors are particularly indebted to Christina Gravert for her support with the laboratory experiment and her valuable comments. We acknowledge further comments and suggestions from Natalia Danzer, Martin Dufwenberg, Jan Marcus and Guido Neidhöfer. Additionally, we are grateful to Arne J. Warnke for providing us access to the data from the Unusual Uses Task. We also thank Daniel Arnold and Maresa Sprietsma for their support with conducting the survey to obtain the external raters' creativity score.

1. Introduction

Many organizations use performance evaluations to determine performance pay, salary bonuses and promotions. Most large firms rely on 360° Feedback for this purpose (Bracken et al. 2016) which includes both self-evaluations of one's own performance and peer evaluations of colleagues. While these evaluations can provide managers with information about performance in settings where they can't observe it, 360° Feedback may also increase unethical behavior. Dufwenberg et al. (2024) develop a theory suggesting that individuals have incentives to manipulate self- and peer evaluations to their own advantage. Empirical evidence from Carpenter et al. (2010) and Leibbrandt et al. (2018) documents that peer evaluations are vulnerable to sabotage, manipulation and cheating. In high-stakes tournaments, sabotage frequently occurs more generally (Charness et al. 2014, Flory et al. 2016, Harbring and Irlenbusch 2011). We are the first to analyze the extent to which there is a gender gap in self- and peer evaluations in a high-stake, competitive tournament laboratory setting. Thus, this study contributes not only to the literature on individuals' evaluation behavior, but also to the understanding of the gender wage gap. If men and women systematically differ in the extent to which they overrate own performance and underrate their peers' performance in evaluations, this could explain the gender pay gap either directly, through smaller performance bonuses, or indirectly, through the lower proportion of women in leadership positions.[1] The primary research questions are: (i) Do men and women differ in self-evaluations? and (ii) Is there a gender gap in peer evaluations?

The previous literature finds that men respond more strongly to competitive incentives than women, although this can vary depending on the setting. Exley and Kessler (2022) show that, conditional on actual performance, men are more likely to self-promote than women in a setting that provides incentives to rate oneself favorably. Charness et al. (2018) document that men are more overconfident than women when overreporting provides a strategic advantage to win a tournament. With regard to our first research question, we hypothesize the existence of a gender gap in self-evaluations. Leibbrandt et al. (2018) and Dato and Nieken (2014), using laboratory experiments, find that men are more likely to sabotage their peers when sabotage is costly. Similarly, Flory et al. (2016) observe in a field experiment conducted in a call center that men tend to engage in peer sabotage more frequently. In contrast, Schwieren and Weichselbaumer (2010) do not find evidence supporting this pattern in a laboratory setting where sabotage incurs no costs. Regarding our second research question, these results give rise to the hypothesis that peer evaluations might also differ by gender.

These gender differences in unethical behavior may be rooted in social norms or psychological factors. Men and women were shown to differ, for example, in their behavior regarding competition (Niederle and Vesterlund 2007, Croson and Gneezy 2009), risk aversion (Eckel and Grossman 2008), overconfidence (Bengtsson et al. 2005), altruism (Andreoni and Vesterlund 2001) and cooperation (Kuhn and Villeval 2015).[2] Price (2012) shows that male managers anticipate this by

[1] In 2021, the share of female CEOs in the Fortune 500 was 8 percent (see https://fortune.com/2021/06/02/female-ceos-fortune-500-2021-women-ceo-list-roz-brewer-walgreens-karen-lynch-cvs-thasunda-brown-duckett-tiaa/).

[2] The reason for these differences could be gender roles or societal gender norms (Gneezy et al. 2009).

choosing a tournament compensation scheme for a female worker significantly less often than for a male worker. Additional factors that may help explain gender differences in immoral behavior include concerns about reputation, self-image or potential social norms that sanction women more than men (Braddy et al. 2020).

As an additional contribution to the literature, our study suggests a way to reduce this gender gap. In their theory, Dufwenberg et al. (2024) show that cheating can occur more often when the evaluation setting is ambiguous. This is because higher ambiguity conceals cheating. In ambiguous settings, observers of selfish evaluation behavior are more likely to attribute it to an error in perception rather than to intentional cheating. For example, when evaluating an easy-to-solve math task, incorrect evaluations that increase the raters' winning probability are more likely perceived as cheating compared to evaluations of a task that depends on taste or lack of the rater's expertise (e. g. a cooking task). Reducing ambiguity therefore limits cheating by individuals who would otherwise cheat in more ambiguous settings. Because we suspect that men cheat more than women, manipulating the ambiguity by our treatments should reduce their behavior to a greater extent. Our third research question is: (iii) Does the gender gap diminish when manipulating the experimental design?

This question contributes to the existing empirical literature documenting that the design of how the evaluations were performed matters for individuals' evaluations. Leibbrandt et al. (2018) show that the likelihood of raters underreporting their competitors‘ performance increases when a gender quota determines the winner of a tournament in mixed-gender teams. Kuhnen and Tymula (2012) find that performance and self-evaluations improve when private and anonymous feedback is announced. Botelho and Gertsberg (2022) show that status awards change the way in which raters evaluate others. In contrast to our study, none of these studies analyze the gender gap. Dato and Nieken (2020) examine this question and find that the gender gap in sabotage disappears when revealing a signal about the competitors' level of sabotage. However, they examine sabotage in a setting that differs substantially from ours. In conclusion, none of the previous studies explain how the design of the evaluations can reduce the gender gap in self- and peer evaluations.

Our analysis is based on data from a laboratory experiment conducted in Gothenburg, Sweden, in which five players competed against each other to win the "winner takes it all" prize of 500 Swedish krona (about 50 dollars). After performing a creativity task, the experimenter informed the players that the winner would be the one who received the highest performance evaluations, measured as the sum of each players' self-evaluations and the evaluations of the four competitors. Then, each player had to evaluate their own and their competitors' performance. Using performance evaluations for determining the winner provides incentives of strategic ratings to win the tournament. Players were randomly assigned to a treatment without validation (providing no further information) or a treatment with public validation. In the latter case, the experimenter told the players that they would calculate an objective creativity score for each player, meaning including their own objective performance. Although this score had no relevance in determining the winner, the participants could see it at the end of the experiment. This treatment was chosen to reduce the ambiguity of the evaluation, because the objective score made potential strategic and egoistic evaluation behavior of every individual more visible to one's competitors. The likelihood to be perceived as cheater increased. Dufwenberg et al. (2024) suggests that the higher the probability of detection of unethical behavior, the less likely the behavior is to occur. Harbring and Irlenbusch (2011) find that labeling actions as sabotage helps in reducing such behavior. Harbring et al. (2007) demonstrate that sabotage occurs less frequent when the saboteur's identity is known. Similarly, Eckel and Grossman (2008) show that men act more selfishly in no-risk situations, while there are no behavioral differences between men and women in risky situations. In our study, we hypothesize that men perceive the objective creativity score in the public validation treatment as a risk to be perceived as a cheater, as it could expose their unethical behavior to the other participants, thus, decreasing cheating.

Our results indicate that men engage more strategically in performance evaluations in the treatment without validation. They overrate their own performance more than women do which is a statistically significant finding. On average, they also rate their competitors' performance worse which is, in turn, statistically insignificant, but their lower average ratings still contribute to men's higher winning probability. These results are robust to controlling for the objective creativity score and the share of men per session, taking serial correlation of the standard errors of the participants into account as well as applying session fixed effects that absorb conditions common to all participants at the session level, e. g., weather or day of the week effects. To avoid that the estimated effects are due to women receiving worse evaluations than men[3], the gender of each players' competitors was kept anonymous throughout the experiment. In the treatment with public validation, the gender difference vanishes completely. Here, men's performance evaluations are the same as those of women whose ratings are mainly the same in both treatments. This finding would be in line with an interpretation that mainly men cheated in the treatment without validation. When cheating becomes more obvious by introducing public validation, men cease to cheat which is why their evaluations converge to those of women. These findings imply, on the one hand, that the design of performance evaluations has an impact on the gender gap and, on the other hand, that even minor changes in the design can eliminate it.

The remainder of the study is as follows. The next section presents the experimental design, after which the third section presents the results. Section 4 discusses the results and the last section summarizes the main findings and derives recommendations for management practices.

2. Experimental design

Each tournament consisted of five players competing against each other in a creativity task for a high-stakes prize. The experiment took place with students in Gothenburg, Sweden, in the years 2016/17. The tournaments were part of a session of ten players that lasted up to 60 minutes. The participants were randomly assigned to treatments that varied at the session level. Every session

[3] For example, women receive worse peer evaluations in the scientific referee process (Card et al. 2020) and worse teaching evaluations (Zölitz et al. 2019) than men, even after keeping actual performance constant.

was gender-mixed. Because the participants were kept anonymous throughout the experiment, they did not know who they were competing with and, thus, the gender of their competitors. This is important because we are interested in how men and women differ in evaluating themselves and others and not in how they are evaluated by others in settings where gender is revealed.

The experiment had the following design: First, respondents were informed that they were participating in a tournament where the winner received a prize of 500 Swedish kronor, equivalent to 50 dollars (including the show-up fee). The other players who did not win received only the show-up fee of 50 Swedish kronor. Second, they performed the creativity task. Third, the experimental instructions informed them that self- and peer evaluations determine the winner of the tournament. This step varied by treatment status. The treatment without validation provided no further information. In the treatment with public validation, participants got to know that the experimenters calculated an objective creativity score of each player's performance and that all players can see it after completing the experiment. Fourth, each player had to self-evaluate their own performance and the performance of each of their competitors. Afterward, participants filled out a questionnaire and then players from the treatment with public validation could see the objective creativity score. Last, each player received their money. 70 players participated in the treatment without validation and 70 players in the treatment with public validation, of which two persons did not reveal their gender in the questionnaire.

The creativity task follows Guilford's Test of Unusual (or Alternate) Uses (1956, 1967). It is part of the Torrance Test of Creative Thinking (Torrance 1966), a widely used and reliable test of creative thinking. The test requires individuals to list as many alternative uses as possible for a common object (e. g., a newspaper) within a given time period. In our experiment, participants should name as many unusual uses for a sheet of paper as they could think of within three minutes. They should list new and unfamiliar uses and they could employ larger or smaller sizes of the paper or more than one sheet in a use. To get familiar with the task, individuals participated in a practice round where they should name unusual uses for a rubber tire. The practice round was not part of the competition.

After finishing the test, participants received the information that the winner of the tournament is the player who will obtain the best evaluations. These evaluations come from all players who should perform the self-evaluation of their own performance and the peer evaluations of their four competitors in the next step of the experiment, each on a scale from zero (worst) to ten (best). They should evaluate with regard to originality, considering (i) the number of answers, (ii) their degree of being unusual and (iii) the number of distinct categories from which the answers came. The player with the highest total score which is the sum of the points given to oneself and those awarded by each player's four competitors, wins the tournament. By definition, this total score has a minimum of zero and a maximum of 50.

While this is the information the players in the treatment without validation received, players in the treatment with public validation were additionally informed that the experimenters created an objective creativity score based on the answers of more than 100 test persons doing the same task prior to this experiment. This score takes into consideration: (i) the number of answers, (ii) their degree of being unusual and (iii) the number of distinct categories. They would be able to see these objective creativity scores for each player in their tournament (including themselves) at the end of the experiment which were irrelevant for winning the tournament. Appendix A and Appendix B contain the instructions of the experiment for both treatments.

Calculating the objective creativity score follows Guilford (1956) who assessed creative performance based on the following dimensions: quantity, originality and flexibility. To assess quantity, every valid answer gave one point. Originality was assigned, if fewer participants named the same idea.[4] Flexibility refered to the number of distinct categories the responses fall into, where each category is associated with one additional point. For example, if participants mentioned only a necklace and a bracelet, they would receive one additional point for flexibility, since the responses fell into the category "jewelry". The average of the objective creativity score is 22.33 in the treatment without validation and 22.49 in the treatment with public validation.

To test the sensitivity of the objective creativity score, we conducted a survey involving 274 other students at the University of Applied Labour Studies in Mannheim in 2023, Germany. Each student obtained typed answers from the creativity task of a group of five players from the original experiment. The reason for providing typed answers instead of copies of the original handwritten answers is that the handwriting of each participant should not be seen by the students in our survey. The students were asked to evaluate the answers of the five players on a scale from zero (worst) to ten (best) according to the same criteria as used in the experiment and to measure the objective creativity score: (i) the number of answers, (ii) their degree of being unusual and (iii) the number of distinct categories. The purpose of conducting this external raters' creativity score was to assess the validity of our objective score. Because the creative performance of every player in the experiment was rated by nine to ten survey participants (= 274 students x 5 players / 140 players in the experiment), the external raters' creativity score reflects the mean of these ratings. Figure A-1 in the Appendix shows that the external raters' creativity score is highly correlated with the objective creativity score.

In the survey, we additionally analyzed the extent to which the handwritten responses in the experiment could reveal information about gender, potentially influencing the peer evaluations given by players. This is important because we are interested in how men and women differ in their evaluation behavior, and we are not interested in how they get evaluated by them. To investigate this, we asked the 274 students to additionally assess the gender based on the handwriting of five answers from the creativity task. These answers were different from those used to generate the external raters' creativity score. Each student received on paper the answers of five players from

[4] This required information on the answers of all respondents which the experimenter obtained from data of Bradler et al. (2019) who conducted the unusual uses tasks for a sheet of paper. Based on this, the experimenter compiled a list of answers and calculated the proportion of participants who mentioned them. While more than 60 percent of the participants named a "paper airplane," there are some answers that were mentioned only once, such as a doormat, a tinsel or a megaphone. Every answer got an additional point for originality (on top of the quantity point for this answer), if fewer than eight percent of participants gave the same response. If fewer than one percent gave the same answer, the experimenter assigned two additional points for that answer instead.

the experiment who competed against each other in a group. They had to indicate whether the responses were written by a man, a woman or whether they could not identify the gender. As before with the external raters' creativity score, the handwriting of every player in the experiment was assessed by nine to ten survey participants which is why we calculated its mean for the regression analysis.[5] Descriptive statistics document that survey participants only assigned 46 percent of the gender of the players correctly, 28 percent incorrectly and 25 percent were not identifiable. There were only six players whose gender was assigned correctly from every of the nine to ten survey participants. In conclusion, handwriting is not a good predictor of gender.

3. Results

The experiment allowed participants to increase their probability of winning by overreporting their own performance and by underreporting the performance of their competitors. The probability of winning was highest when one awarded oneself ten points and gave zero points to every other participant. Only six men and two women chose this "10-0-0-0-0" strategy which are too few observations to meaningfully analyze it further by treatment status.[6] As a comprehensive measure that combines over- and underreporting, we define *performance evaluations* as the difference between the points each participant gave to themselves (self-evaluation) and the mean of the points they awarded to their four competitors (peer evaluations). The higher this measure was, the more likely a participant was to win the tournament.

3.1. Gender differences in the treatment without validation

Figure 1 documents that there are gender differences in performance evaluations in the treatment without validation in Panel A. Men have an average performance evaluation of 4.6 which is more than twice the score given by women (2.2). This difference is statistically significant at the 5 percent level ($z = 2.21, p = 0.027$, Mann-Whitney test).[7] Panels B and C show that this finding is a result of men awarding themselves more points in self-evaluations on average ($z = 2.46, p = 0.014$, Mann-Whitney test) and (to a lesser extent) giving lower average peer evaluations to their competitors ($z = –1.48$, $p = 0.138$, Mann-Whitney test), albeit this latter result is statistically insignificant.[8] The findings of the self-evaluations remain statistically significant when calculating t-tests with and without Welch correction. These gender differences are not a result of men's better

[5] We assigned survey participants' answers as 1 (if participants assigned the handwriting as male), 0 (for female handwriting) and 0.5 (for not identifiable).

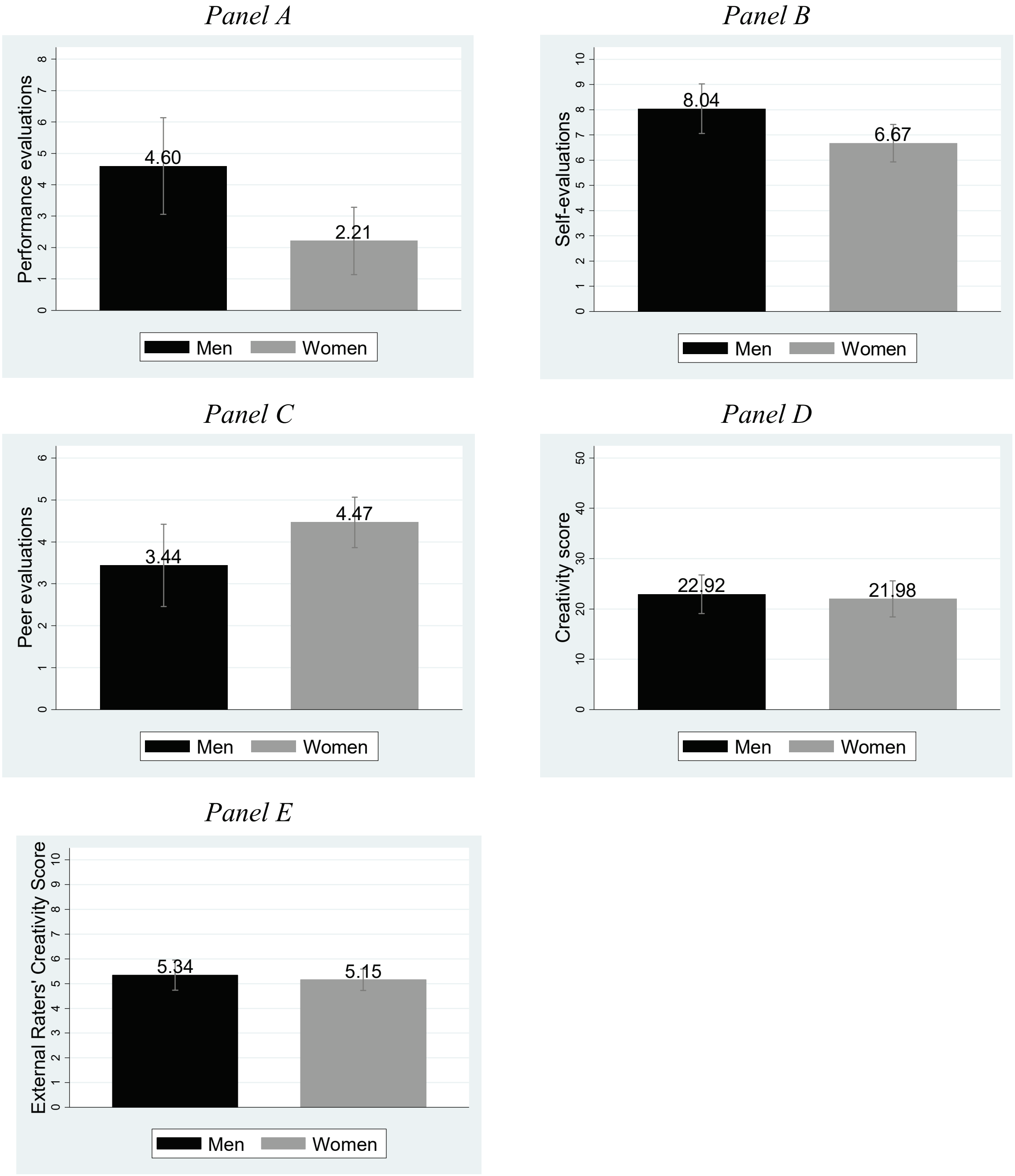
[6] Men used this strategy four times in the treatment without validation and twice in the treatment with public validation, while women used it only in the treatment without validation.

[7] The Mann–Whitney test always refers to a two-sided test throughout this study.

[8] Figure A-2 in the Appendix illustrates the kernel densities of self- and peer evaluations by gender, documenting how much the distribution differs by gender. The difference is quantitatively assessed using Cohen's U1, a measure that calculates the percentage of non-overlap between two populations. For performance evaluations by gender, Cohen's d value is 0.65. This indicates a non-overlap of approximately 40 percent in the distributions which suggest them to differ substantially.

creative performance as Panel D documents that the average objective creativity scores are similar between men and women and that the difference is not statistically significant ($z = 0.539$, $p = 0.590$, Mann-Whitney test). Panel E shows external raters’ creativity score, revealing again that men and women do not differ with regard to their creativity ($z = 0.483$, $p = 0.629$, Mann-Whitney test).

Figure 1. Gender gap in evaluations and creativity in the treatment without validation


Panel A

Panel B

Panel C

Panel D

Panel E

Note: Panel A documents participants' performance evaluations which is measured as the difference between self-evaluations and the mean of the peer evaluations given to one's competitors. Panel B shows the participants' self-evaluations and Panel C the mean of the peer evaluations. Panel D illustrates the objective creativity score that is higher, (i) the more valid answers the participants gave, (ii) the more original these answers were and (iii) the more categories they came from. Panel E also displays external raters' creativity score. The standard error bars show the 95% confidence interval.

Table 1 presents the OLS estimates where we regress performance evaluations on a dummy for men. Because each session consisted of two tournaments, the same conditions applied to each of the ten session participants (e. g., a sunny versus a rainy day), resulting in serially correlated standard errors at the session level. To account for this, we cluster the standard errors at the session level (Moulton 1990) and show them in parentheses. For reason of robustness, we apply wild cluster bootstrap (Roodman et al., 2019) and show this inference in brackets. Column (1) demonstrates that the average gender gap in performance evaluations shown in Figure 1 remains statistically significant even after taking session clusters into account. Although Figure 1 confirms that the creative performance does not explain the gender gap in performance evaluations, column (2) additionally controls for the objective creativity score which leaves the main conclusions unchanged. Column (3) repeats the analysis of the previous column but uses the external raters' creativity score instead of the objective creativity score for reasons of robustness. The conclusion remains robust again.

Next, we provide evidence that these findings represent that men and women differ in how they evaluate others and not that men and women are evaluated differently by gender. Column (4) shows results using the specification from column (1), but further controls for the share of men per session. This neither affects the coefficient to a large extent nor does it change the significance level. Column (5) illustrates similar findings from another robustness check by controlling for the proportion of men per group. Column (6) accounts for the gender based on the handwriting that was assessed by the survey participants to ensure that peer evaluations are not influenced by the rater's perception of the ratee's gender. Also in this specification, the main results hold. These robustness checks from column (4) to (6) confirm our conclusion that the results present differences in how men and women evaluate others. Column (7) applies in addition to the objective creative score also session fixed effects to account for any contemporary effects common to all participants at the session level (e. g. weather conditions, day of the week effects or the share of men per sessions). Even in this specification, the results remain robust. Because this specification accounts for all sources of potential bias, it is our main specification in the following.

Further results indicate that gender differences in performance evaluations are driven by both self- and peer evaluations (see column 1, 2, 4 and 5 in Table A-1 in the Appendix). The specification with self-evaluation as the dependent variable shows that men award themselves approximately 1.2 more points than women do. This difference is statistically significant. Using peer evaluations as the dependent variable indicates that men give their competitors on average one point less compared to women. Even though this effect is not statistically significant, it contributes to the gender gap in performance evaluations that is calculated using both self- and peer evaluations.

Table 1. Regression results of the gender gap in performance evaluations in the treatment without validation

	Performance evaluations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Men	2.387**	2.291**	2.185**	2.249**	2.610*	2.281**	2.196**	2.195*
	(0.878)	(0.810)	(0.715)	(0.877)	(1.085)	(0.819)	(0.843)	(1.086)
	[0.047]	[0.047]	[0.016]	[0.047]	[0.063]	[0.047]	[0.047]	[0.078]
Creativity score	No	Yes	No	Yes	Yes	Yes	Yes	Yes
External raters' creativity score	No	No	Yes	No	No	No	No	No
Share of men per session	No	No	No	Yes	No	No	No	No
Share of men per group	No	No	No	No	Yes	No	No	No
Gender based on handwriting	No	No	No	No	No	Yes	No	No
Session fixed effects	No	No	No	No	No	No	Yes	Yes
Controlling for justification	No	No	No	No	No	No	No	Yes
Adj. R^2	0.0787	0.1518	0.2398	0.1391	0.1463	0.1389	0.2006	0.2251
R^2 overall	0.0922	0.1768	0.2622	0.1770	0.1840	0.1768	0.2946	0.328
Observations	69	69	69	69	69	69	69	69

Notes: The dependent variable is performance evaluations, i.e. the difference between self-evaluations and the mean of the peer evaluations given to one's competitors. The data refer to individuals in the treatment without validation. The table shows the results when regressing performance evaluations on a dummy for men (column 1). Further control variables include individuals' creativity score (column 2, 4, 5, 6, 7), the external raters' creativity score from the survey (column 3), the share of men per session (column 4), the share of men per group (column 5) and the gender based on handwriting per group (column 6). Column 7 shows the results when applying session fixed effects and column 8 controls for justification (for further information see next section). Standard errors clustered at the session level (with seven clusters) are shown in parentheses. The p-values from wild cluster bootstrapped errors clustered at the session level (seven clusters) are shown in brackets. Statistical significance: p<0.1 *, p<0.05 **, p<0.01 ***.

3.2. Gender differences in the treatment without validation and with public validation

Figure 2 shows the gender gap in performance evaluations in the treatment without validation (as already shown in Figure 1) compared to the treatment with public validation. For the latter, the gender gap is small in magnitude and not statistically significant (z = –0.670, p = 0.503, Mann–Whitney test). This is due to men whose performance evaluations are significantly lower in the treatment with public validation compared to the treatment without validation (z = – 2.038, p = 0.042, Mann–Whitney test). For women, the difference by treatment status is not statistically significant (z = –0.672, p = 0.501, Mann–Whitney test). When performing the t-test with and without Welch correction for reason of robustness, the conclusions are the same. One reason for that could be that mainly men cheated in the treatment without validation. When cheating became more obvious by providing the public validation, men stop cheating and their evaluation results normalize to that of women which is the same across both treatments.

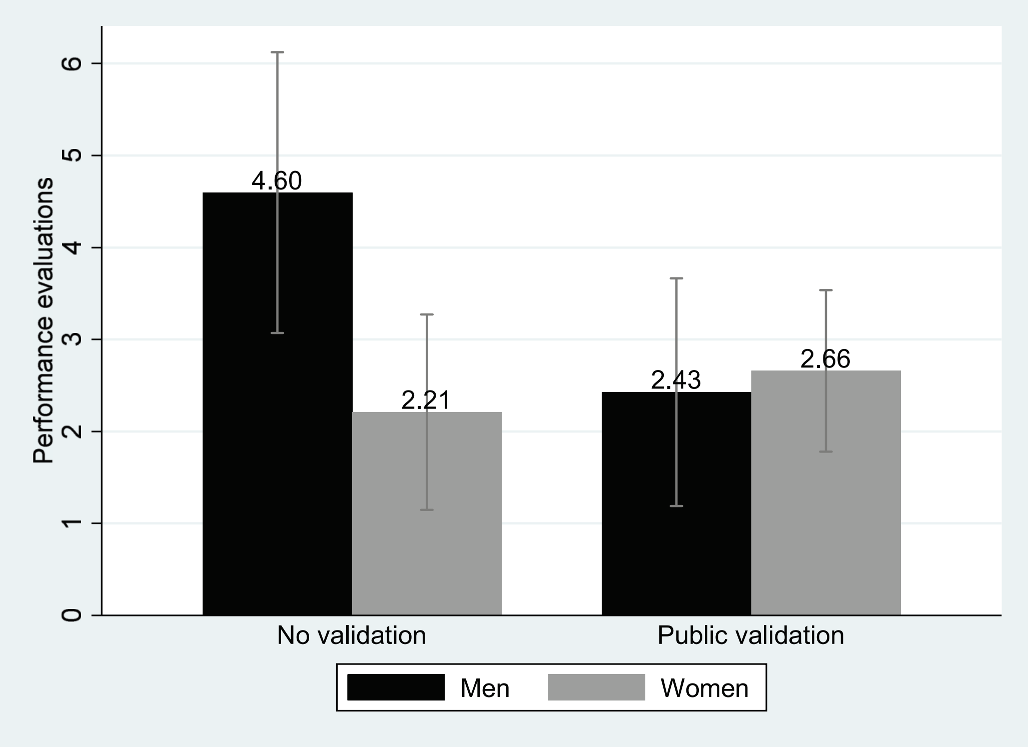
Table A-2 in the Appendix shows the balancing by treatment status using the variables available in the questionnaire.[9] Of the 13 variables, we only find one difference to be statistically significant. This variable refers to the self-assessment of whether one finds cheating justifiable in different situations (i. e., claiming government tax you are not entitled to, avoiding fare on public transport and cheating on taxes). There are two interpretations of this finding. On the one hand, it could hint at problems with the balancing. On the other hand, these self-assessments could be affected by the treatment status itself, meaning that respondents in the treatment without validation indicate cheating as more justifiable in other domains in order to justify their own selfish evaluation choices. This is possible because the corresponding questions had to be asked after the experiment was conducted to avoid this question to bias the findings of the experiment through framing. We consider this interpretation as likely because we know from the question on choices at the end of the experiment that some respondents used the possibility to win by evaluating opportunistically.[10] Additionally, none of the other variables differed from each other on a conventional level of significance which one would suggest in case of nonrandom selection of the respondents. Because justification is similar to other variables like individuals' opinion that most people are fair or pursue their own interests, these variables should differ by treatment status as well which none of them does. However, because it is not clear to which extent this explanation is true, we run additional

[9] As these characteristics cover more than 30 items, we decided to reduce the number of variables by aggregating questions that measure the same characteristic into one variable based on an factor analyses. This was the case for the four questions concerned with competition, the 21 personality questions and the three questions on the justification of cheating. Competition represents one factor only (with positive factor load) which is why we have generated one variable for by taking the sum of the corresponding questions. The same is true for justification. Personality traits were segregated into five factors, representing the big five "OCEAN" characteristics. So, we generated five traits by summing up the answers to the corresponding questions (with negative loads entering the scale in reverse order) and dividing it by the number of questions per factor. Robustness checks show that the conclusions remain the same when using the non-aggregated characteristics, but the precision declines due to collinearity.

[10] Here you find some examples of these answers: „I gave points to the others according to the instruction. But then I gave myself higher scores than the others to improve my own total score but without cheating too much." or „I tried to compare them with each other. Which one was the best. It was hard. Especially giving the others more points than myself. So, I just gave a little bit more than myself. Maybe I should haven given more." or „First I looked which ones where the „worst". Gave them a zero. Gave the others somewhat better. And then I sat about 2 min to think whether I should give myself a fair number or higher. I gave myself a 10. (Not fair)".

sensitivity checks and interpret our results more conservatively. First, we re-estimate the gender gap in evaluation behavior controlling for this variable that we denote “justification” henceforth. The eighth column of Table 1 shows the corresponding results, confirming our main conclusions. Table A-1 shows that the gender gap in self-evaluations remains statistically significant (see column 3), while it remains insignificant for peer evaluations (see column 6). In conclusion, the gender gap documented in the treatment without validation remains robust when controlling for the justification variable.

Figure 2. Gender gap in performance evaluations by treatment.



Notes: The figure illustrates the gender gap in performance evaluations over treatments. The standard error bars show the 95% confidence interval.

Table 2 presents the regression results that analyze the gender difference between treatments controlling for the creativity score, applying session fixed effects and clustering the standard errors at the session level. Column (1) illustrates the results when regressing performance evaluations on a dummy for men and an interaction of the dummy for men and a dummy indicating the public validation treatment. Because the treatments are nested within sessions, a dummy for the public validation treatment drops out of the regression in this specification. The dummy for men indicates that the gender effect in the treatment without validation is statistically significantly positive (as before). The interaction term documents that the gender gap is statistically significantly smaller in the treatment with public validation. The joint interpretation of both terms confirms the findings from Figure 2, documenting that the gender gap in performance evaluations disappears in the case of public validation. To check the robustness of our findings, column (2) presents the results when controlling for the external raters' creativity score instead of the creativity score. The conclusions remain unaffected. Columns (3) and (4) employ the same regression as in column (1), but they include the share of men per group or the gender based on handwriting per group as further control variables, respectively. The results remain consistent again.

Column 5 additionally takes the variable "justification" into account. This changes the previous results only in that the interaction term renders statistically insignificant. As we have already mentioned, that could be because justification is an outcome of the treatment status (and not a problem with the balancing) in which controlling for this variable would bias our results. Because we cannot fully prove this explanation, we conclude that the observed gender gap in Figure 2 does not differ between treatment status on a statistically significant level. Column (6) and (7) account for justification as well, illustrating that men assess their performance better in self-evaluations and their peers worse in the treatment without validation, confirming our previous conclusions. These results also indicate that the gender difference in self-evaluations disappears in the treatment with public validation on a statistically significant level and that the difference in peer evaluations becomes smaller albeit it is statistically insignificant.

This shift towards more equitable evaluations in the public validation treatment is further reinforced by the actual tournament results, which reveal a notable change in the proportion of objectively deserving winners, particularly among women. Of the women who objectively should have won the tournament in the treatment without validation, only 29 percent actually did. For men, the "fair" success rate was 43 percent. In the treatment with public validation, the gap in the objective winner rate between men and women was halved. The proportion of "fair" winners among women increased to 42 percent, while for men, it experienced only a slight increase to 50 percent. Although this difference is not statistically significant due to the small number of observations, the observed pattern indicates that the treatment with public validation provides a significantly greater and fairer chance for women to win the tournament.

Table 2. Regression results of gender gap in evaluation behavior by treatment status

	Performance evaluations					Self-evaluations	Peer evaluations
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Men	2.296 **	2.310 **	2.527 **	2.452 **	2.479 *	1.238 *	-1.057
	(0.876)	(0.882)	(0.995)	(1.005)	(1.209)	(0.577)	(0.677)
	[0.060]	[0.058]	[0.057]	[0.059]	[0.072]	[0.119]	[0.042]
Men x Public validation	-2.352*	-2.394*	-2.495*	-2.499*	-2.307	-1.667**	0.492
	(1.284)	(1.334)	(1.348)	(1.344)	(1.430)	(0.745)	(0.805)
	[0.068]	[0.071]	[0.069]	[0.083]	[0.123]	[0.199]	[0.165]
Creativity score	Yes	No	Yes	Yes	Yes	Yes	Yes
External raters' creativity score	No	Yes	No	No	No	No	No
Share of men per group	No	No	Yes	No	No	No	No
Gender based on handwriting	No	No	No	Yes	Yes	No	No
Session fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controlling for justification	No	No	No	No	Yes	Yes	Yes
Adj. R^2	0.1355	0.1485	0.1359	0.1504	0.183	0.0620	0.1103
R^2 overall	0.2364	0.2479	0.2431	0.2558	0.2903	0.1784	0.2207
Observations	138	138	138	138	138	138	138

Notes: Column (1) illustrates the results when we regress performance evaluations on a dummy for men, an interaction of the dummy for men and a dummy indicating the public validation treatment and the creativity score, including session fixed effects (that eliminates the dummy for public validation because it is nested within session). Column (2) shows the corresponding results that differ only by including the external raters' creativity score instead of the creativity score. Using the same specification as in column (1), columns (3) and (4) additionally control for the share of men per group and the gender based on handwriting per group, respectively. Column 5 controls for justification. Columns (6) and (7) document the results when using self-evaluations or peer evaluations as the dependent variable, respectively. Standard errors clustered at the session level (14) are shown in parentheses and p-values from wild cluster bootstrapped errors clustered at the session level (14) are shown in brackets. Statistical significance: p<0.1 *, p<0.05 **, p<0.01 ***.

4. Discussion of the Results

This section is twofold. First, it sheds light on the determinants of the evaluation behavior by exploiting the individual characteristics from the questionnaire. Second, it addresses the extent to which differences in these characteristics can explain the gender gap in the treatment without validation. If men more frequently exhibit characteristics that are associated with strategic evaluations, this could explain why they evaluate differently from women. Because there is only a gender gap in the treatment without validation, we restrict the following analyses to this treatment. Table A-3 shows which characteristics are associated with the evaluation behavior. While religion becomes statistically significant when analyzing performance and peer evaluations, this is true for higher levels of competitiveness when analyzing self-evaluations. Importantly, in the regressions using performance evaluations and self-evaluations the gender dummy renders statistically insignificant. Please note that this does not imply that the previous results were incorrect. Rather, it complements these findings by demonstrating that some of the considered variables that differ by gender are correlated with the evaluation behavior.

To further explore this issue, we conduct decomposition analyses following the basic idea of Oaxaca (1973) and Blinder (1973) which reveal the extent to which individual characteristics contribute to explaining the observed gender gap in evaluation behavior. The first column of Table 3 presents the results of decomposing performance evaluations by running a pooled regression over men and women and then using these coefficients as the reference (Oaxaca and Ransom 1994). Column (2) illustrates the findings using the same outcome from a model that incorporates an additional indicator variable for gender (Jann 2008). This robustness check is important to rule out the possibility that the explained part of the decomposition is overestimated. Both columns show that competitiveness is the most important factor in explaining the gender gap in evaluation behavior. While it accounts for 37 to 61 percent of the explained part, all other factors are of lesser importance. Column (3) and (4) illustrate the decomposition results for self-evaluations, finding that competitiveness explains 64 to 97 percent which is again much higher than for all other variables. Since the explanatory power of competitiveness is higher when analyzing self-evaluations and because performance evaluation was calculated based on self- and peer evaluations, we suggest that competitiveness explains the gender gap in performance evaluations through self-evaluations. This suggestion is also in line with the results provided in Table A-3, which show that competitiveness is the most influential factor in explaining self-evaluations. In contrast, religion plays a more important role for peer evaluations.

Thus, one might find it puzzling that religion has no explanatory power in the decomposition analyses. However, it is important to note that this is not the only condition required for decomposing the gender gap. Only if characteristics also differ by gender, they can help explain it. While it is a well-established fact that men and women differ in their attitudes towards competition (Niederle and Vesterlund 2007, Croson and Gneezy 2009), there is no reason to believe that there is a gender gap in religiosity. Our data confirms this. The only robust, statistically significant gender difference in individual characteristics is found for competitiveness at the 5 percent level, showing that men report being more competitive.[11] In contrast, there is no gender difference in being religious. Since section 2 did not find a statistically significant gender gap in peer evaluations and because competitiveness operates through self-evaluations, we abstain from running a decomposition analysis for peer evaluations. In conclusion, competitiveness is the most influential characteristic in explaining self-evaluations. How should this be interpreted? If the evaluation instructions do not include a validation mechanism that has the potential to increase honest reporting (as in the treatment without validation), they might be interpreted as an opportunity to win the tournament. And being more competitive means being more willing to win which could lead to higher levels of manipulation of the evaluations in the treatment without validation. However, the model is far from perfect as the unexplained part remains between 60 to 79 percent which suggests that competitiveness is not the only avenue explaining the gender gap.

[11] Extraversion also differs by gender at the ten percent level of significance according to the Mann–Whitney test, but it is insignificant in the t-tests with and without Welch correction.

Table 3. Decomposition analyses

	Explaining the gender gap in performance evaluations				Explaining the gender gap in self-evaluations			
	(1)		(2)		(3)		(4)	
	Coeffi-cient	% from explained	Coeffi-cient	% from explained	Coeffi-cient	% from explained	Coeffi-cient	% from explained
Difference	-2.458		-2.458		-1.373		-1.373	
Explained Part	-0.985	40%	-0.521	21%	-0.547	40%	-0.287	21%
Unexplained Part	-1.473	60%	-1.937	79%	-0.826	60%	-1.086	79%
Creativity score	-0.008	1%	-0.008	0%	-0.004	0%	-0.004	0%
Risk aversion	0.016	-1%	0.053	-3%	0.091	-11%	0.112	-10%
Do you believe most people are fair? (y/n)	-0.158	11%	-0.089	5%	-0.002	0%	0.037	-3%
Do people usually pursue their interests? (y/n)	-0.031	2%	-0.052	3%	-0.020	2%	-0.031	3%
Positive attitude towards competition	-0.896	61%	-0.715	37%	-0.800	97%	-0.698	64%
Personality traits	0.159	-11%	0.324	-17%	0.179	-22%	0.272	-25%
I justify cheating more often	-0.002	0%	-0.002	0%	-0.001	0%	-0.001	0%
Religion (y/n)	-0.092	6%	-0.065	3%	-0.012	1%	0.003	0%
Liquidity constraints (y/n)	0.027	-2%	0.033	-2%	0.021	-3%	0.025	-2%
Observations	62		62		62		62	

Notes: The table shows the results of decompositions of the gender gap in performance and self-evaluations. Columns (1) and (3) contain the results of running a pooled regression over gender and using these coefficients as the reference (Oaxaca and Ransom 1994). Columns (2) and (4) document the findings from a decomposition that includes additionally a binary variable distinguishing men and women (Jann 2008).

5. Conclusion

This study uses a laboratory experiment to examine gender effects in performance evaluations. Participants in the experiment took part in a high-stake “winner-takes-all” tournament where they had to perform a creativity task. After finishing the task, they were informed that only self- and peer evaluations determined the winner. All players had to evaluate their own performance and the performance of their four competitors. There were two randomly assigned treatments. The treatment without validation did not announce an objective authority. In the treatment with public validation, participants were informed that the experimenters would generate an objective creativity score which they could see at the end of the tournament, but that it would have no relevance for determining the winner.

The results show that men give themselves significantly better self-evaluations than women in the treatment without validation. With regard to peer evaluations, men rate their competitors on average worse than women do. Even though this latter effect is not statistically significant, it still raises a players’ winning probability in the tournament. We can rule out that differences in creative performance explain this gender gap because the conclusions remain the same after controlling for the objective creativity score. Our results reinforce the findings of Exley and Kessler (2022) who demonstrate that men tend to evaluate themselves more positively than women even after accounting for actual performance. They interpret their results as a gender gap in self-promotion. To the extent that evaluations positively affect promotions or wage negotiations, our findings contribute to explaining the gender pay gap.

In line with Exley and Kessler (2022), we present some evidence that the setting of the experiment can change the results, even though this result is not as robust as the previous findings. These findings indicate that the gender evaluation gap disappears when providing public validation. The reason for this finding is that men give higher points in self-evaluations in the treatment without validation compared to public validation. In conclusion, minor changes in the design of the evaluations have the potential to mitigate the gender gap. This public validation setting is similar to an evaluation settings where managers use the support of an internal control authority (e. g., HR department, audit) or an external HR consulting firm as public validation. It might also hold true when the peer evaluations are made public within a team which is sometimes the case when using 360° Feedback. Most importantly, as we find that the gender gap occurs most severely in self-evaluations, firms should use them either more carefully or even ignore them when evaluating the performance of their employees based on the 360° Feedback. Especially when these performance evaluations are used to decide upon promotions or wages, ignoring individuals’ self-evaluations has the potential to reducing or even avoiding gender biases. Because we find that the gender gap in the treatment without validation is due to differences in competitive attitudes, all measures that reduce the gender gap in competitiveness seem promising as well. This includes, e. g., informing women about this gender gap (Kessel et al. 2021) or providing relative performance feedback (Wozniak et al. 2014) that managers and policymakers could implement.

References

Andreoni, J. and L. Vesterlund (2001). Which is the Fair Sex? Gender Differences in Altruism. The Quarterly Journal of Economics, 116 (1), 293-312.

Bengtsson, C., M. Persson and P. Willenhag (2005). Gender and overconfidence. Economics Letters 86, 199-203.

Blinder, A.S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. Journal of Human Resources, 8, 436-455.

Botelho, T. L. and M. Gertsberg (2022). The Disciplining Effect of Status: Evaluator Status Awards and Observed Gender Bias in Evaluations. Management Science, 68 (7), 5311-5329.

Bracken, D. W., D. S. Rose and A. H. Church (2016). The Evolution and Devolution of 360° Feedback. Industrial and Organizational Psychology, 9 (4), 761-794.

Braddy, P. W., R. E. Sturm, L. Atwater, S. N. Taylor and R. A. McKee (2020). Gender Bias Still Plagues the Workplace: Looking at Derailment Risk and Performance With Self-Other Ratings. Group & Organization Management, 45 (3), 315-350.

Bradler, C., S. Neckermann and A. J. Warnke (2019). Incentivizing Creativity: A Large-Scale Experiment with Tournaments and Gifts. Journal of Labor Economics, 37 (3), 793-851.

Card, D., S. DellaVigna, P. Funk and N. Iriberri (2020). Are Referees and Editors in Economics Gender Neutral? The Quarterly Journal of Economics, 135 (1), 269-327.

Carpenter, J., P. H. Matthews and J. Schirm (2010). Tournaments and Office Politics: Evidence from a Real Effort Experiment. American Economic Review, 100 (1), 504-517.

Charness, G., D. Masclet and M. C. Villeval (2014). The Dark Side of Competition for Status. Management Science, 60 (1), 38-55.

Charness, G., A. Rustichini and J. van de Ven (2018). Self-confidence and strategic behavior. Experimental Economics, 21, 72-98.

Croson, R. and U. Gneezy (2009). Gender Differences in Preferences. Journal of Economic Literature, 47 (2), 448-474.

Dato, S. and P. Nieken (2014). Gender differences in competition and sabotage. Journal of Economic Behavior & Organization, 100, 64-80.

Dato, S. and P. Nieken (2020). Gender differences in sabotage: the role of uncertainty and beliefs. Experimental Economics 23, 353-391.

Dufwenberg, M., Görlitz, K. and C. Gravert (2024). Peer Evaluation Tournaments. CEBI Working Paper Series No. 20/24.

Eckel, C. C. and P. J. Grossman (2008). Men, Women and Risk Aversion: Experimental Evidence. In C. R. Plott & V. L. Smith (Eds.), Handbook of Experimental Economics Results (Volume 1, pp. 1061–1073). Amsterdam, New York, Oxford, Tokyo: North Holland.

Exley, C. L. and J. B. Kessler (2022). The gender gap in self-promotion. The Quarterly Journal of Economics, 137 (3), 1345-1381.

Flory, J. A., A. Leibbrandt and J. A. List (2016). The Effects of Wage Contracts on Workplace Misbehaviors: Evidence from a Call Center Natural Field Experiment. NBER Working paper 22342.

Gneezy, U., K. L. Leonard and J. A. List (2009). Gender Differences in Competition: Evidence from a Matrilineal and Patriarchal Society. Econometrica, 77 (5), 1637-1664.

Guilford, J. P. (1956). The structure of intellect. Psychological Bulletin, 53 (4), 267-293.

Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Harbring, C. and B. Irlenbusch (2011). Sabotage in Tournaments: Evidence from a Laboratory Experiment. Management Science, 57 (4), 611-627.

Harbring, C., B. Irlenbusch, M. Kräkel and R. Selten (2007). Sabotage in Corporate Contests – An Experimental Analysis. International Journal of the Economics of Business, 14 (3), 367-392.

Jann, B. (2008). The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal, 8 (4), 453-479.

Kessel, D., J. Mollerstrom and R. van Veldhuizen (2021). Can simple advice eliminate the gender gap in willingness to compete? European Economic Review, 138, 103777.

Kuhn, P. and M. C. Villeval (2015). Are Women More Attracted to Co-operation Than Men? The Economic Journal, 582 (1), 115-140.

Kuhnen, C. M. and A. Tymula (2012). Feedback, Self-Esteem and Performance in Organizations. Management Science, 58 (1), 94-113.

Leibbrandt, A., L. C. Wang and C. Foo (2018). Gender quotas, competitions and peer review: Experimental evidence on the backlash against women. Management Science, 64 (8), 3501-3516.

Moulton, B. R. (1990). An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit. Review of Economics and Statistics, 72, 334-338.

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

Oaxaca, R.L. (1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review, 14, 693-709.

Oaxaca, R. and M. Ransom (1994). On discrimination and the decomposition of wage differentials. Journal of Econometrics, 61 (1), 5-21.

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

Roodman, D., J. G. MacKinnon, M. D. Webb, M. A. Nielsen (2019). Fast And Wild: Bootstrap Inference In Stata Using Boottest. The Stata Journal, 19, 4-60.

Schwieren, C. and D. Weichselbaumer (2010). Does competition enhance performance or cheating? A laboratory experiment. Journal of Economic Psychology, 31 (3), 241-253.

Torrance, E. P. (1966). Torrance tests of creative thinking—norms technical manual research edition—verbal tests, forms A and B—figural tests, forms A and B. Princeton: Personnel Pres. Inc.

Wozniak, D., T. Harbraugh and U. Meyer (2014). The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices. The Journal of Labor Economics, 32 (1), 161-198.

Zölitz, U., F. Mengel and J. Sauermann (2019). Gender bias in Teaching Evaluations. Journal of the European Economic Association 17 (2): 535-566.

Appendix

Table A-1. Regression results of gender gap in self- and peer evaluations in the treatment without validation

	Self-evaluations			Peer evaluations		
	(1)	(2)	(3)	(4)	(5)	(6)
Men	1.213**	1.181**	1.180**	-1.036	-1.016	-1.015
	(0.475)	(0.459)	(0.475)	(0.622)	(0.648)	(0.816)
	[0.047]	[0.047]	[0.031]	[0.125]	[0.125]	[0.172]
Creativity score	Yes	Yes	Yes	Yes	Yes	Yes
Share of men per session	Yes	No	No	Yes	No	No
Session fixed effects	No	Yes	Yes	No	Yes	Yes
Controlling for justification	No	No	Yes	No	No	Yes
Adj. R^2	0.1059	0.1002	0.0859	0.0460	0.1196	0.1950
R^2 overall	0.1454	0.2061	0.2069	0.0881	0.2232	0.3014
Observations	69	69	69	69	69	69

Notes: The table shows the results when regressing self-evaluations (peer evaluations) on a dummy for men, individuals' creativity score and the share of men per session in column 1 (column 4), respectively. Columns 2 and 5 show the corresponding results when applying session fixed effects. Columns 3 and 6 control for justification. Standard errors clustered at the session level (with seven clusters) are shown in parentheses. The p-values from wild cluster bootstrapped errors clustered at the session level (seven clusters) are shown in brackets. Statistical significance: p<0.1 *, p<0.05 **, p<0.01 ***.

Table A-2. Balancing by treatment status

Variables	Mean		p-value of t-test		Mann–Whitney test
	No validation	Public validation	equal variance	unequal variance	
Men	0.377	0.449	0.391	0.391	0.389
Creativity score	22.333	22.493	0.932	0.932	0.908
Self-assessment of risk aversion (scale: 1-10)	5.913	5.855	0.869	0.869	0.990
Do you believe most people are fair? (y/n)	0.493	0.435	0.503	0.503	0.501
Do people usually pursue their interests? (y/n)	0.455	0.391	0.448	0.448	0.446
Positive attitude toward competition (higher value in sum of the competition questions)	18.28	18.07	0.6619	0.6619	0.6759
Openness	5.1643	5.223	0.7168	0.7166	0.5899
Neuroticism	4.2947	4.3578	0.8094	0.8094	0.8459
Extraversion	4.8603	4.709	0.4449	0.4452	0.4855
Agreeableness	4.8696	4.75	0.4805	0.48	0.363
Conscientiousness	3.8043	3.7647	0.8182	0.818	0.9966
Attitude towards justification of cheating (higher value in sum of cheating questions)	11.23	9.66	0.0742	0.0742	0.0742
Binary variable for being religious	0.246	0.358	0.164	0.163	0.163
Being liquidity-contrained (scale: 1-4)	1.957	2043	0.610	0.610	0.332

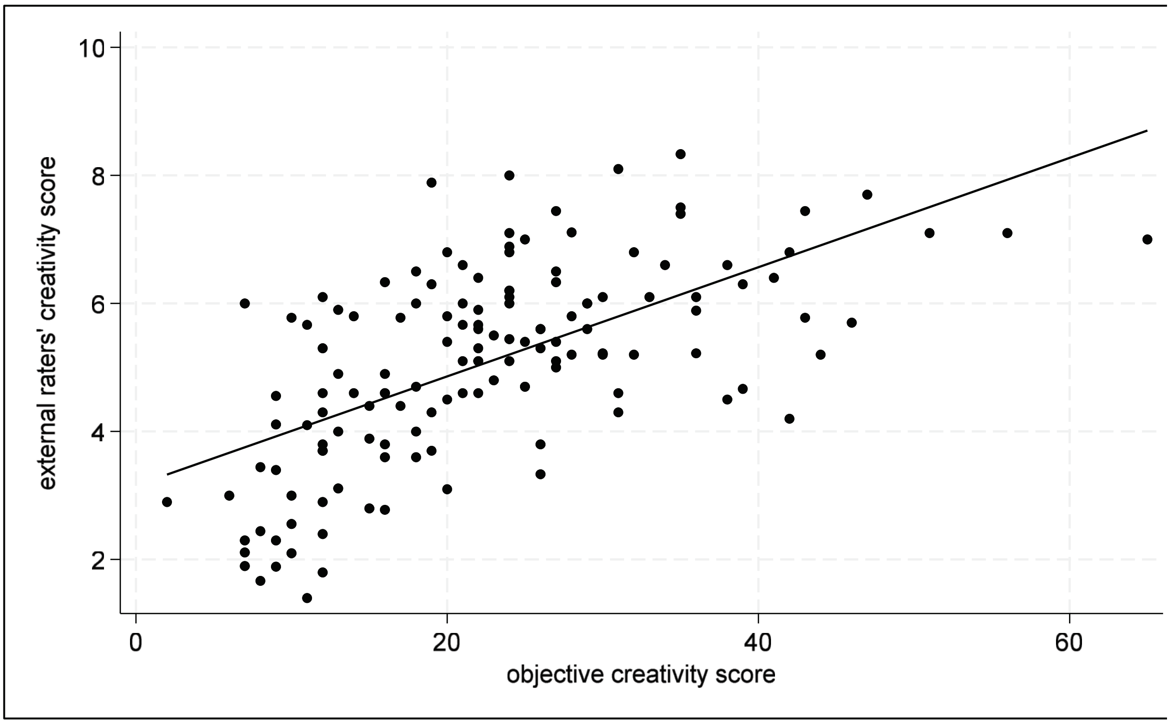
Notes: The table presents the mean of individuals' characteristics by treatment status, p-values of differences in the means with t-tests with and without Welch correction and with a Mann–Whitney test. Statistical significance: p<0.1 *, p<0.05 **, p<0.01 ***.

Table A-3. Determinants of evaluation behavior

	Performance evaluations (1)		Self- evaluations (2)		Peer evaluations (3)	
	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Men	1.937	(1.466)	1.086	(0.781)	-0.851	(0.781)
Creativity score	0.079	(0.084)	0.037	(0.057)	-0.042	(0.040)
Risk aversion	0.121	(0.346)	0.254	(0.131)	0.133	(0.311)
Do you believe most people are fair?	-0.472	(1.594)	0.194	(0.943)	0.666	(0.995)
Do people pursue their interests?	-0.620	(0.598)	-0.376	(0.631)	0.244	(0.428)
Positive attitude toward competition	0.200	(0.176)	0.195 *	(0.098)	-0.005	(0.085)
Openness	0.111	(0.766)	0.364	(0.380)	0.254	(0.585)
Neuroticism	0.067	(0.260)	-0.114	(0.117)	-0.181	(0.197)
Extraversion	0.745	(0.643)	0.488	(0.316)	-0.257	(0.451)
Agreeableness	-0.366	(0.621)	-0.297	(0.366)	0.069	(0.340)
Conscientiousness	-0.553	(0.391)	-0.518	(0.417)	0.035	(0.215)
I justify cheating more often	0.039	(0.101)	0.025	(0.066)	-0.014	(0.042)
Religious (y/n)	-1.191 *	(0.521)	0.059	(0.350)	1.250 *	(0.623)
Liquidity-contrained	-0.763	(1.059)	-0.565	(0.670)	0.198	(0.702)
Creativity score	Yes		Yes		Yes	
Share of men per session	Yes		No		No	
Session fixed effects	No		Yes		Yes	
R^2 overall	0.3811		0.3942		0.2655	
Observations	62		62		62	

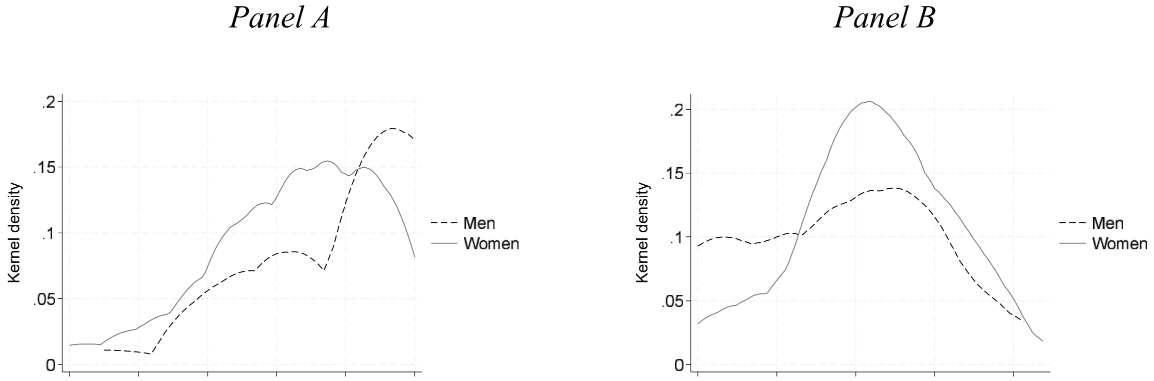
Notes: The table shows OLS results regressing performance evaluations, self-evaluations and peer evaluations on individual characteristics. Standard errors clustered at the session level (with seven clusters) are shown in parentheses. Statistical significance: p<0.1 *, p<0.05 **, p<0.01 ***.

Figure A-1. Correlation between the objective and the external raters’ creativity score



Note: The figure shows the correlation between the objective and the external raters’ creativity score. The solid straight line presents the linear prediction of an OLS regression.

Figure A-2. Kernel densities of self- and peer evaluations by gender in the treatment without validation


Panel A

Panel B

Note: The figure shows the kernel densities of participants’ self-evaluations (Panel A), and the mean of peer evaluations (Panel B) by gender in the treatment without validation.

Appendix A: Instructions for treatment without validation

Group: Player:

1

Thank you for participating in our experiment. We will begin shortly. Today's experiment will last up to 60 minutes.

In this experiment you will have the opportunity to earn some money. Therefore, it is in your interests to pay attention to the instructions and make careful choices. Your earnings will be added to your show up fee of 50 SEK and will be paid out to you after the session.

Anonymity:

Should you agree to participate your name will not be connected to any decision that you make here today or any answer that you provide. All of your actions and information you provide are kept completely anonymous.

Some Rules:

Please switch your cell phone on completely silent (no vibration) and put away anything else that you have brought with you. Please do not talk to other participants during the experiment or attempt to look at the questionnaires of other participants. Do not skip ahead in the papers. Wait for the instructions by the experimenter.

If you have any questions at this point, or at any later point during the experiment, please simply raise your hand. A member of the research team will come to you and answer them in private.

Anyone violating these rules may be excluded from the experiment. In this case he/she will forfeit any earnings.

Structure of the experiment:

The experiment consists of solving a task and of filling out a questionnaire. Before solving the task, there will be a detailed description of the task and a practice round. Please find more instructions on the task and the practice round on the next page.

Please write your group and your player number on top of each page!

Please take your time to read all instructions carefully before making any decisions!

Group: Player:

2

Task description

In the experiment, you will be randomly assigned to a group of five players. You will all do the same task. The group winner of this task will receive 500 SEK (including show-up fee). The other four players receive their show-up fee of 50 SEK. You will never be informed by the research group whom you compete with.

Before starting the real task, you will first do a practice round on the example below. You will not be able to earn money in the practice round, but you can prepare yourself for the real task that will start afterwards.

Explanation of the task for the practice round:
Please list as many different and unusual uses for a **rubber tire** (gummi däck) as you can think of. Do not restrict yourself to a specific size of a tire. You can also list uses that require several tires. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible!

Please write unusual uses. Using a tire as a "car tire" is not an unusual use, using it as a swing is. Answers are also invalid when they would be impossible to create. Using a tire as a flower pot is possible, using it as a space rocket is not. Try to think of different categories for usage. A "garden seat" and a "swing seat" both fall under the category "tires as seats" while using a tire as a "flower pot" is a different usage category ("container"). You should not name any categories.

In order to win it is in your best interest to write as many, original uses that come from as many different categories as you can think of.

Please quietly raise your hand, if there are any questions.

Do not turn to the next page before instructed, otherwise you will be excluded from the experiment and you will not receive any earnings.

Group: Player:

3

Practice Round

The Task

Please list up as many different and unusual uses for **a rubber tire** as you can think of. Do not restrict yourself to a specific size of a tire. You can also list uses that require several tires. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible!

You have three minutes for the task.

□ ○

Please write your ideas below in **CAPITAL LETTERS**!

Group: Player:

4

Official Round

The Task

Please list up as many different and unusual uses for **a sheet of paper** as you can think of. Do not restrict yourself to a specific size of the paper. You can also list uses that require several sheets of paper. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible!

You have three minutes for the task.

Please write your ideas below in **CAPITAL LETTERS**!

□ ○

Group: Player:

5

The Scoring Rules

The winner of the 500 SEK will be determined by the following procedure:

You will now evaluate your own answer and the answers of the other four players with whom you compete for the 500 SEK. Please evaluate the answers with respect to their originality. Originality is scored for each person on a scale from 0 to 10 where 0 indicates overall “not at all original answers” and 10 “very original” answers. For scoring, take into consideration i) the number of answers, ii) their degree of being unusual and iii) the number of different categories they come from. The other players in your group will also do the same scoring.
The points given and the points received are kept anonymous by the research team as well as the information who are the players in your group.

For each player, the research team will add up the points given to a TOTAL SCORE of a minimum of 0 and a maximum of 50. The person with the highest TOTAL SCORE out of your group will receive the 500 SEK (including show-up fee). Who has won will be announced by the research team at the end of the experiment. The others receive their show-up fee of 50 SEK. If there is a tie, the winner will be chosen randomly.

You will now receive the five sheets of paper with the answers of each player of your group. Please write your score from 0 to 10 in the BOX above the text field. Please write your player number in the CIRCLE.

Group: Player:

6

Questionnaire

1. **How do you see yourself: Are you generally willing to take risks (risk-prone), or do you try to avoid risks (risk-averse)? Please answer on a scale from 0 to 10 where 0 means risk-averse and 10 means risk-prone.**

Risk-averse **1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	Risk-prone **10**
□	□	□	□	□	□	□	□	□	□

2. **Do you believe that most people…**

- would use you if they had the chance…….. □

- or that they would try to be fair to you?... □

3. **Would you say that people usually…**

- try to be helpful... □

- or try to pursue their own interests?……… □

4. **Please state how much the following statements describe you…**

	Strongly disagree **1**	**2**	**3**	**4**	**5**	**6**	Strongly agree **7**
- I enjoy working in situations involving competition with others…	□	□	□	□	□	□	□
- It is important to me to perform better than others on a task…..	□	□	□	□	□	□	□
- I feel that winning is important in both work and games…………….	□	□	□	□	□	□	□
- I try harder when I am in competition with other people…………..	□	□	□	□	□	□	□

Group: Player:

7

5. **What kind of personality are you? People can have many different qualities—some are listed below. You will probably find that some of these descriptions fit you completely, some not at all and others may fit to a certain extent. Please answer on a scale from 1 to 7, where 1 means "does not describe me at all", and 7 meaning "describes me perfectly".**

I am someone who…	Does not describe me at all **1**	**2**	**3**	**4**	**5**	**6**	Describes perfectly **7**
- does a thorough job	□	□	□	□	□	□	□
- is talkative	□	□	□	□	□	□	□
- believes that laws and policies should change to reflect the needs of a changing world	□	□	□	□	□	□	□
- I often tries new things just for trying	□	□	□	□	□	□	□
- is sometimes rude to others	□	□	□	□	□	□	□
- is original, comes up with new ideas	□	□	□	□	□	□	□
- worries a lot	□	□	□	□	□	□	□
- prefers to spend time in familiar surroundings	□	□	□	□	□	□	□
- has a forgiving nature	□	□	□	□	□	□	□
- tends to be lazy	□	□	□	□	□	□	□
- is curious about many different things	□	□	□	□	□	□	□
- outgoing, sociable	□	□	□	□	□	□	□
- values artistic, aesthetic experiences	□	□	□	□	□	□	□
- gets nervous easily	□	□	□	□	□	□	□
- believes that religious authorities should be involved when deciding moral issues	□	□	□	□	□	□	□
- does things efficiently	□	□	□	□	□	□	□
- is reserved, quiet	□	□	□	□	□	□	□
- is considerate and kind to almost everyone	□	□	□	□	□	□	□
- has an active imagination	□	□	□	□	□	□	□
- is relaxed, handles stress well	□	□	□	□	□	□	□
- believes that ideals and principles are more important than open-mindedness	□	□	□	□	□	□	□

Group: Player:

8

6. **For each of the following actions or activities, please indicate whether you think that it can always be justified, never be justified, or something in between. You may use any response from 1 to 10 to reflect the strength of your feeling. "1" indicates that it is never justifiable and "10" that it is always justifiable. Make a cross in each of the three rows.**

	Never justifiable									Always justifiable
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Claiming government benefits to which you are not entitled………………………………	□	□	□	□	□	□	□	□	□	□
Avoiding a fare on public transport………	□	□	□	□	□	□	□	□	□	□
Cheating on taxes if you have a chance..	□	□	□	□	□	□	□	□	□	□

7. **Do you belong to a religion or religious denomination?**

Yes…………………….. □

No …………………….. □

Prefer not to answer ……………….. □

8. **With which one of the following statements do you agree most?**

The basic meaning of religion is:

To follow religious norms and ceremonies………………… □

To do good for other people……………………………… □

9. **Assume that an essential commodity of daily use gets broken. How easily would you be able to afford 2,000 SEK to replace the commodity within two weeks without having to borrow the money?**

Very easily	Rather easily	Rather difficult	Not at all able
□	□	□	□

Group: Player:

9

10. What is your gender?

Male…………….. □

Female………… □

Prefer not to answer………… □

Thanks for participating in the questionnaire! Please wait for the next instructions.

Group: Player: 10

Objective Scores

Prior to this experiment, we had more than 100 test persons do the same task as you. With these results, we created a catalogue of possible answers.

We used this catalogue to calculate an OBJECTIVE SCORE (with scores between 0 and 10) for each participant in your group taking into consideration the number of answers, their degree of being unusual and the number of different categories they come from.

Important: These OBJECTIVE SCORES are not used to determine the winner. The winner is solely determined by the TOTAL SCORE (composed of the scores given by one's players in the group and one's own assessment of the task). The TOTAL SCORES will not be revealed except for announcing the winner of each group.

You now have the chance to see the OBJECTIVE SCORES we have calculated. It is for information only.

You do not need to look at the OBJECTIVE SCORES, if you don't want to. But you will have to wait until all people are finished. Your choice has, thus, no influence on the time you stay here for the experiment.

Please circle your answer below. "Yes", if you want to see the OBJECTIVE SCORES and "No", if you do not want to see them.

In the case of "No" you will receive an empty piece of paper.

YES **NO**

Group: Player:

11

Choices

We would like to understand how you made your decisions in the experiment. Please tell us why you chose to allocate the points in the way you did.

____

____

____

____

____

____

____

Appendix B: Instructions for treatment with public validation

Group: Player:

1

Thank you for participating in our experiment. We will begin shortly. Today's experiment will last up to 60 minutes.

In this experiment you will have the opportunity to earn some money. Therefore, it is in your interests to pay attention to the instructions and make careful choices. Your earnings will be added to your show up fee of 50 SEK and will be paid out to you after the session.

Anonymity:

Should you agree to participate your name will not be connected to any decision that you make here today or any answer that you provide. All of your actions and information you provide are kept completely anonymous.

Some Rules:

Please switch your cell phone on completely silent (no vibration) and put away anything else that you have brought with you. Please do not talk to other participants during the experiment or attempt to look at the questionnaires of other participants. Do not skip ahead in the papers. Wait for the instructions by the experimenter.

If you have any questions at this point, or at any later point during the experiment, please simply raise your hand. A member of the research team will come to you and answer them in private.

Anyone violating these rules may be excluded from the experiment. In this case he/she will forfeit any earnings.

Structure of the experiment:

The experiment consists of solving a task and of filling out a questionnaire. Before solving the task, there will be a detailed description of the task and a practice round. Please find more instructions on the task and the practice round on the next page.

Please write your group and your player number on top of each page!

Please take your time to read all instructions carefully before making any decisions!

Group: Player:

2

Task description

In the experiment, you will be randomly assigned to a group of five players. You will all do the same task. The group winner of this task will receive 500 SEK (including show-up fee). The other four players receive their show-up fee of 50 SEK. You will never be informed by the research group whom you compete with.

Before starting the real task, you will first do a practice round on the example below. You will not be able to earn money in the practice round, but you can prepare yourself for the real task that will start afterwards.

Explanation of the task for the practice round:
Please list as many different and unusual uses for a **rubber tire** (gummi däck) as you can think of. Do not restrict yourself to a specific size of a tire. You can also list uses that require several tires. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible!

Please write unusual uses. Using a tire as a "car tire" is not an unusual use, using it as a swing is. Answers are also invalid when they would be impossible to create. Using a tire as a flower pot is possible, using it as a space rocket is not. Try to think of different categories for usage. A "garden seat" and a "swing seat" both fall under the category "tires as seats" while using a tire as a "flower pot" is a different usage category ("container"). You should not name any categories.

In order to win it is in your best interest to write as many, original uses that come from as many different categories as you can think of.

Please quietly raise your hand, if there are any questions.

Do not turn to the next page before instructed, otherwise you will be excluded from the experiment and you will not receive any earnings.

Group: Player:

3

Practice Round

The Task

Please list up as many different and unusual uses for **a rubber tire** as you can think of. Do not restrict yourself to a specific size of a tire. You can also list uses that require several tires. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible!

You have three minutes for the task.

□ ○

Please write your ideas below in **CAPITAL LETTERS**!

Group: Player:

4

Official Round

The Task

Please list up as many different and unusual uses for **a sheet of paper** as you can think of. Do not restrict yourself to a specific size of the paper. You can also list uses that require several sheets of paper. Do not restrict yourself to uses you are familiar with, but think of as many new uses as possible!

You have three minutes for the task.

□ ○

Please write your ideas below in **CAPITAL LETTERS**!

Group: Player:

5

The Scoring Rules

The winner of the 500 SEK will be determined by the following procedure:

You will now evaluate your own answer and the answers of the other four players with whom you compete for the 500 SEK. Please evaluate the answers with respect to their originality. Originality is scored for each person on a scale from 0 to 10 where 0 indicates overall “not at all original answers” and 10 “very original” answers. For scoring, take into consideration i) the number of answers, ii) their degree of being unusual and iii) the number of different categories they come from. The other players in your group will also do the same scoring.

The points given and the points received are kept anonymous by the research team as well as the information who are the players in your group.

For each player, the research team will add up the points given to a TOTAL SCORE of a minimum of 0 and a maximum of 50. The person with the highest TOTAL SCORE out of your group will receive the 500 SEK (including show-up fee). Who has won will be announced by the research team at the end of the experiment. The others receive their show-up fee of 50 SEK. If there is a tie, the winner will be chosen randomly.

Prior to this experiment, we had more than 100 test persons do the same task as you. With these results, we created a catalogue of possible answers.

We will use this catalogue to calculate an OBJECTIVE SCORE (with scores between 0 and 10) for each participant in your group taking into consideration the number of answers, their degree of being unusual and the number of different categories they come from.

You will be able to see these OBJECTIVE SCORES for each player in your group at the end of the experiment.

Important: These OBJECTIVE SCORES are not used to determine the winner. The winner is solely determined by the TOTAL SCORE (composed of the scores given by one’s players in the group and one’s own assessment of the task).

You will now receive the five sheets of paper with the answers of each player of your group. Please write your score from 0 to 10 in the BOX above the text field. Please write your player number in the CIRCLE.

Group: Player:

6

Questionnaire

1. **How do you see yourself: Are you generally willing to take risks (risk-prone), or do you try to avoid risks (risk-averse)? Please answer on a scale from 0 to 10 where 0 means risk-averse and 10 means risk-prone.**

Risk-averse									Risk-prone
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
□	□	□	□	□	□	□	□	□	□

2. **Do you believe that most people…**

- would use you if they had the chance…….. □

- or that they would try to be fair to you?... □

3. **Would you say that people usually…**

- try to be helpful... □

- or try to pursue their own interests?……… □

4. **Please state how much the following statements describe you…**

	Strongly disagree						Strongly agree
	1	**2**	**3**	**4**	**5**	**6**	**7**
- I enjoy working in situations involving competition with others…	□	□	□	□	□	□	□
- It is important to me to perform better than others on a task…..	□	□	□	□	□	□	□
- I feel that winning is important in both work and games…………….	□	□	□	□	□	□	□
- I try harder when I am in competition with other people…………..	□	□	□	□	□	□	□

Group: Player:

7

5. **What kind of personality are you? People can have many different qualities—some are listed below. You will probably find that some of these descriptions fit you completely, some not at all and others may fit to a certain extent. Please answer on a scale from 1 to 7, where 1 means "does not describe me at all", and 7 meaning "describes me perfectly".**

I am someone who…	Does not describe me at all						Describes perfectly
	1	**2**	**3**	**4**	**5**	**6**	**7**
- does a thorough job	□	□	□	□	□	□	□
- is talkative	□	□	□	□	□	□	□
- believes that laws and policies should change to reflect the needs of a changing world	□	□	□	□	□	□	□
- I often tries new things just for trying	□	□	□	□	□	□	□
- is sometimes rude to others	□	□	□	□	□	□	□
- is original, comes up with new ideas	□	□	□	□	□	□	□
- worries a lot	□	□	□	□	□	□	□
- prefers to spend time in familiar surroundings	□	□	□	□	□	□	□
- has a forgiving nature	□	□	□	□	□	□	□
- tends to be lazy	□	□	□	□	□	□	□
- is curious about many different things	□	□	□	□	□	□	□
- outgoing, sociable	□	□	□	□	□	□	□
- values artistic, aesthetic experiences	□	□	□	□	□	□	□
- gets nervous easily	□	□	□	□	□	□	□
- believes that religious authorities should be involved when deciding moral issues	□	□	□	□	□	□	□
- does things efficiently	□	□	□	□	□	□	□
- is reserved, quiet	□	□	□	□	□	□	□
- is considerate and kind to almost everyone	□	□	□	□	□	□	□
- has an active imagination	□	□	□	□	□	□	□
- is relaxed, handles stress well	□	□	□	□	□	□	□
- believes that ideals and principles are more important than open-mindedness	□	□	□	□	□	□	□

Group: Player:

8

6. **For each of the following actions or activities, please indicate whether you think that it can always be justified, never be justified, or something in between. You may use any response from 1 to 10 to reflect the strength of your feeling. "1" indicates that it is never justifiable and "10" that it is always justifiable. Make a cross in each of the three rows.**

	Never justifiable **1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	Always justifiable **10**
Claiming government benefits to which you are not entitled…………………………………	□	□	□	□	□	□	□	□	□	□
Avoiding a fare on public transport……….	□	□	□	□	□	□	□	□	□	□
Cheating on taxes if you have a chance..	□	□	□	□	□	□	□	□	□	□

7. **Do you belong to a religion or religious denomination?**

Yes…………………….... □

No …………………….. □

Prefer not to answer ……………….. □

8. **With which one of the following statements do you agree most?**

The basic meaning of religion is:

To follow religious norms and ceremonies……………….. □

To do good for other people……………………………… □

9. **Assume that an essential commodity of daily use gets broken. How easily would you be able to afford 2,000 SEK to replace the commodity within two weeks without having to borrow the money?**

Very easily	Rather easily	Rather difficult	Not at all able
□	□	□	□

Group: Player:

9

10. What is your gender?

Male…………….. □

Female………… □

Prefer not to answer………… □

Thanks for participating in the questionnaire! Please wait for the next instructions.

Group: Player: 10

Objective Scores

Prior to this experiment, we had more than 100 test persons do the same task as you. With these results, we created a catalogue of possible answers.

We used this catalogue to calculate an OBJECTIVE SCORE (with scores between 0 and 10) for each participant in your group taking into consideration the number of answers, their degree of being unusual and the number of different categories they come from.

Important: These OBJECTIVE SCORES are not used to determine the winner. The winner is solely determined by the TOTAL SCORE (composed of the scores given by one's players in the group and one's own assessment of the task). The TOTAL SCORES will not be revealed except for announcing the winner of each group.

You now have the chance to see the OBJECTIVE SCORES we have calculated. It is for information only.

You do not need to look at the OBJECTIVE SCORES, if you don't want to. But you will have to wait until all people are finished. Your choice has, thus, no influence on the time you stay here for the experiment.

Please circle your answer below. "Yes", if you want to see the OBJECTIVE SCORES and "No", if you do not want to see them.

In the case of "No" you will receive an empty piece of paper.

YES **NO**

Group: Player:

11

Choices

We would like to understand how you made your decisions in the experiment. Please tell us why you chose to allocate the points in the way you did.