Journal of Economic Psychology 90 (2022) 102505

Contents lists available at ScienceDirect

Journal of Economic Psychology

journal homepage: www.elsevier.com/locate/joep

A gender bias in reporting expected ranks when performance feedback is at stake☆

Iván Barreda-Tarrazona [a], Aurora García-Gallego [a,b], Jaume García-Segarra [a,*], Alexander Ritschel [c]

[a] LEE & Department of Economics, Universitat Jaume I. Av. Vicent Sos Baynat, s/n 12071 Castelló de la Plana, Spain
[b] ICAE, Universidad Complutense de Madrid, Campus de Somosaguas, 28223 Pozuelo de Alarcón, Spain
[c] Zurich Center for Neuroeconomics (ZNE), Department of Economics, University of Zurich, Blümlisalpstrasse 10, 8006 Zurich, Switzerland

ARTICLE INFO

JEL classification:
C91
D91
J16

Keywords:
Relative ranking
Overplacement
Beliefs
Real-effort task

ABSTRACT

We introduce a mechanism for eliciting beliefs that combines the simple use of monetary incentives with the desire to know the own performance. In our experiment, participants performed a real-effort task that naturally reinforced the desire to know their relative performance. In two treatments, differing in the degree of ex-post transparency, we elicited the belief about one's own standing. In the Baseline, the performance ranking was always revealed. In the Treatment, we deprived the subjects of learning their relative performance if they did not accurately report their actual rank. This simple manipulation creates a bias in behavior that goes in opposite directions for men and women. Under the manipulation, men overplace even more, and women underplace themselves compared to the Baseline.

1. Introduction

Humans feel the need to know their relative position compared to their peers in many contexts to be aware of their status concerning a wide variety of dimensions. On top of the intrinsic motivation for learning where oneself belongs (status), comparison to know the relative ability is a powerful tool to update beliefs that modulates decision-making in many contexts.

This paper focuses on the underpinnings and motives of the subjects for overplacing or underplacing themselves when forming beliefs about their relative position compared to others. Thus, we do not focus on the willingness to compete *per se*. Nevertheless, there is a clear link between the beliefs about one's relative performance and the willingness to compete. Indeed, according to rationality, beliefs about one's own ability and expected performance are the cornerstones over which subjects build their choices when facing some decision-making dilemmas related to the willingness to compete. This way, subjects must apply for a job, promotion, etc., if they think they will beat the competitors. However, the key question we approach in this research is related to the underpinnings of the gender differences in overplacement (according to Moore and Healy (2008), overplacement is the difference between the elicited expected rank and the actual one). More concretely, it is well known that both men and women share a very intense preference for learning their raking after performing specific real effort tasks, but do men and women have a

☆ We thank A. Achtziger, C. Alós-Ferrer, J. Buckenmaier, and S. Hügelschäfer, for helpful comments and discussion. Jaume García-Segarra acknowledges financial support from his research assignation from the University of Cologne, Germany, and grant UJI-B2020-16. This work is supported by the Spanish Ministry of Science, Innovation and University, Spain (grant RTI2018-096927-B-100) and Universitat Jaume I, Spain (grant UJI-B2018-76). We also thank the Associate Editor and two anonymous referees for their thoughtful reports. Data is available at https://osf.io/98vxa/?view_only=bafddbd98cfd43cca978964c681161d6.

* Corresponding author.
E-mail addresses: ivan.barreda@uji.es (I. Barreda-Tarrazona), mgarcia@uji.es (A. García-Gallego), jagarcia@uji.es (J. García-Segarra), alexander.ritschel@econ.uzh.ch (A. Ritschel).

https://doi.org/10.1016/j.joep.2022.102505
Received 18 September 2021; Received in revised form 5 February 2022; Accepted 20 February 2022
Available online 11 March 2022

common motivation for learning this feedback? or alternatively, do men and women have different motives for wanting to know such feedback? (as we conjecture). Those different motives could result in a gender bias in reporting expected ranks when the performance feedback is at stake.

1.1. Related literature

Recent research shows that performance curiosity (the desire to know the own – relative – performance) is a powerful driver that under certain circumstances overcomes inequality aversion (Alós-Ferrer et al., 2018). In the aforementioned paper, subjects performed the well-known real effort task consisting of solving additions of five two-digit numbers for a period of time, introduced by Niederle and Vesterlund (2007). Alós-Ferrer et al. (2018) show that, for every level of actual performance, women report lower beliefs about expected performance than men. The reader may think that subjects, especially those who reported low expectations, can protect their self-image by avoiding performance feedback whenever possible. However, that paper also shows that when the feedback about performance is at stake, the subject's desire to protect their self-image avoiding to know is not a factor since only 3 out of 270 subjects rejected to receive feedback (see appendices A1 and A2 in Alós-Ferrer et al., 2018). Thus, subjects strongly prefer to have feedback independently of how well or poorly they believe they performed. Indeed, in Alós-Ferrer et al. (2018), the incentives were displayed such that every subject in each group of 5 members who was ranked on the 3th, 4th, and 5th position obtained 15%, 10%, and 5% of the joint proceeds (respectively) when implementing the *performance based* distribution (that always came with the ranking feedback). In contrast, they received 20% of the common pot in case of implementing the *equal split* distribution. The only difference across treatments was that in the baseline, subjects obtained their feedback (including the ranking) no matter which distribution they chose, while in the "no info" treatment, the equal split distribution came with no ranking feedback at all). The proportion of people who chose the performance based distribution was significantly higher in the "no info" treatment, even for those who did not expect to perform above the average. This fact reveals that a significant proportion of the subjects tacitly renounced (or were willing to pay) 25%, 50%, or 75% of their rewards in case of being ranked the 3th, 4th, or 5th, just because they wanted to know their feedback. If we convert it to money, they were willing to pay EUR 3.31, EUR 6.62, and EUR 9.93, respectively. Therefore, the findings in Alós-Ferrer et al. (2018) established two points. First, women report lower beliefs about expected performance than men for every level of actual performance. This is in line with many previous studies suggesting that women are less overconfident than men (Barber & Odean, 2001; Beyer, 1990; Beyer & Bowden, 1997; Hackett & Betz, 1981; Hügelschäfer & Achtziger, 2014; Jakobsson, 2012; Ludwig et al., 2017; Reuben et al., 2012). Second, subjects show a strong desire to know their ranking when performing this specific task (a significant fraction of them are even willing to sacrifice from 25% to 75% of their earnings in that experiment to know their ranking jointly with the number of correct sums performed). Then, it is clear that subjects want to know their feedback. What remains unclear are the motives why subjects show this intense desire to know it.

Based on the previous literature, we conjectured that the motives why subjects want to know their ranking might differ between gender. For instance, in an experiment observing trainee truck drivers' behavior, where subjects were roughly 90% males, Burks et al. (2013) report that individuals show a desire to confirm positive beliefs and that they enjoy doing that. The authors state that their design "provides incentives to report one's self-assessment of relative performance truthfully". This point is crucial since they found that truck drivers do not deliberately underplace to feel happier when receiving better feedback than reported. On the contrary, confirming positive beliefs means receiving better feedback compared with what one truly expects. Moreover, relying on a personality profile test, Burks et al. (2013) suggest that social image concerns can be the main driver behind misreported beliefs. They also rule out two existing theories focused on self-image. According to these theories, individuals with optimistic beliefs about their abilities should avoid new information about their absolute or relative performance (Köszegi, 2006; Weinberg, 2009).

Interestingly, this general pattern of behavior (a preference for confirming positive beliefs) reported in Burks et al. (2013) is at odds with many previous studies reporting that women are more pessimistic than men (De Paola et al., 2014; Jacobsen et al., 2014; Küçükaslan & Çelik, 2010), women feel more shame than men regarding overplacing themselves, i.e., about reporting themselves in a higher rank than they actually are (Ludwig et al., 2017), and that women are, in general, more sensitive to social cues and negative feedback than men (Gilligan, 1982; Johnson & Helgeson, 2002; Roberts & Nolen-Hoeksema, 1989). Therefore, all these features do not seem to be compatible with the behavior reported in Burks et al. (2013) and it could be that their findings were biased due to the fact that roughly 90% of the subjects in their trainee truck drivers sample were males. In contrast, according to what has been reported, the women's behavior seems to be more compatible with the opposite behavior, i.e., instead of confirming positive beliefs as truck drivers did, it could be that women prefer to rule out negative beliefs. That is, it could be that women prefer to confirm that they are not worst than they genuinely think.

1.2. The present research

We rely on performance curiosity, i.e., that subjects have a strong preference for knowing their feedback (including their ranking), to introduce a mechanism for eliciting beliefs that combines the simple use of monetary incentives with the desire to know the feedback on their own performance. This mechanism allows us to investigate whether men and women have identical or different motives for knowing such feedback. The idea is that, in the Treatment condition, on top of the pecuniary incentive, we deprive subjects of their performance's feedback (therefore, from some positive utility) if they do not report their expected beliefs closely enough to their actual performance. In contrast, in the Baseline treatment, failing to provide an accurate belief about the actual performance only results in losing the pecuniary incentive of this expected-ranking elicitation but not the feedback about the real effort task.

Theoretically, the economic incentive and the desire to know the feedback in this experiment are aligned in such a way that not revealing the true beliefs is strictly dominated (by revealing actual beliefs). This is because subjects have preferences for earning as much money as possible and for knowing their performance when facing the proposed real effort task. Therefore, by design there is no theoretical trade-off between feedback and economic incentives. However, by revealing the actual beliefs, subjects might risk not being accurate enough, and then, they would lose both the economic incentive for being accurate and, in general, also the information about the feedback (not knowing whether they were not accurate enough because they overplaced or underplaced too much). Note that, in the absence of any cognitive bias or anomalies related with the desire to know the feedback, there should be no difference in behavior across treatments. But the design we propose in this paper allows to go one step further, since we can check whether there are differences in the motives between men and women for changing behavior across treatments.

Regarding our predictions, based on the results reported in Alós-Ferrer et al. (2018) we expect differences in behavior across treatments due to the interaction with the feedback manipulation. Moreover, we also expect (for the reasons explained above) that the value of the feedback for the subjects overcomes the EUR 2 we pick by design to reward the accurate reported expected ranks. Therefore, the subjects would prefer to risk the monetary incentive for being accurate (EUR 2) rather than the possibility to infer *something* about their ranking in case they failed to be accurate enough. But the key point here is that men and women will risk the EUR 2 in a completely different way across gender.

We illustrate this point with the examples of John and Lisa. Imagine John truly expects to be ranked in the 10th position inside a group of 20 participants after performing a real effort task. Imagine that he can also achieve an additional monetary reward of EUR 2 and the feedback about his ranking (we will refer to this situation as the Treatment) if he can provide an expected rank accurate enough (accurate here means that his reported rank is contained inside a window of the 5 positions closer to the actual rank, i.e., if he reports expecting being ranked the 10th, then he would receive the monetary reward and the feedback in case his actual rank is the 8th, 9th, 10th, 11th, or 12th position). Theoretically, if John only focuses on the monetary reward for being accurate, then he should behave identically in the situation described above and in another one in which the feedback about the ranking is going to be disclosed independently of how accurate is the reported expected rank (Baseline). That is, to maximize his chance to obtain the EUR 2, he must report expecting being ranked on the 10th position. However, in the Treatment, John could be deprived of his feedback if he is not accurate enough, and this fact can influence his behavior in the following way. Imagine that John prefers to update his beliefs regarding his rank rather than maximizing his chance of achieving 2 additional euros. If in the Treatment, John naively reports expecting to be ranked on the 10th position, it could happen that at the end of the day, his actual position is better than expected (7th or better rank), or worse than expected (13th or worse rank). In this case, John would have no clue whether he failed to be accurate because he overplaced o underplaced himself too much, and he would end up empty-handed without the reward and without the feedback. As a consequence, he did not only fail to achieve the EUR 2 but also the feedback and *the possibility to infer "something" about his actual rank*. This circumstance would never happen in the Baseline, where John could focus on achieving the monetary reward since the feedback will be revealed independently of how accurate he reports his expected rank. Thus, under the Treatment, John can try to increase the chance of being able to get some information about his actual rank if he behaves strategically focusing on the upper or lower bound of the 5 positions window that allows him to achieve both the monetary reward for being accurate and the feedback regarding his actual position. Based on the behavior described in Burks et al. (2013) (whose experimental subjects were roughly 90% men), John has "a desire to confirm positive beliefs and he enjoys doing that". Then, suppose John strategically misreports his expected rank, revealing that he expects to be ranked on the 8th position when he actually expects to be ranked on the 10th. In that case, he can observe what happened in the position he hoped to achieve and four additional positions ahead. Therefore, if John is better than expected, he would "enjoy" confirming that belief. Thus, if he is lucky and at the end of the day his actual rank is from the 6th to the 10th position, he obtains everything, that is, he achieves the monetary reward, the feedback regarding the ranking, and the joy of confirming positive beliefs. But it would be even more tellingly for him, the fact that if he fails to achieve the feedback, then John can infer that he failed because he overplaced too much, and probably his actual rank is the 11th or worst.

Imagine now that Lisa also truly expects to be ranked on the 10th position, and she faces exactly the very same circumstances described in the example of John above. Imagine that Lisa also prefers to update her beliefs regarding her rank rather than maximizing her chance of achieving 2 additional euros. If in the Treatment she naively reports expecting to be ranked on the 10th position, she might also end up empty-handed without the reward and without the feedback. Thus, Lisa can try to increase the chance of being able to get some information about her actual rank if she focuses on the upper or lower bound of the 5 positions window that allows her to achieve both the monetary reward and the feedback. Contrary to John, who focused on the upper bound of such 5 positions window, and based on the literature reporting that women show a higher fear of failure, are more pessimistic, and tend to be more underconfident than men; we conjectured that women in the Treatment would prefer to focus on the lower bound of such 5 positions window to rule out negative beliefs. That is, women would want to be sure that they do not perform worse than they truly expect. Then, suppose now that, under the Treatment, Lisa strategically misreports her expected rank revealing that she expects to be ranked on the 12th position when she actually expects to be ranked on the 10th. In that case, she can observe what happened in the position she hoped to achieve and four additional positions below. Therefore, if Lisa is not worse than expected, she would "enjoy" confirming that belief. Thus, if at the end of the day her actual rank is from the 10th to the 14th position, she achieves the monetary reward and the feedback. This way, she can check whether her rank is worse than expected. In addition, if she obtains exactly the 10th position, she obtains everything, i.e., the monetary incentive, the feedback, and the joy, because she can rule out being worse than expected. But on top of that, if she fails to be accurate enough, then she can guess that this is because she underplaced too much, and probably her actual rank is the 9th or better, allowing her to hold the beliefs that she is not worse

than expected. This strategic behavior is not needed in the Baseline since she would know her feedback no matter how accurately she reports her expected rank.

In a nutshell, the examples above summarize the research question we approach, the hypotheses, and the design of the experiment to check those hypotheses. The rest of the paper is organized as follows. Section 2 explains the details of the design and procedures. Section 3 shows the results. Finally, in Section 4 we include a discussion of the main findings.

2. Design and procedures

2.1. Design

The experiment was conducted in the Laboratori d'Economia Experimental (LEE) at Universitat Jaume I and programmed in z-Tree (Fischbacher, 2007). Five sessions of 40 subjects were conducted with a total of $N = 200$ participants (104 females). Every session and every treatment was perfectly balanced between males and females, except for one session where four males did not show up, and the session was completed with females from the list of reserves. Participants did not know the gender composition of each group neither the identities of the rest of the players inside the group. This is a relevant point since the gender composition of a group might affect performance and decision making within a group (Apesteguía et al., 2012; Gneezy et al., 2003). The experiment was a between-subjects design with 54 females and 46 males in the control condition and 50 females and 50 males in the treatment condition. Sessions lasted around 90 minutes, and the average payoff was EUR 20.63 (USD 25.5 at the time of the experiment).

2.2. Procedures

Participants were allocated to groups of twenty players (two groups per session, one under each treatment condition). Before starting, participants were informed that the experiment consisted of performing three tasks and answering two questionnaires. The only difference between the control (Baseline) and the treatment condition (Treatment) was in the design of Task 1.

In Task 1, participants were informed that they were assigned to a group with other 19 participants and had to face a real-effort task, in particular, they had to solve five two-digit numbers sums (e.g., 23 + 45 + 62 + 51 + 17= □) for eight minutes. This task was firstly proposed by Niederle and Vesterlund (2007) and has been used in many experiments (e.g., Alós-Ferrer et al., 2018; Azmat & Iriberri, 2016; Ludwig et al., 2017).[1] Subjects could use scratch paper, but not calculator or electronic devices. No feedback about the correctness of the answers was provided until the end of the experiment. After performing the task, participants reported their expected rank.

Each correct sum generated EUR 0.50 to a common pot to be equally shared by all the participants at the end of the session. In addition, a ranking was built based on the correct sums performed by each player. In case of a tie, the computer would randomly tie-break. Additionally, each participant obtained EUR 2 if she was able to predict her ranking accurately enough. The prediction was considered accurate enough if it coincided with the actual position in the ranking or with one of the four closest positions to the actual one. For instance, if a player reported expecting to be in the 10th position, she achieved the additional EUR 2 in case her actual position in the ranking was the 8th, 9th, 10th, 11th, or 12th. If she reported expecting to be 1th, 2th, or 3th position, she would get the extra money in case her actual position in the ranking was 1th, 2th, 3th, 4th, or 5th. Analogously for the other tail at the bottom of the ranking. All three examples were detailed and explicitly explained in the instructions. The ranking served as our information manipulation. In the Baseline this ranking was always revealed at the end of the session. In the Treatment the ranking was only revealed when the participant also managed to predict her rank accurately. Note that the monetary and the informational incentives are intentionally aligned, and participants did not face a trade-off between money and information.

All the rest of the tasks and questionnaires were identical for participants in both conditions (Baseline and Treatment). The experiment continued with Task 2, where participants faced the well-known multiple price list proposed by Holt and Laury (2002) to measure risk attitudes.

After the second task, participants were asked to answer a version of the Achievement Motive Scale (Lang & Fries, 2006) translated into Spanish using a back-translation procedure with three different translators working in parallel and agreeing on the final translation. The AMS-scale consists of two subscales of five items each to measure the constructs of "Hope of Success" and "Fear of Failure". Right after, we ran the NEO-FFI-3 questionnaire proposed by McCrae and Costa (2010) to control for the personality traits analyzed in the Big Five Model, i.e., neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (McCrae & Costa, 1985). The NEO-FFI-3 is a short version of other scales used to measure the Big Five Model, consisting of 60 items, 12 items for each personality trait. We paid EUR 5 to every participant for answering these questionnaires.

Finally, the third task consisted of performing the abstract reasoning part of the Differential Aptitude Test (Bennett et al., 1974) in its Spanish adaptation by Cordero and Corral (2006). In this task, participants have 20 min to answer 40 multiple-choice items to complete a matrix of a logical sequence of figures. The task was incentivized with EUR 0.25 for each correct answer. Thus, the total amount of money participants obtained was the sum of Tasks 1, 2, and 3 plus the additional fixed amount of EUR 5 for answering the questionnaires.

[1] Lundeberg et al. (1994) stated possible gender differences in this task since it belongs to a masculine domain. However, in line with Niederle and Vesterlund (2007) we found no gender difference in the number of correct sums (means: 12.57 for men and 11.71 for women, Mann–Whitney–Wilcoxon test, $N = 200$, $z = 0.668$, $p = 0.504$.) nor in the distributions of these correct sums (Kolmogorov–Smirnov test, $N = 200$, combined $D = 0.113$, $p = 0.547$).

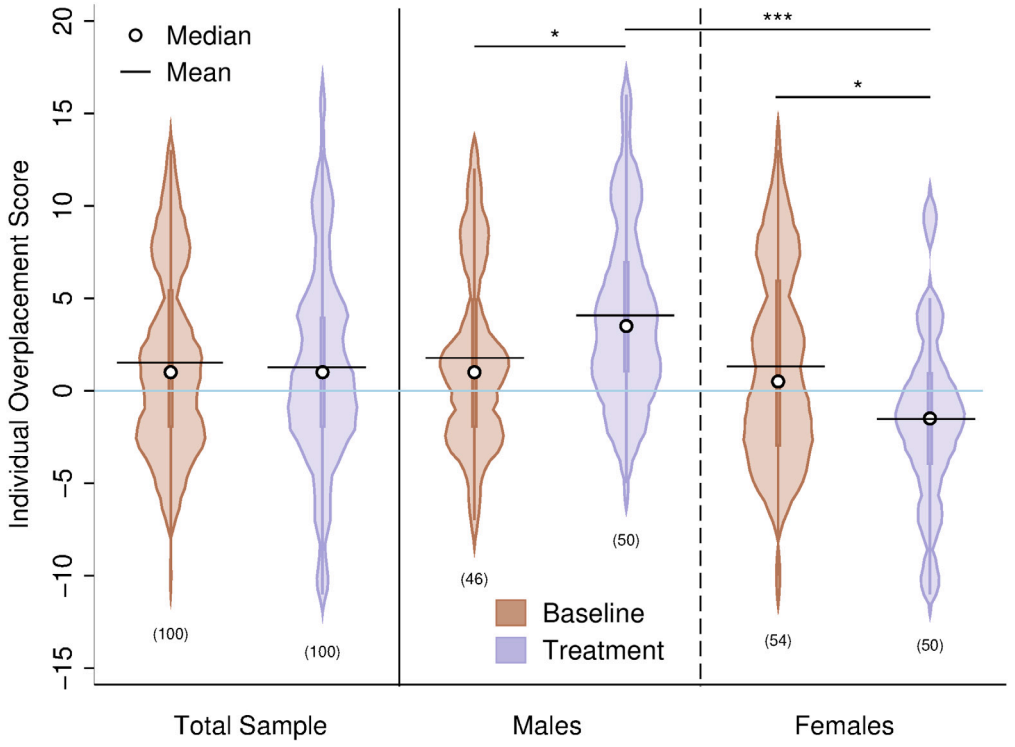

Fig. 1. Violin plots of the Individual Overplacement Score. On the left-hand side, the IOS by treatment. On the right-hand side, the IOS by treatment split by gender. Stars indicate the significance of Mann–Whitney–Wilcoxon tests. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

3. Results

We conducted an ex ante power analysis to determine the sample size using the software G*Power 3.1. Considering the conventional levels of $\alpha = 0.05$ and medium effect size (Cohen's $d = 0.5$) for one-tailed Mann–Whitney–Wilcoxon tests, a priory power above 96% is achieved with 184 observations (92 per condition) that we rounded up to $N = 200$ (100 independent observations in each treatment). Since we were interested in analyzing differences in behavior by gender, we balanced the number of males and females in each group and treatment. Therefore, the analysis of the subsamples of 50 subjects (approx.) still has power above 78%.

3.1. Individual Overplacement Score

We build the Individual Overplacement Score (IOS) to measure how many positions the participants overplaced or underplaced with respect to the actual ranking. IOS is computed as the difference between the actual rank participants achieved and their reported expected rank.

As can be seen in Fig. 1 left-hand side, neither the means nor the distribution of the IOS appear to be significantly different. According to a Mann–Whitney–Wilcoxon test (MWW), we cannot reject the null hypothesis that the medians in the Baseline and the Treatment are equal ($N = 200$, $z = 0.141$, $p = 0.888$). However, this does not imply that our manipulation did not work. Indeed the manipulation creates a significant effect in opposite directions for males and females. Fig. 1, right-hand side, displays the IOS mean and the IOS distribution split by gender in each treatment. On average, both males and females overplace themselves in the Baseline compared to the actual ranking about 1.74 and 1.30 positions, respectively. Interestingly, after introducing our manipulation, males overplace themselves with about 4.04 positions on average, while females underplace themselves with about 1.52 positions. Thus, in the Treatment, males significantly increase their IOS compared to the Baseline (MWW test, $N = 96$, $z = -2.315$, $p = 0.020$), and women significantly shift from showing overplacement in the Baseline to showing underplacement in the Treatment (MWW test, $N = 104$, $z = 2.423$, $p = 0.015$). Note that for those subjects whose actual rank is top 3 and reported a top 3 expected position (4 observations in the Baseline and 7 in the Treatment) or last 3 positions and reported it that way too (only one observation in the whole experiment), the mechanism is not incentive-compatible in the sense that those top 3 subjects get the same monetary payment and same info if they report the 1st, 2nd, or 3rd position (analogously, for these last 3 subjects when reporting the 18th, 19th, or 20th position). Our conclusions and qualitative results remain the same when excluding these observations.

We also observe that the difference in behavior between males and females in the Baseline is not significant (MWW test, $N = 100$, $z = 0.489$, $p = 0.625$) while it is highly significant in the Treatment (MWW test, $N = 100$, $z = 5.165$, $p < 0.001$). This result suggests that the manipulation inspired radically different behavior between males and females.

Table 1
OLS regressions of Individual Overplacement Score.

IOS	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Treatment	−0.2400	−0.3604	2.3009*	2.3383*	2.2871*	2.0155*
	(0.7520)	(0.7219)	(0.9957)	(1.0056)	(1.0247)	(1.0058)
Female		−3.0096***	−0.4428	−0.4315	−0.0333	−0.3799
		(0.7218)	(1.0094)	(1.0151)	(1.0829)	(1.1392)
Treat. × Female			−5.1172***	−5.1623***	−5.0844***	−4.9183***
			(1.3998)	(1.4162)	(1.4181)	(1.4372)
Response Time				0.0140	0.0159	0.0140
				(0.0216)	(0.0212)	(0.0207)
Constant	1.5000**	3.1252***	1.7391*	1.3028	1.9746	−7.1511
	(0.5046)	(0.6344)	(0.7105)	(1.0331)	(4.0302)	(5.4473)
Controls	No	No	No	No	Yes	Yes
Big 5	No	No	No	No	No	Yes
adj. R^2	−0.0045	0.0718	0.1262	0.1235	0.1214	0.1366
LinCom: Treat. +			−2.8163**	−2.8240**	−2.7973**	−2.9028**
Treat. × Female			(0.9840)	(0.9912)	(0.9742)	(1.0180)
LinCom: Female +			−5.5600***	−5.5938***	−5.1177***	−5.2982***
Treat. × Female			(0.9698)	(0.9801)	(1.0274)	(1.0432)
Observations	200	200	200	200	200	200

Standard errors in parentheses.
*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

Regarding the accuracy in the reported expected rank, data show that 54.35% of the males report accurately in the Baseline and only 40.00% in the Treatment. The difference is not statistically significant and only suggests a trend (test of proportions, $N = 96$, $z = 1.407$, one-sided $p = 0.080$). However, only 31.48% of the females report accurately in the Baseline, and 52.00% do so in the Treatment. Therefore, the proportion of females reporting accurately enough is significantly higher in the Treatment (test of proportions, $N = 104$, $z = -2.123$, one-sided $p = 0.017$). This observation is a trivial consequence of our main hypotheses, i.e., humans, in general, are overconfident and tend to overplace, but men usually display this bias with higher strength than women. Therefore, if our hypotheses state that men will overplace even more, and women will underplace under our manipulation, then it is not surprising that, as a consequence, the manipulation also produces a higher proportion of accurate predictions only for women.

Finally, we find that women scored significantly higher (13.60) on the Fear of Failure scale than men (10.89; MWW test, $N = 200$, $z = 5.568$, $p < 0.001$). This result remains robust and significant when analyzing the gender differences in each treatment. In addition, there are no statistically significant treatment differences for each gender. At the same time, there is generally no statistically significant difference in Fear of Failure scores between the Baseline (12.44) and the Treatment (12.15; MWW test, $N = 200$, $z = 0.695$, $p = 0.487$). Hope of Success and risk aversion according to the multiple price list (Holt & Laury, 2002) do not show any statistically significant difference between treatments or gender.

3.2. Regression analysis

To complement our non-parametric analysis we performed OLS regressions. Table 1 shows the regression models. The regression models focus on the effect of the Treatment, gender, and its interaction on the average IOS while controlling for other individual characteristics collected during the experiment.

Model 1 simply introduces the Treatment dummy, which is not statistically significant. This reflects our previous finding that in the pooled sample (i.e., if we do not split data by gender) it seems that the Treatment does not affect behavior. Model 2 introduces the Female dummy, which is negative and highly significant. This indicates that generally, women have a lower IOS score than men. The Treatment dummy remains statistically non-significant. Model 3 adds an interaction term between Treatment and Gender (Treat. × Female). This interaction changes the interpretation of the Treatment and Female dummy. The interaction term itself is negative and highly significant, which means that the treatment effect is more intense for women (in the sense that the coefficient for women is negative, resulting in a much lower IOS score) than for men. However, how the treatment affects each gender can be seen by the Treatment dummy and the linear combination of the Treatment dummy and the interaction term. The Treatment dummy by itself represents the treatment effect for men. The dummy is positive and significant, which means that men overstated their rank much more in the Treatment than in the Baseline. The treatment effect for women can be seen by a linear combination test of Treatment and the Treatment and Gender interaction term (Treat. + Treat.×Female). The bottom of Table 1 shows this linear combination test which is negative and highly significant, indicating that women underplaced themselves much more in the Treatment than in the Baseline. Therefore, we confirm our non-parametric results and show that the Treatment shifts the stated rank in opposite directions for women and men.

In addition to analyzing the treatment effect, we can also compare the differences between women and men within the Baseline or Treatment. The Female dummy by itself (and after introducing the interaction term) represents the difference between women and men in the Baseline. The dummy is not significant suggesting that there might be no statistically significant difference between

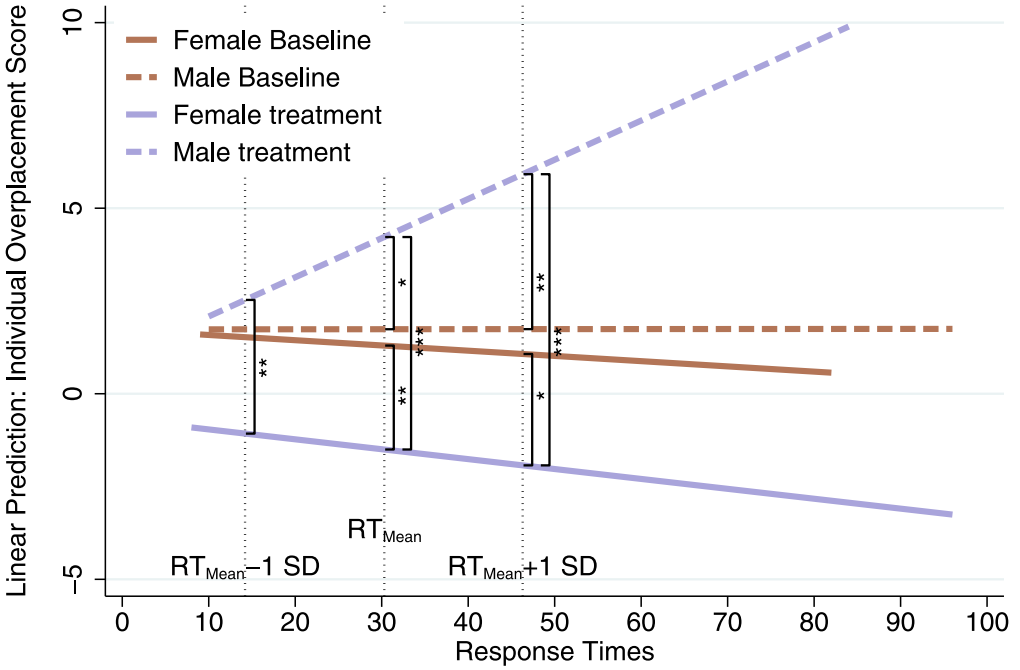

Fig. 2. Linear prediction of IOS on Response Times (RT). Linear combination tests at three different points in time: mean RT, mean RT-1 Standard Deviation (SD), mean RT+1 SD. Stars indicate the significance of Linear Combination tests. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

genders in the overplacement score. The linear combination test Female + Treat. × Female shows the difference between women and men in the Treatment. The linear combination is negative and highly significant, indicating that women have a much lower IOS score than men in the Treatment.

Models 4 to 6 check our results for robustness while controlling for multiple other variables collected during the experiments. Model 4 controls for response times. Our previous results remain robust. Model 5 controls for the number of safe lotteries chosen in the Holt and Laury multiple price list (a proxy for risk aversion), Hope of Success, and Fear of Failure score. Our results remain robust. Model 6 adds the individual scores from the Big 5 questionnaire (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness). Again, our results remain robust and our conclusions are the same.

In summary, the regression models indicate a significant treatment effect, leading to an increase for men and a decrease for women in the IOS score.

3.3. Strategic reasoning and decision times

The fact that subjects fall into strategic reasoning should increase the time required to report their expected ranking. This is why we recorded the participants' decision times when making the actual decision about reporting their expected rank. We intended to use decision times to indicate the strength of preferences between choices when choosing their exact expected position. In this way, we should observe a higher overplacement for males and higher underplacement for women when the response time is longer.

Fig. 2 displays the linear prediction of the overplacement score on the vertical axis and the response time on the horizontal axis. The figure clearly shows that the slope in the Baseline is flat and, according to a Linear Combination test, non-significantly different from 0 (Linear Combination (LinCom) test; men Baseline, $\hat{\beta} = 0.0001$, $p = 0.998$; women Baseline, $\hat{\beta} = -0.0141$, $p = 0.746$). Whereas the slope for men in the Treatment is positive and significant (LinCom test, $\hat{\beta} = 0.1056$, $p = 0.021$). For women in the Treatment, the slope is not significantly different from zero (LinCom test, $\hat{\beta} = -0.0267$, $p = 0.529$), however, Fig. 2 reveals a downward shift. The figure also indicates the mean response time (30.27 s) in our experiment and also the mean response time plus/minus one standard deviation from the mean (14.22 s and 46.31 s). At the mean response time, we find a significant difference for the predicted IOS between Baseline and Treatment for men (LinCom test, $\hat{\beta} = -2.484$, $p = 0.015$) and women (LinCom test, $\hat{\beta} = 2.800$, $p = 0.004$). There is also a highly significant difference for the predicted IOS between men and women in the Treatment (LinCom test, $\hat{\beta} = 5.724$, $p > 0.001$). We find similar results for the later time point mean+1SD (LinCom tests; Baseline vs. Treatment: men $\hat{\beta} = -4.175$, $p = 0.004$; women $\hat{\beta} = -3.001$, $p = 0.028$; men vs. women in the Treatment: $\hat{\beta} = 7.845$, $p > 0.001$). At time point mean-1SD, there are no statistically significant differences between the Baseline and Treatment neither for men (LinCom test, $\hat{\beta} = 0.791$, $p = 0.576$) nor women (LinCom test, $\hat{\beta} = 2.598$, $p = 0.063$), however, we still find a significant difference for the predicted IOS between

men and women in the Treatment (LinCom test, $\hat{\beta} = 3.602$, $p = 0.010$).[2] Therefore, the linear prediction of response times on the Individual Overplacement Score shows that, in the Treatment, the longer time it takes to report the expected rank, the higher is the overplacement of men and underplacement of women. This fact provides additional support to our rationale about the opposite strategic misreporting (in the Treatment) of men and women in line with our hypothesis.

4. Discussion

The reader might notice that given the incentives scheme, if we only focus on the monetary part of the incentives, a subject might be tempted to free-ride on the rest of the participants by not solving a single correct sum. This way, she would still collect the money coming from her group-mates effort, on top of assuring a high probability of obtaining two additional euros for being accurate in reporting an expected rank at the bottom of the distribution. But our experiment deals with something else than money. We offer subjects the chance to update beliefs about their actual ability in performing a task where previous research (Alós-Ferrer et al., 2018) already established that they want to know their feedback. Data show that we had only one subject in the whole experiment who predicted to obtain one of the last 3 positions, and her performance matched her reported expected rank. Indeed, this subject answered 4 correct sums, and no single subject, out of 200, solved 0 correct sums. Other 4 subjects in the whole experiment solved less than 4 correct sums (2 observations in the Baseline and 2 in the Treatment, but they expected the 10th, 12th, 15th, and 16th rank, respectively.) Hence, the combination of incentives worked as intended (i.e., no subject decided to free-ride on the rest of group-mates) because, as we learned in Alós-Ferrer et al. (2018), people have an intrinsic motivation to know their performance in this task. Therefore, if a subject would free-ride on her group-mates to significantly improve her chance to obtain 2 additional euros, this movement would come at the cost of not knowing her actual ability to perform the mentioned task (both in the Baseline and in the Treatment).

This experiment shows that a simple manipulation, i.e., depriving subjects of learning their actual relative performance if they are not accurately guessing their rank, is enough to affect elicited beliefs. The manipulation introduced in this experiment induces men to report higher expected ranks in the Treatment than in the Baseline. In contrast, this very same manipulation induces the opposite effect in women, i.e., they report lower expected ranks in the Treatment than in the Baseline. Thus, the manipulation impacts the generalized overplacement bias in opposite directions for women and men. The overplacement is strengthened for men, but completely disappears for women (indeed, women show underplacement in the Treatment). Data also suggest that subjects possibly stated their belief strategically.

A potential explanation for these findings would be that, on the one hand, men strategically overplace themselves to know whether they are better than they truly expected. In contrast, on the other hand, women strategically underplace themselves to know the opposite, i.e., whether they are worse than they truly expected. A policy implication of this study is that an accurate self-assessment among peers will have to be carefully designed, accounting for the gender of the participants.

Regarding potential limitations of our study, the reader might think that since we measure the IOS inside groups of 20 subjects, potential different sample assumptions might induce different beliefs motivating differences in behavior. This would yield an alternative explanation of our results. However, by design, we randomly assigned subjects to two groups of 20 participants (40 per session) in a between-subject design. Every session in every treatment contained homogeneous samples of students majorly balanced across gender. Thus, everyone faced very similar circumstances, observing a group of similar age university students with a mostly balanced composition of men and women. Therefore, there should not be objective reasons to develop different "priors" or ex-ante beliefs due to the design of the experiment. On top of that, we ran the experiment in a big lab, and we had two groups in each session. Then, if by chance a subject recognizes an extremely skillful/unskillful participant on the way to the lab (for instance, some other known student from the same faculty), even this issue would be diluted by the fact that there is a 50% chance that such subject is assigned to the other group. Thus, it is hard to justify that the results can be interpreted as a confound generated by different "beliefs" due to the different samples instead of by the rationale explained in Section 1 and illustrated by the examples of John and Lisa.

Finally, our design intentionally allows for this strategic behavior since we hypothesize that men and women would use different strategies to update beliefs in the Treatment. Further avenues of research should help measure which part of this behavior is merely strategic and which part is because when subjects are asked to think about their rank, suddenly, men genuinely become more optimistic and women more pessimistic.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.joep.2022.102505.

[2] We thank an anonymous reviewer for suggesting this analysis.

References

Alós-Ferrer, C., García-Segarra, J., & Ritschel, A. (2018). Performance curiosity. *Journal of Economic Psychology*, *64*, 1–17.

Apesteguía, J., Azmat, G., & Iriberri, N. (2012). The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science*, *58*(1), 78–93.

Azmat, G., & Iriberri, N. (2016). The provision of relative performance feedback: An analysis of performance and satisfaction. *Journal of Economics & Management Strategy*, *25*(1), 77–110.

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics*, *116*(1), 261–292.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1974). *DAT: Differential Aptitude Test*. New York, NY: The Psychological Corporation.

Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology*, *59*(5), 960–970.

Beyer, S., & Bowden, E. M. (1997). Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, *23*(2), 157–172.

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2013). Overconfidence and social signalling. *Review of Economic Studies*, *80*(3), 949–983.

Cordero, A., & Corral, S. (2006). *DAT–5 tests de aptitudes diferenciales*. Madrid: TEA Ediciones.

De Paola, M., Gioia, F., & Scoppa, V. (2014). Overconfidence, omens and gender heterogeneity: Results from a field experiment. *Journal of Economic Psychology*, *45*, 237–252.

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

Gilligan, C. (1982). *In a different voice*. Harvard University Press.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, *118*(3), 1049–1074.

Hackett, G., & Betz, N. E. (1981). A self-efficacy approach to the career development of women. *Journal of Vocational Behavior*, *18*(3), 326–339.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Hügelschäfer, S., & Achtziger, A. (2014). On confident men and rational women: It's all on your mind (set). *Journal of Economic Psychology*, *41*, 31–44.

Jacobsen, B., Lee, J. B., Marquering, W., & Zhang, C. Y. (2014). Gender differences in optimism and asset allocation. *Journal of Economic Behaviour and Organization*, *107*, 630–651.

Jakobsson, N. (2012). Gender and confidence: are women underconfident? *Applied Economics Letters*, *19*(11), 1057–1059.

Johnson, M., & Helgeson, V. S. (2002). Sex differences in response to evaluative feedback: A field study. *Psychology of Women Quarterly*, *26*(3), 242–251.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, *4*(4), 673–707.

Küçükaslan, A., & Çelik, S. (2010). Women feel more pessimistic than men: Empirical evidence from turkish consumer confidence index. *Journal of Business Economics and Management*, *11*(1), 146–171.

Lang, J. W., & Fries, S. (2006). A revised 10-item version of the achievement motives scale: Psychometric properties in german-speaking samples. *European Journal of Psychological Assessment*, *22*(3), 216–224.

Ludwig, S., Fellner-Röhling, G., & Thoma, C. (2017). Do women have more shame than men? An experiment on self-assessment and the shame of overestimating oneself. *European Economic Review*, *92*, 31–46.

Lundeberg, M. A., Fox, P. W., & Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments.. *Journal of Educational Psychology*, *86*(1), 114–121.

McCrae, R. R., & Costa, P. T. (1985). *The NEO personality inventory manual*. Odessa, FL: Psychological Assessment Resources.

McCrae, R. R., & Costa, P. T. (2010). *NEO inventories for the neo personality inventory-3 (NEO-PI-3), NEO five-factor inventory-3 (NEO-FFI-3), NEO personality inventory-revised (NEO PI-R): professional manual*. PAR.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

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

Reuben, E., Rey-Biel, P., Sapienza, P., & Zingales, L. (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behaviour and Organization*, *83*(1), 111–117.

Roberts, T.-A., & Nolen-Hoeksema, S. (1989). Sex differences in reactions to evaluative feedback. *Sex Roles*, *21*(11–12), 725–747.

Weinberg, B. A. (2009). A model of overconfidence. *Pacific Economic Review*, *14*(4), 502–515.