HEIDELBERG UNIVERSITY
DEPARTMENT OF ECONOMICS

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Against the Standard

Maria Cubel

Santiago Sanchez-Pages

Christiane Schwieren

Cosima-Valerie Steck

AWI DISCUSSION PAPER SERIES NO. 764
July 2025

Against the Standard[1]

Maria Cubel[2], Santiago Sanchez-Pages[3], Christiane Schwieren[4], Cosima-Valerie Steck[5]

Abstract:

Extensive research has documented gender differences in the willingness to compete against others. Less attention has been given to situations where individuals must meet a standard of excellence to obtain rewards, such as promotions, grants, and publications. This paper investigates gender differences in competing against such standards through a laboratory experiment. Participants completed two rounds of a multiple-choice test. After the first round, they received feedback on whether they met a top-quartile performance threshold set by a reference group. Before the second round, they had to choose between a piece rate payment or a higher rate contingent upon surpassing the threshold. We compared choices across a control treatment with no feedback and three feedback conditions with varying standards: objective peer performance, peer expectations, and expert expectations. Results show that without feedback, women are less likely than men to benchmark against the standard. Feedback closes this gap when the standard is set by peers, but not when set by experts. A theoretical model and an out-of-experiment study suggest these differences stem from gendered priors about ability and asymmetric belief updating. These findings offer insights into gender differences in self-promotion and suggest ways feedback might mitigate these differences.

JEL Classification: C91, D91, J16

Keywords: gender, competitive behavior, experiment, information provision

[1] We thank Larbi Alaoui, Libertad Gonzalez, Gianmarco León, Humberto Llavador, Rosemarie Nagel, Paola Profeta, Ana Tur Prats for their help and support, and seminar audiences at ASFEE Lyon, AXA Gender Lab Bocconi, Bath, EDGE network, ESA Bologna, Exeter, HeiKaMax, Jornadas de Economia Laboral, Kiel, SAE 2024, Sheffield, Trier, UAB, and the Workshop on Gender and Work for their comments.
[2] Department of Economics, City University London. E-mail: maria.cubel@gmail.com. URL: https://sites.google.com/site/mariacubel/home
[3] Department of Political Economy, King's College London. E-mail: sanchez.pages@gmail.com. URL: http://www.sanchezpages.com/
[4] Alfred Weber Institute for Economics, University of Heidelberg. E-mail: christiane.schwieren@awi.uni-heidelberg.de
[5] Alfred Weber Institute for Economics, University of Heidelberg. E-mail: cosima-valerie.steck@uni-heidelberg.de

1. Introduction

Gender differences in labor market outcomes persist despite extensive research and policy interventions. Over 150 years after a letter to the editor in the New York Times discussed the unfairness of gender pay differences for the first time, gender gaps remain in wages (e.g. Blau & Kahn, 2020), pensions (e.g. Hammerschmid & Rowold, 2019), representation in leadership positions (e.g., Eckel et al., 2021), patents (Rosser, 2013), and high-visibility publications (Sá et al., 2020), to name just a few areas.

While much attention has focused on interpersonal competition as an explanatory factor for these gaps, individuals often face situations where career advancement depends on meeting standards of excellence. Examples include college and grant applications, awards, performance reviews, and academic journal submissions. In these settings, candidates must evaluate themselves and decide whether they believe they can surpass a standard set by others, often experts, rather than directly competing against peers. This type of competition has received comparatively less attention in the economics literature.

Competition against standards differs from interpersonal competition in several important ways. First, passing the threshold guarantees the reward, eliminating the personal aspect of competition and interpersonal rankings. Second, the source of the standard, whether it comes from peers or experts, may significantly influence how individuals evaluate themselves relative to that standard and how they interpret feedback about their performance.

The feedback participants receive in these settings is often indirect. They simply learn whether they passed or not, rather than receiving precise information about their relative standing. This indirect feedback can significantly affect subsequent choices, such as decisions to reapply or to pursue similar opportunities in the future. In many competitive situations, the first step is a decision whether to submit an application at all. This decision can be framed as contemplating whether one actually performs above the "standard" to be considered any further. If more men than women consider themselves to meet the standards required for labor market rewards, the sample of men participating in these competitions will generally be larger than the sample of women, leading to an overrepresentation of men among those who win, beyond any starting gender differences in ability.

This paper uses a laboratory experiment to investigate how gender differences manifest in competition against standards of excellence and how feedback influences these differences.

Participants completed two rounds of a multiple-choice test on principles of economics. They received feedback after the first round on whether they met a top-quartile performance threshold set by a reference group. Before the second round, they chose between a piece rate per correct answer and a scheme where the piece rate quadrupled if they surpassed the threshold, with no payment otherwise. We compared choices across a control treatment with no feedback and three feedback conditions with varying standards: the top quartile performance threshold in a previous group of peers, the top quartile performance threshold expected by peers, and the threshold expected by a group of experts, a group of professors from the same university as the participants in this case.

Our results show that women are less likely than men to benchmark their performance against the standard in the absence of feedback. Feedback closes the gap when the standard is set by peers, either objectively or as an expectation. However, the gap re-emerges when experts set the standard. Women are more likely than men to under-benchmark their performance, opting for the piece rate scheme despite having met the standard. This phenomenon is particularly pronounced in the expert treatment, where women who met the standard are as likely to benchmark their performance as those who did not.

To better understand the mechanisms driving these findings, we develop a theoretical model of belief updating and conduct an additional out-of-experiment study with the same population. They suggest that the gender differences observed in the main experiment stem from differential priors about ability in the task rather than perceptions about the strictness of standards. A second driver is asymmetric belief updating; women might be discounting positive expert signals, possibly due to perceptions of gender bias in expert evaluation or attributing success to luck rather than ability.

Taken together, our findings offer insights into gender differences in self-promotion and suggest ways feedback might help mitigate these differences. By implementing clear, objective performance information from peer-based standards, institutions and organizations can reduce gender gaps in labor market outcomes and promote the advancement of high-performing women, ensuring they attain key career milestones at rates comparable to men.

The remainder of the paper is structured as follows. In Section 2, we present the related literature. Section 3 describes our experimental design and hypotheses. Next, Section 4 presents our main results. Section 5 discusses the mechanisms behind our results, and Section 6 concludes.

2. Related Literature

Behavioral economists have identified gender differences in competition as a key explanatory factor for persistent labor market gaps (e.g., Niederle and Vesterlund, 2007, Niederle, 2016, Schram et al., 2019, Lozano et al, 2022, Backus et al., 2023). This literature typically employs some variant of the experimental design from the seminal paper by Niederle and Vesterlund (2007). In this setup, players usually start out with a real-effort task with piece-rate incentives, followed by the same task in some kind of competitive setting. In a third round, players choose whether they want to perform under piece-rate or competitive incentives. Often, this is followed by belief elicitation about one's own relative performance. This design allows researchers to observe both the choice of incentive scheme (competition vs piece-rate) and the performance effects of playing under tournament incentives. Additionally, it allows linking choices to beliefs so that confidence can be assessed.

This approach shows that women generally choose competitive incentives less often than men, even when they could gain from doing so, while men tend to be overconfident and choose competition too frequently. This gap varies depending on the task and the precise setting. Gender-homogeneous settings reduce gender differences, while age and culture have limited and mixed effects. The reasons given for women's lower participation in tournaments are manifold, ranging from risk aversion and general dislike of competition to lower confidence and feedback aversion (Lozano et al., 2022).

Most of these factors are specific to interpersonal competition. Women tend to avoid direct competition, especially with men, when the gender of the opponent is known (Datta, Gupta, Poulsen and Villeval, 2005), and respond differently to status rankings (Schram et al., 2019, Brandts et al., 2020). However, many career-advancing opportunities involve competition against standards rather than direct interpersonal competition, a distinction central to our study.

The closest paper to our approach is Coffman et al. (2024a), who conduct both a field experiment and an online experiment to investigate gender differences in the willingness to apply for higher-return and more challenging work. Participants choose between two tasks of similar nature, but one is more demanding and offers higher rewards. The authors introduce three experimental treatments in which they vary the information provided to participants regarding the selection cutoff for the more demanding task, referred to as "the bar". They observe that in the absence of information, women are less likely than equally qualified men to apply for the high-return task.

In their design, clearing the bar is merely an entry requirement for a subsequent interpersonal competition. In contrast, in our setting, it guarantees rewards, thus eliminating any interpersonal competition and creating a purer test of willingness to compete against standards of excellence.

Booth and Nolen (2022) also study competition against standards. Subjects in their experiment are confronted with a known "target" which they must exceed to get paid. The design varies the source of this target: either it is anonymously set via an unknown mechanism or based on prior performance of anonymous others of either known or unknown gender. Different from our setting, the performance of one other participant in the same laboratory session sets the target, making it a *de facto* interpersonal competition. Our approach differs by using standards set by either previous participants or experts with no direct competitive relation to subjects.

Exley and Kessler (2022) study the role of subjective assessments in competitive settings. They include performance feedback of varying quality to see whether the availability of objective information improves participants' estimation of their own performance. They find a gender gap in self-evaluation, especially when the aim is self-promotion (to a possible employer) and in stereotypically male tasks. When participants are informed about their performance both in absolute and relative terms, the gender gap persists. In our study, the evaluation is based on a quantitative scale. We also examine how the source of standards (peer vs. expert) influences competitive choices beyond the mere presence of performance information.

When it comes to the decision whether to enter a competition or not, performance feedback has been studied in a variety of settings as a way to improve entry decisions, both for underconfident women and overconfident men. The evidence is mixed. Some studies find that feedback helps women to improve their entry decisions (e.g. Ertac and Szentes, 2011, Wozniak, 2012). Others find that women internalize negative feedback more than men do (Berlin and Dargnies, 2016), suggesting asymmetric updating processes. Brandts et al. (2015) provide advice, not feedback, from participants from previous rounds of the experiment, and induce strongly performing women to enter more, while reducing the entry of weakly performing men and intermediately performing women; as a result, the entry gap does not reduce substantially.

3. Experimental design

We conducted a laboratory experiment in which participants had to solve a task under a *piece rate* scheme first. Participants were then asked to select under which payment scheme they

wanted to perform the task a second time, under piece rate or under an *excellence* scheme. In the excellence scheme, participants whose performance was above a given standard of excellence were paid at a quadruple rate, and zero otherwise. We implemented several treatments (details below), that varied the nature of the standard of excellence and whether participants received feedback after the first round of the task.

The task in our experiment was to answer multiple-choice questions on basic economic principles. We compiled a question bank consisting of 360 items sourced from basic economics textbooks used in high school and introductory university courses in Spain, where the experiment took place. The questions presented to the participants were randomly selected from this bank (without replacement). Each question had four possible answers, with only one of them being correct. Participants had five minutes to answer as many questions as they could.

We selected this task because it requires expertise and knowledge. It is relevant to academic and professional settings, where hierarchies of knowledge exist, and the pursuit of excellence is integral. By using a task in which success demands knowledge, we can approximate how individuals approach and strive for excellence within these contexts.

The experiment was conducted at the Behavioral Experimental Sciences Laboratory (BESLab) of the Universitat Pompeu Fabra (UPF) using standard recruiting procedures. From the BESLab subject pool, we contacted all students, both undergraduate and postgraduate, who had taken a course on Principles of Economics. From this pool, 293 individuals participated in the experiment; 53.6% of them were women. This sample comprised students in Economics (34%), Business Administration (33%), International Business (19%), Business Science (8%) and other degrees (6%). There was a total of 30 experimental sessions. They took place during the Fall of 2019 and the Fall of 2021 as BESLab was closed in the interim period due to COVID. Gender was not discussed at any time during the sessions. Each participant received a 3€ show-up fee. Participants were told that they would be asked to complete two tasks. In addition, they were told they would be asked a series of questions about their beliefs and that all these questions would be incentivised. The experiment was implemented through computer terminals using zTree (Fischbacher, 2007).

Figure 1: Order of tasks and payment choices.

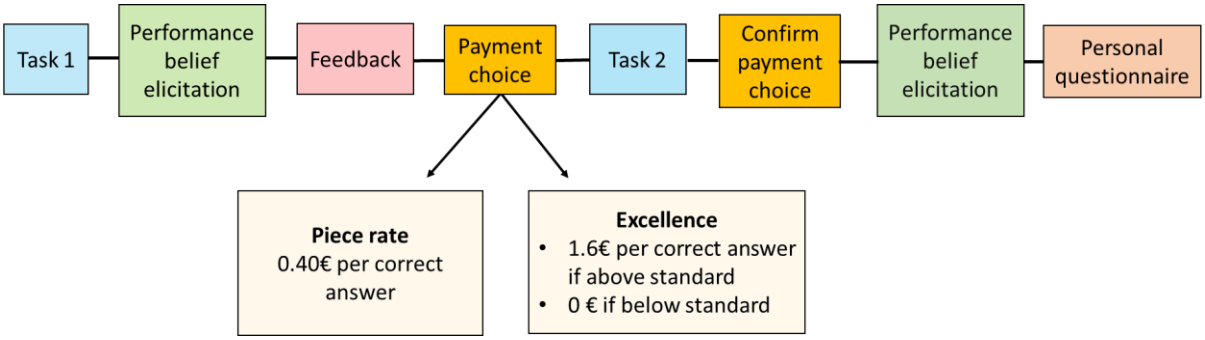


Figure 1 summarizes the timing of the experiment. Participants performed a first round of the task (*Task 1*) that we employ as a measure of ability. Then, they had to state whether they believed their performance was above or below the standard of excellence, and which quartile they thought their performance fell into within their session. They received 50 eurocents per correct answer to these questions. In the feedback treatments (details below), participants were informed of their relative performance in Task 1, i.e. whether they had met or not the standard, but they were not informed about the actual standard, i.e. the number of correct answers defining an excellent performance, nor about their absolute performance, i.e., the number of questions they answered correctly, until the end of the experiment.

We deliberately withheld information about the actual threshold and participants' absolute performance to mirror real-world competitive contexts, where standards are often opaque. In many professional settings, individuals receive only binary feedback (accepted/rejected, passed/failed) without knowing how far above or below the standard they performed. Potential applicants typically have only a vague sense of whether the standard is low or high, and their interpretation of success or failure depends on their beliefs about the difficulty of passing. Furthermore, these interpretations are often influenced by the perceived authority of those setting the standards. By providing only the source of the standard and binary feedback, our design captures this uncertainty while allowing us to observe how participants' decisions respond to standards from different sources.

Before undertaking the second round of the task (*Task 2*), participants had to choose whether they preferred to perform under the piece rate or the excellence scheme. In the former case, they were paid 40 eurocents per correct answer. In the latter, they were paid 1.60€ per correct answer if their performance in Task 2 was equal to or above the standard, and zero otherwise. After

taking the second round of the task, participants had to confirm their payment choice; they could change their initial choice by paying 40 eurocents.[6] Before filling a post-experiment personal questionnaire, participants were once again asked about the quartile they believed their performance in Task 2 occupied within their session.

The standard of excellence was the threshold defining the top quartile performance in the task, i.e., an "excellent" performance, as determined by a group of reference. The payoffs we set per correct answer imply that a participant with a 25% chance of beating the standard, i.e. being in the top quartile according to the group of reference, had the same expected payoff under the excellence payment scheme and under piece rate. We implemented four treatments that varied the group of reference and the availability of feedback. Participants were randomly allocated to one of these four treatments:

Control treatment: The standard of excellence in this treatment was defined as the top quartile performance threshold of a previous group of peers, that is, a group of subjects drawn from the same population and who undertook the same task under the same incentives in a pilot session. The other main feature of this treatment is that participants did not receive any feedback after performing Task 1. In other words, they did not know whether their performance in Task 1 had met or fallen short of the standard at the point of making their first payment scheme choice and performing in Task 2.

Objective treatment: The only difference with respect to *Control* was that participants in this treatment were told whether their performance in Task 1 was above or below the standard. This standard was again determined as the top quartile performance threshold of a previous group of participants. With this treatment, we could thus study whether participants of similar ability but different gender adjusted their payment choices differently after receiving the same feedback, and whether there were any gender differences in how participants responded to feedback (being above or below the standard in Task 1).

Peer treatment: The standard of excellence in this treatment was the performance threshold that a previous group of peers believed defined the top quartile. We asked participants in the pilot session what number of correct answers they believed corresponded to the top 25% performance in that session. The standard of excellence in the *Peer* treatment was the median guess within the

[6] Only 16.7% of subjects used this option; 85.8% of those changed their first choice from excellence to piece rate.

pilot group[7] (8 correct answers)[8]. With this treatment we can then study whether men and women made different payment choices when the standard of excellence was socially rather than objectively defined.

Expert treatment: The standard was once again defined as an expectation about what constitutes an excellent (top quartile) performance. The group of reference in this treatment was drawn from the Economics faculty at UPF. We consulted five professors (three male, two female) who are familiar with the Principles of Economics course taught at that university. They were asked to estimate the number of correct answers that would define a top 25% performance in the task. We used the median of their estimates as the standard of excellence for this treatment. These experts set a slightly higher standard than students in the pilot group (9 vs. 8 correct answers), but they were correct. Importantly, recall that we did not inform participants of the actual standard nor their own number of correct answers. Therefore, this treatment allows us to study whether male and female participants adjusted their payment choices differentially when the expectation of excellence was established by a group with superior rather than similar expertise to their own.

The post-experiment questionnaire included some demographic questions alongside the assessment of relevant factors such as risk tolerance, subject of study and seniority in the degree, grade achieved in the Principles of Economics course, and self-efficacy (e.g., Bandura, 1977; Scherbaum *et al.*, 2006), which relates to effort, motivation, and goal setting in tasks.[9] The experimental sessions lasted about 45 minutes. Total earnings averaged 10.30€.

4. Experimental results

In this section, we study whether, conditional on their performance and the type of standard, male and female participants differed in their payment scheme choice, piece rate versus excellence. We first explore whether there were any gender differences in performance in Task 1, which we take as a measure of ability in the task, and in beliefs about their level of performance relative to the standard and participants in their session. We then examine the participants' payment scheme

[7] The pilot group was composed by 15 students of which 12 were women. The average number of correct answers was 6.73, similar to 6.71 in the main study (Mann-Whitney test, exact p-value = 0.811)..

[8] This guess was slightly lower than the correct top 25% performance threshold: 9 correct answers.

[9] We implemented the Spanish version of the self-efficacy inventory as developed by Sanjuan-Suarez *et al.* (2000).

choices and whether they differed conditional on the standard of excellence and their performance in the first round of the task.

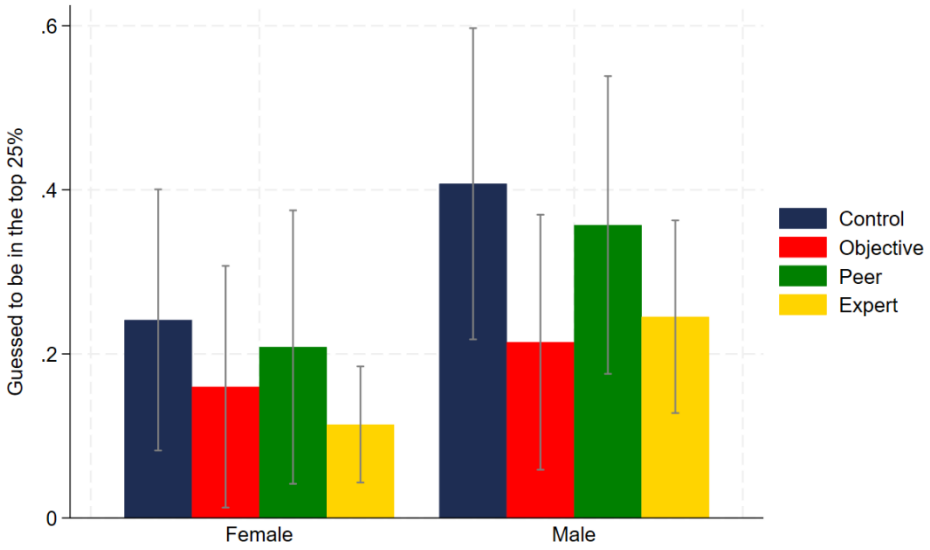
Performance and beliefs in Task 1

The average number of correct answers was 6.26 for women and 7.23 for men. A Mann–Whitney test detected a significant gender difference in average performance in Task 1 ($p = 0.005$). This is not something that we expected, nor did it show up in either the pilot study or the out-of-experiment study discussed in Section 5.

Because this difference in performance might have been driven by factors which correlated with gender, such as field of study or risk tolerance, we ran an OLS analysis with performance in Task 1 as dependent variable (see first column of Table A1 in appendix). We controlled for participant characteristics such as risk tolerance, academic background, seniority, and grade achieved in the Principles of Economics course. Estimates show that, on average, women answered 0.8 fewer questions correctly than men. This gender difference was only weakly significant after accounting for the aforementioned factors. It disappeared entirely when we applied the same analysis to performance in Task 2, now controlling for performance in Task 1 and initial payment scheme choice (fourth column of Table A1).

After performing Task 1, we asked participants whether they believed their performance was above the standard of excellence in their treatment. As Figure 2 below shows, men across all treatments stated more optimistic beliefs than women, measured by the fraction of them who expected to have passed the standard. These differences were not significant except in the *Expert* treatment (proportion test, $p = 0.0471$). This difference vanished when we ran a linear probability model with the belief of having beaten the standard as dependent variable and controlling for relevant characteristics (second column in Table A1). That said, men had less accurate guesses than females. A 36.0% of male participants versus a 26.1% of females incorrectly guessed whether their performance in Task 1 had been above or below the standard; this difference in accuracy was significant (one-sided proportion test, $p = 0.033$). As described below, males were less accurate because they were overoptimistic.

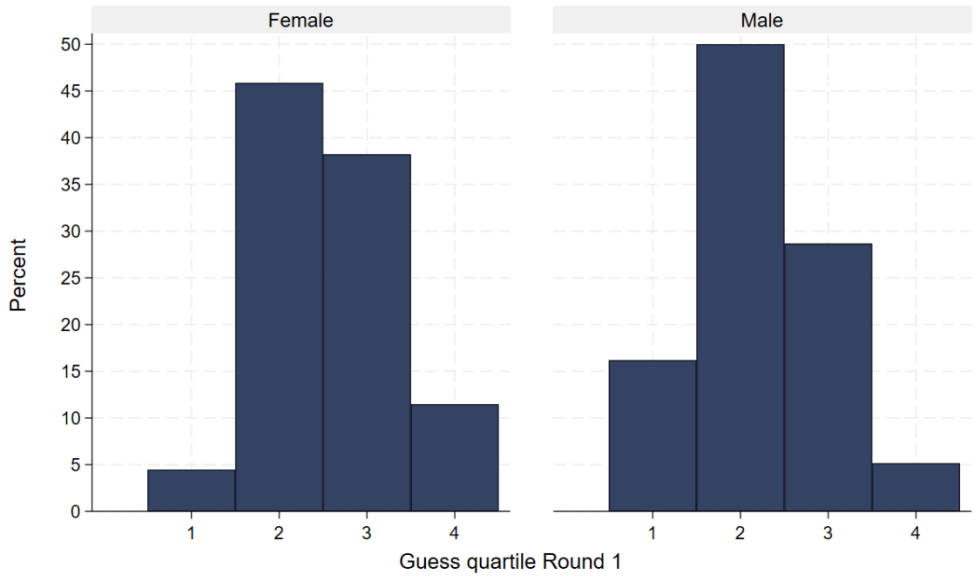
Figure 2: Proportion of guesses above the standard after Task 1 by treatment and gender



Men held higher performance expectations than women within their session. Figure 3 shows that, while most participants, regardless of gender, anticipated their performance to fall into the second or third quartile, men's distribution of guesses was shifted towards higher quartiles compared to women's.[10] This observation is supported by an ordered logit regression (third column, Table A1), whose estimates show that women believed their performance occupied worse (higher) quartiles than men, even after controlling for performance and other relevant personal characteristics. This difference in within-session performance beliefs remained in Task 2 (fifth column, Table A1)

[10] This picture remains almost identical after Task 2. See Figure A1 in the appendix.

Figure 3: Guessed performance quartile in Task 1 by gender.



In sum, there were minimal to no gender differences in task performance or beliefs regarding the likelihood of performing above the standard of excellence. Consistent with prior findings from (e.g. Niederle and Vesterlund, 2007), men tended to be more optimistic, on average, about their performance within their respective sessions.

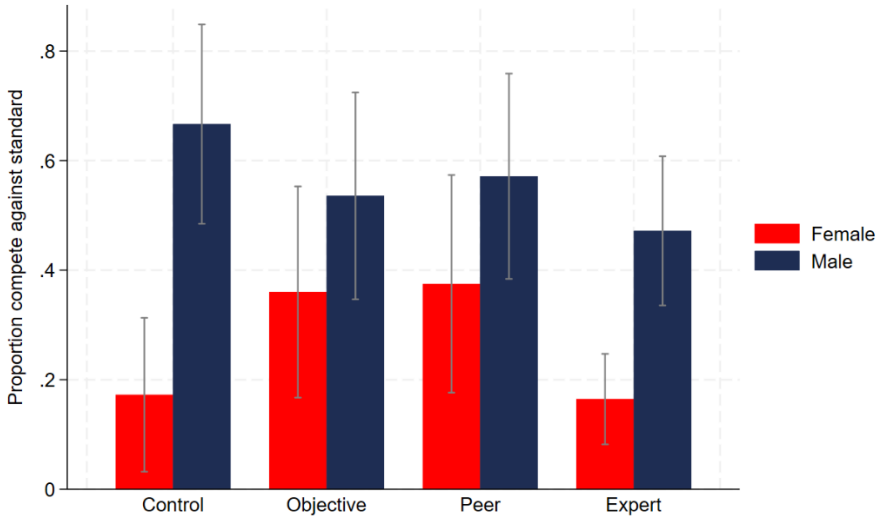
Gender differences in payment scheme choices

We now turn our attention to our first main result. Having experienced the task, participants were asked which payment scheme they wanted to apply to their performance in Task 2, either *piece rate* or *excellence*. Participants who chose the latter received a fourfold reward per correct answer compared to piece rate if their performance equated or surpassed the corresponding standard. Note that this was a purely individual decision; its outcome did not depend on the choices of other participants. The payment choice should have only depended on participants' beliefs about their own ability and the strictness of the standard of excellence.

Despite comparable performances and similar beliefs about their ability level relative to the standard, women chose the excellence payment scheme less often than men across all treatments. As Figure 4 below shows, most women chose the piece rate, while most men preferred to benchmark their performance against the standard of excellence. This gender gap in the choice

of the excellence payment scheme varied notably across treatments. It was widest in the *Control* treatment (66.7% vs 17.2%), and narrowest under the *Objective* condition (53.6% vs 36.0%), where the standard was exactly the same as in the *Control* but feedback was provided after Task 1. The substantial gender gap in payment scheme choices observed in the *Control* and *Expert* treatments (exceeding 30 percentage points) is statistically significant (proportion test, $p < 0.001$ in both cases). These differences became even more pronounced for confirmed payment choices after Task 2 (see Figure A2 in the appendix).

Figure 4: Payment scheme choices by treatment and gender.



Regression analyses confirm this result. Table 1 presents the estimates of a series of linear probability models with the choice of the excellence payment scheme as the dependent variable. The negative and significant coefficients for the variable *Female* in columns (1) and (2) show that female participants were about 20pp less likely than men to benchmark their performance against the standard, even after controlling for their performance in Task 1 and for relevant personal characteristics.

Table 1: Regressions on payment scheme choices, gender, and treatment

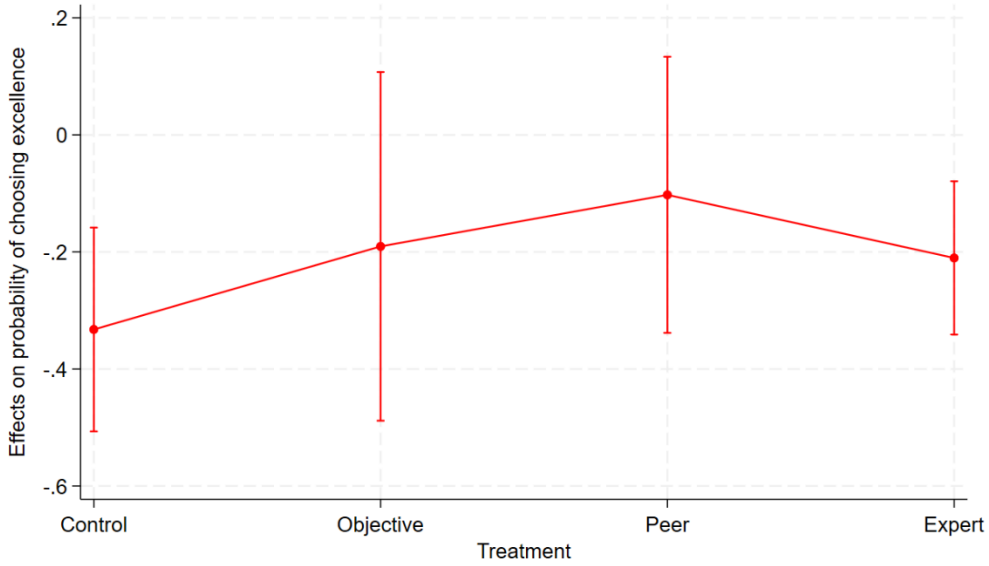
	(1)	(2)	(3)
Score Task 1	0.044***	0.0422***	0.0418***
	(0.008)	(0.00879)	(0.00845)
Guessed quartile	-0.119***	-0.114***	-0.114***
	(0.035)	(0.0335)	(0.0330)
Female	-0.206***	-0.201***	-0.386***
	(0.056)	(0.0565)	(0.0965)
Objective x Female			0.223
			(0.133)
Peer x Female			0.267*
			(0.144)
Expert x Female			0.214*
			(0.112)
Treatments	YES	YES	YES
Academic characteristics	NO	YES	YES
Personal characteristics	NO	YES	YES
Observations	293	293	293
R-squared	0.269	0.271	0.280

Standard errors clustered at session level. *** p < 0.01, ** p < 0.05, * p < 0.1. Academic characteristics include subject of study, seniority, and grade in Principles of Economics. Personal characteristics include risk tolerance, age, and self-efficacy score. All regressions include treatment and COVD dummies.

In column (3), we explore whether this gender gap varied across treatments. We employ the resulting coefficients to estimate the marginal effect of being female in each treatment. Figure 5 depicts them. In *Control*, women were 36pp less likely to choose the excellence payment scheme, and 17pp in *Expert.* These marginal effects are significantly different from zero ($p < 0.001$ and

p = 0.012 respectively) unlike those for *Objective* and *Peer*. These results hold as well for confirmed payment choices after Task 2 (see Table A2 and Figure A3 in the appendix).

Figure 5: Marginal effect of being female on payment scheme choice by treatment.



In summary, we observe a sizable and significant gender gap in participants' choice of benchmarking their performance against a standard of excellence when they did not receive feedback on whether they had met the standard. This gap disappeared when feedback was provided and the standard was determined by a group of peers, either objectively or based on their expectations. However, the gender gap in payment choices reappeared when experts set the standard. Women were significantly less likely than men to choose to be paid according to whether their performance would meet a standard of excellence as defined by a group with superior expertise.

Table 2: Regressions on payment scheme choices, feedback, gender, and treatment

	(1)	(2)	(3)
Score Task 1	0.00619	0.00609	0.00378
	(0.0135)	(0.0134)	(0.0135)
Guessed quartile	-0.112***	-0.113***	-0.115***
	(0.0380)	(0.0387)	(0.0378)
Female	-0.102	-0.127*	-0.269*
	(0.0606)	(0.0644)	(0.132)
Passed standard	0.435***	0.397***	0.314*
	(0.0901)	(0.0918)	(0.165)
Passed standard x Female		0.0904	0.342
		(0.114)	(0.225)
Peer x Female			0.179
			(0.221)
Expert x Female			0.171
			(0.144)
Passed standard x Peer			0.187
			(0.238)
Passed standard x Expert			0.0899
			(0.156)
Passed standard x Peer x Female			-0.260
			(0.376)
Passed standard x Expert x Female			-0.420
			(0.294)
Treatments	YES	YES	YES
Academic characteristics	YES	YES	YES
Personal characteristics	YES	YES	YES
Observations	237	237	237
R-squared	0.380	0.382	0.391

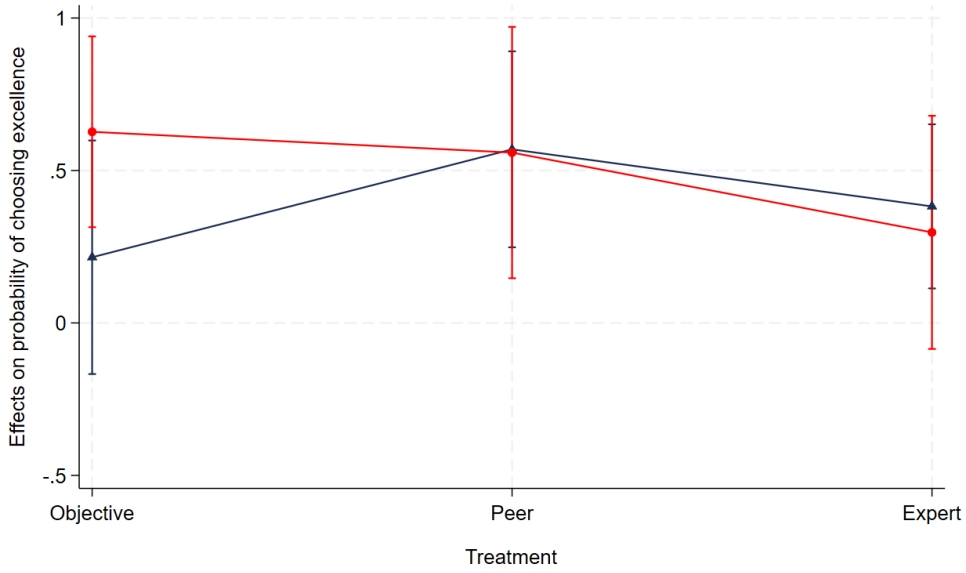
Standard errors clustered at session level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Academic characteristics include subject of study, seniority, and grade in Principles of Economics. Personal characteristics include risk tolerance, age, and self-efficacy score. All regressions include treatment and COVID dummies.

Gender differences in response to feedback

We next study whether men and women responded differently to the same feedback. Table 2 below presents regressions on the choice of the excellence payment scheme, now controlling for the type of feedback received after Task 1 (passed the standard vs. not). The sizable and significant positive coefficients for the variable *Passed standard* across all regressions indicate that participants were much more likely (between 30 and 40pp) to choose the excellence payment scheme after receiving positive feedback. The insignificant coefficient of its interaction with *Female* in column (2) indicates that women did not react differently to feedback when all treatments are pooled.

However, the estimates in column (3), which feature the triple interaction between gender, feedback, and treatment, show that women and men did react differently across treatments to passing the standard in Task 1. Figure 6 depicts the estimated marginal effects on payment choice of having passed the standard for men (blue) and women (red) by treatment. All marginal effects are positive and significant except for males in the *Objective* treatment ($p = 0.067$) and women in *Expert* ($p = 0.183$). In other words, receiving positive feedback did not influence males' payment choices when excellence was benchmarked against the performance of previous participants, nor did it affect females' payment choices when the standard of excellence was determined by the expectation of a group of experts.

Figure 6: Marginal effect of positive feedback on payment choice by gender and treatment.



It is important to note that in the *Objective* treatment, men and women who passed the standard in Task 1 responded differently to receiving good news. The marginal effects for men (31pp) and for women (65pp) are different from each other at the 95% confidence level. In short, women incorporated feedback into their payment choices to a greater extent than men when the standard of excellence was based on an objectively defined benchmark rather than an expectation.

Gender differences in incorrect payment choices

The results above would not be concerning if differences in payment choices simply reflected differences in performance; women (men) choosing less (more) often to benchmark their performance against a standard of excellence would be natural if they were simply less (more) likely to excel. This pattern would be concerning, though, if many excellent women chose piece rate and/or many mediocre men chose the excellence payment scheme. In that case, both these individuals and society as a whole would be better off if they chose differently.[11] Moreover, these mistakes would be distributed differently, as the cost of incorrect payment choices was higher for those who opted for the piece rate when they should have aimed for excellence, as they were foregoing a fourfold increase in earnings.

The results in the previous section already indicate that there were gender differences in incorrect payment choices. To summarize the evidence accumulated thus far: men were more likely than women to choose the excellence payment scheme in the *Control* and *Expert* treatments. Additionally, women who had passed the standard in Task 1 of the *Expert* treatment, and thus could be considered excellent, were as likely to choose the excellence payment scheme as women who had not passed the standard in that treatment. Taken together, this suggests that we should expect women to have incorrectly chosen the piece rate payment more often than men, and that more mediocre men than mediocre women mistakenly aspired to excellence in the *Control* and *Expert* treatments.

Table 3 below displays the frequency of incorrect payment choices conditional on Task 1 performance and their associated costs. In this table, we are assuming that performance costs (which cannot be directly measured) are identical across men and women and that performance was independent of the payment scheme chosen (recall that it could be changed after Task 2).

[11] Note that, unlike in experiments on tournament entry, payment scheme choices have no externalities in our setting, so individual and social welfare fully align.

The table reports ex-ante mistakes and costs, meaning it uses initial payment choices and Task 1 performance as a predictor of Task 2 performance. Table A2 in the appendix reports instead ex-post mistakes and costs, that is, using confirmed payment choices and Task 2 performance. The qualitative results fully remain.

Table 3: Mistakes in payment choices and their cost by gender and treatment

	Control		Objective		Peer		Expert	
	F	M	F	M	F	M	F	M
Correct choice	56.7%	48.1%	56.0%	53.6%	66.7%	67.9%	67.1%	58.5%
Under benchmarking								
Should benchmark	13	14	18	10	11	13	19	21
Did not benchmark	**76.9%**	**35.7%**	55.6%	40.0%	45.5%	23.1%	**84.2%**	**42.9%**
Average cost (€)	11.9	10.8	12.4	12.6	12.7	10.4	12.5	12.8
Over benchmarking								
Should not benchmark	16	13	7	18	13	15	60	32
Did benchmark	**12.5%**	**69.2%**	14.3%	50.0%	23.1%	40.0%	**16.7%**	**40.6%**
Average cost (€)	2.2	2.3	1.2	2.0	1.5	2.2	1.8	2.2

Note: Pairs of percentages in bold are statistically different under a proportions test at the 95% level.

Most participants made the right payment choice, even in the *Control* treatment where they received no feedback on whether their Task 1 performance had met the standard. However, there were significant differences in the errors women and men made, and consequently, in the earnings they forewent. Across all treatments, men were more likely than women to choose the excellence payment scheme when they should not have done so. Similarly, women were more likely than men to shy away from benchmarking their performance against the standard when they should have. In the *Control* and *Expert* treatments, these differences were statistically significant, with twice as many men incorrectly choosing excellence and four times as many women incorrectly choosing piece rate. Let us highlight one particularly striking result we observe: In the *Expert* treatment, 84% of the women whose Task 1 performance had been excellent according to the standard still chose not to benchmark their Task 2 performance against that standard.

Table 2 also presents the expected costs of incorrect payment choices. These costs are calculated as the difference between the potential earnings under the two payment schemes. On average,

the cost of under-benchmarking is approximately 12 euros, about 115% of the average earnings in the experiment. In contrast, participants who incorrectly chose the excellent payment scheme forewent 2 euros in earnings, about 20% of the average earnings. Since under-benchmarking was more prevalent among women, mistakes in payment choices hurt them considerably more. While error rates remained similar in final payment choices (as shown in Table A3 in the appendix), the cost of under-benchmarking tripled. This increase occurred because participants generally improved at the task, so mistakenly choosing the piece rate scheme became more costly.

To summarize, we find that women in our experiment shied away from benchmarking their performance against a standard of excellence compared to men. More excellent women than excellent men chose to be paid piece rate, whereas more mediocre men than mediocre women aspired to excellence. This pattern was particularly significant in the absence of feedback and when the standard of excellence was set by a group of experts. These mistakes reduced women's earnings disproportionally more as the cost of under-benchmarking against the standard was more than five times higher than the cost of over-benchmarking.

5. Mechanisms

In this section, we discuss the mechanisms that might be driving our experimental results. We first present a theoretical model that fleshes out two possible mechanisms: Gender differences in priors and biases in belief updating. Then, we present results of an out-of-experiment study we conducted on the same population that sheds light on some of these mechanisms. We then revisit the results of our main experiments in light of this additional theoretical and experimental evidence.

Theoretical framework

To identify the mechanisms behind our experimental results, we developed a theoretical model where an agent's success probability in each attempt at the task depends on their ability and a standard of excellence. Both of them are imperfectly known, so the agent updates their beliefs after observing their first task outcome. We next describe the predictions of this model while relegating the specifics to Appendix B.

The model predicts that, ceteris paribus, agents who succeed in the first attempt become more optimistic and are more likely to choose the excellence payment scheme. Agents with different

priors about their ability or the standard may choose different payment schemes even after experiencing the same first-attempt outcome. Hence, differences in beliefs about one's own ability and/or the strictness of the standard can naturally explain variations in payment scheme choices.

The model also allows for biased updating. Agents who underreact to signals, i.e., placing too little weight on the first-attempt outcome, tend to stick with their initial beliefs and are therefore less responsive to either success or failure. In contrast, agents who overemphasize the informativeness of the first outcome revise their beliefs sharply, leading to more extreme swings in their decisions.

We relate asymmetries in these deviations from Bayesian updating to well-known decision biases, each with distinct behavioral implications. Confirmation bias leads agents to react more strongly to signals that align with their priors and to underreact to those that contradict them. In contrast, motivated reasoning causes agents to rely more heavily on priors when the outcome confirms them, thereby reducing responsiveness to confirmatory information rather than contradictory information. These biases thus generate opposite predictions: confirmation bias amplifies responses to expectation-congruent outcomes, while motivated reasoning attenuates responses to those same outcomes.

The out-of-experiment study

The theoretical predictions above lay the ground for a better understanding of agents' decisions to compete against the standard of excellence. To establish the influence of beliefs and updating biases on payment choices, we conducted an out-of-experiment study where we directly measured key variables in the model such as performance expectations and beliefs about peer and expert standards. By comparing the model's predictions with the observed behavior in this study, we can assess whether the aforementioned mechanisms can explain the results of our main experiment.

The additional study included 65 undergraduate students at UPF who performed the task once under the same payment conditions as in the main experiment. We elicited incentivised beliefs about the standards set by both experts and peers, as well as beliefs about which gender was better at the task. Unlike in the main experiment, these beliefs could not be polluted by the presence of a second task and a payment choice, so they are better measures of beliefs about the strictness of the standards.

We observed no significant gender differences in actual performance scores (t-test, p = 0.6122), in beliefs about the standard set by peers (rank-sum test, p = 0.5914), or in beliefs about the standard set by experts (rank-sum test, p = 0.4662). However, there were significant gender differences in beliefs about which gender performs better (chi-squared test, p = 0.010; Fisher's exact test, p = 0.009), with women more likely to believe that men perform better, while men's views were more balanced.

Discussion

The gender differences in task perception we observed in the out-of-experiment study lend support to the presence of differential priors about ability, rather than about the strictness of the standard. The fact that women were more likely to experience self-doubt is also in line with our earlier observation of men being more optimistic than women about their performance in the first task relative to others, and with the significant gender gap in the *Control* treatment, as the table below shows.

Table 4: Participants who competed against the standard by gender, treatment and outcome of 1st attempt

	Control	*Objective*		*Peer*		*Expert*	
		Pass	Fail	Pass	Fail	Pass	Fail
Females	17.24% [7.22,35.81]	77.78% [41.91,94.81]	12.5% [3.12,38.78]	85.71% [41.67,98.05]	17.65% [5.77,42.87]	50.00% [19.89,80.11]	12.68% [6.71,22.67]
Males	66.67% [46.88,81.93]	80.00% [45.73,94.99]	38.89% [19.71,.62.25]	92.31% [60.69,98.94]	26.67% [10.32,53.46]	75.00% [52.03,89.24]	30.03% [17.09,47.83]

However, gender differences in beliefs about ability cannot explain why the gender gap varies so much across treatments and outcomes. This suggests the presence of biases in belief updating, such as overreaction or underreaction to the initial outcome.

Passing the test significantly boosted women's likelihood of competing against the standard. In contrast, after failing, the competition rate remained similar to that in *Control*. In line with motivated reasoning, success might have challenged women's less optimistic priors, while failure might have reinforced their initial self-doubt. This resulted in a low (high) weight assigned to the outcome of the first attempt when it was a fail (success). Feedback thus seems to have acted as a *confidence amplifier*: women who passed the first attempt at the task became much

more likely to compete against the standard, while those who failed remained equally reluctant as those who received no feedback.

Men exhibited the opposite pattern, again consistent with motivated reasoning. Their competition rate in the *Objective* treatment after a success remains as high as under *Control*, but decreases after a fail. This suggests that success reinforced their more optimistic priors, inducing little updating, whereas failure contradicted their priors and prompted downward updating. Feedback might have provided a *corrective mechanism*: men who passed the first attempt continued to behave as in *Control*, but those who failed became less likely to compete against the standard.

Next, we focus on within-gender differences in competition rates across treatments for a given outcome. Our theoretical model indicates that these differences, if present, can be safely attributed to differences in beliefs about the standard. For males, differences are not significant, suggesting that they expected the standard of excellence to be similar across treatments. The same applies to women who failed at the task, underscoring that failure just reinforced their less optimistic priors. However, women who passed the first attempt in the *Expert* treatment were less likely to compete against the standard than in the other treatments, suggesting that they expected the standard set by experts to be stricter. Below, however, we argue that that is not the best explanation.

Lastly, let us discuss the gender gaps in competition rates for a given outcome and by treatment. Because the out-of-experiment study shows that there were no gender differences in beliefs about the standard in the *Peer* and *Expert* treatments, any observed gender gaps must stem from differential beliefs about ability or from differences in belief updating.

The significant gender difference in competition rates in the *Expert* treatment is unlikely to reflect differences in beliefs about ability. The lack of gender gaps in the *Objective* and *Peer* treatments -where men and women held comparable beliefs about the standard- indicates that success washed away any differences in priors about ability. The weaker belief updating after a success in the *Expert* treatment – consistent with a low weight given to the outcome of the first attempt– suggest that women discounted positive expert signals.

We next consider two potential, non-mutually exclusive explanations for this observation:

First, expert feedback may have triggered the perception of a gender ability gap in the task uncovered in the out-of-experiment study. The significant gender difference in expected success rates in the *Expert* treatment (11.4% for females vs. 24.5% for males, proportions test $p = 0.0471$)

supports this interpretation. Notably, this gap is absent in the *Peer* or *Objective* treatments, where passing the standard appears to have offset any gender differences in priors, as shown also by the comparable rates at which males and females choose the excellence scheme. The authoritative context of expert evaluation, rather than the task itself, might have heightened female participants' concerns about their ability relative to their male peers.

Second, female participants might have interpreted their success in meeting the expert standard as a result of luck rather than ability. This explanation is consistent with the pattern observed among women who believed they had failed the first attempt but later learnt they had passed it. In the *Expert* treatment, only 40% of these women chose to compete against the standard—notably lower than in the *Objective* (71.4%) and *Peer* (80%) treatments. This muted response to expert-confirmed success suggests that they perceived it as a less reliable indicator of their ability. Such underweighting of positive feedback in the belief updating process could explain why female participants were significantly less likely to compete against the standard in the *Expert* treatment.

6. Conclusion

This paper has provided new insights into gender differences in competitive behaviour, particularly in relation to the setting of standards of excellence and the provision of performance feedback. Our experiment generates several key observations. Firstly, in the absence of feedback, women are less likely than men to benchmark their performance against a standard of excellence. This is inefficient because women who are likely to obtain increased rewards choose a low reward scheme instead. This choice is also surprising since these women knew they had met the standard in their first attempt at the task. When feedback is provided and the standard is set by peers, this gender gap closes. However, the gap re-emerges, and even widens, when the standard of excellence is set by experts.

Our results thus highlight the importance of feedback in influencing competitive behaviour. The hierarchical distance between participants and the source of feedback plays a crucial role in shaping competitive behavior. Carefully designed feedback mechanisms can enhance women's participation in excellence programs, which is critical for career advancement. Institutions aiming to promote gender equity should consider implementing peer-based feedback systems to encourage more women to pursue higher-reward opportunities.

Lastly, our study highlights how the initial decision to enter a competition, influenced by feedback, can have long-term implications for gender representation in various fields. If standards are set by experts and committees are perceived as male-dominated, a gender gap will exist in the award of promotions, grants or recognition. Understanding the differential impact of standards and feedback provision can help to design more inclusive competitive processes and bridge gender gaps in labour market outcomes.

References

Backus, P., Cubel, M., Guid, M., Sanchez-Pages, S., & Manas, E. (2023). Gender, competition and performance: Evidence from real tournaments. *Quantitative Economics,* 14(1), 349-380.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics - Foundations and applications 2*, 69–186. North-Holland/Elsevier.

Berlin, N., & Dargnies, M. P. (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization*, 130, 320-336.

Blau, F. D., & Kahn, L. M. (2020). The gender pay gap: Have women gone as far as they can? In *Inequality in the United States*, Ed, John Brueggemann, Routledge, 345-362.

Booth, A.L., & Nolen, P. (2022). Gender and psychological pressure in competitive environments: A laboratory-based experiment. *Economica*, 89, S71-S85.

Brandts, J., Groenert, V. & Rott, C. (2015). The impact of advice on women's and men's selection into competition. *Management Science*, 61(5), 1018-1035.

Brandts, J., Gërxhani, K., & Schram, A. (2020). Are there gender differences in status-ranking aversion? *Journal of Behavioral and Experimental Economics*, 84, 101485.

Coffman, K.B., Collis, M.R., & Kulkarni, L. (2024a). Whether to apply. *Management Science*, 70(7), 4649-4669.

Coffman, K.B., Collis, M.R., & Kulkarni, L. (2024b) Stereotypes and Belief Updating, *Journal of the European Economic Association*, 22(3), 1011–1054.

Datta Gupta, N., Poulsen, A., & Villeval, M. C. (2013). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry*, 51(1), 816-835.

Eckel, C., Gangadharan, L., Grossman, P. J., & Xue, N. (2021). The gender leadership gap: Insights from experiments. In *A Research Agenda for Experimental Economics*, Ed. Ananish Chaudhuri, Edward Elgar Publishing, 137-162.

Ertac, S., & Szentes, B. (2011). The effect of information on gender differences in competitiveness: Experimental evidence. Working paper, Koc University.

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

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

Hammerschmid, A., & Rowold, C. (2019). Gender pension gaps in Europe are more explicitly associated with labor markets than with pension systems. *DIW Weekly Report*, 9(25), 203-211.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin,* 108(3), 480–498.

Lozano, L., Ranehill, E., & Reuben, E. (2022). Gender and preferences in the labor market: Insights from experiments. *Handbook of Labor, Human Resources and Population Economics*, 1-34.

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science*, 68(11), 7793-7817.

Niederle, M. (2016). Gender. *Handbook of Experimental Economics*, second edition, Eds. John Kagel & Alvin E. Roth, Princeton University Press, 481-553.

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

Rosser, S. (2013). The gender gap in patents. In *Women, Science, and Technology*, second edition, Eds. Mary Wyer, Mary Barbercheck, Donna Cookmeyer, Hatice Ozturk, & Marta Wayne, Routledge, 111-130.

Sá, C., Cowley, S., Martinez, M., Kachynska, N., & Sabzalieva, E. (2020). Gender gaps in research productivity and recognition among elite scientists in the US, Canada, and South Africa. *PloS one*, 15(10), e0240903.

Scherbaum, C. A., Cohen-Charash, Y., & Kern, M. J. (2006). Measuring general self-efficacy: A comparison of three measures using item response theory. *Educational and psychological measurement*, 66(6), 1047-1063.

Schram, A., Brandts, J. y Gërxhani, K. (2019). Social-status ranking: A hidden channel to gender inequality under competition. *Experimental Economics*, 22(2), 396-418.

Wozniak, D. (2012). Gender differences in a market with relative performance feedback: Professional tennis players. *Journal of Economic Behavior & Organization*, *83*(1), 158-171.

Wozniak, D., Harbaugh, W.T., & Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, 32(1), 161-198.

Appendix A: Additional tables and figures

Table A1: Regressions on scores and beliefs

	Score in Task 1	Guess above standard in Task 1	Guessed quartile in Task 1	Score in Task 2	Guessed quartile in Task 2
Score in Task 1		-0.0576***	-0.159***	0.443***	
		(0.0203)	(0.0554)	(0.0677)	
First payment choice				-0.0401	
				(0.0428)	
Score in Task 2					-0.120***
					(0.0419)
Female	-0.790*	0.160	0.571***	0.106	0.897***
	(0.396)	(0.105)	(0.220)	(0.364)	(0.208)
Treatment					
Objective	-0.497	0.216	0.267	0.498	0.771
	(0.498)	(0.150)	(0.234)	(0.547)	(0.488)
Peer	-0.886*	0.0741	0.932**	0.0800	0.549
	(0.517)	(0.151)	(0.402)	(0.581)	(0.454)
Expert	-1.267**	0.255	0.860**	0.0412	1.075**
	(0.600)	(0.154)	(0.373)	(0.466)	(0.436)
Semester of study	0.356*	-0.0215	-0.451***	0.0719	-0.355**
	(0.208)	(0.0695)	(0.162)	(0.209)	(0.161)
Risk tolerance	-0.154*	0.0423	-0.0234	0.0274	-0.0874
	(0.0848)	(0.0277)	(0.0695)	(0.0988)	(0.0611)
Age	0.00918	-0.0519	-0.0170	-0.0166	-0.131
	(0.192)	(0.0401)	(0.0937)	(0.144)	(0.102)
Self-efficacy	-0.249	-0.163	-0.245	0.365	-0.574**
	(0.488)	(0.128)	(0.335)	(0.544)	(0.277)
Grade in principles	YES	YES	YES	YES	YES
Subject of study	YES	YES	YES	YES	YES
Observations	293	293	293	293	293
R-squared	0.129	0.154		0.217	

Standard errors clustered at session level. *** p < 0.01, ** p < 0.05, * p < 0.1. All regressions include COVD dummy.

Table A2: Regressions on final payment scheme choices, gender, and treatment

	(1)	(2)	(3)	(4)	(5)	(6)
Score Task 2	0.011	0.00904	0.00903	0.000610	-0.000213	-0.00135
	(0.007)	(0.00680)	(0.00703)	(0.00780)	(0.00773)	(0.00826)
Guessed quartile	-0.241***	-0.237***	-0.234***	-0.150***	-0.152***	-0.155***
	(0.032)	(0.0333)	(0.0331)	(0.0381)	(0.0374)	(0.0373)
Female	-0.164**	-0.153**	-0.325**	-0.0791	-0.0534	-0.132
	(0.062)	(0.0626)	(0.123)	(0.0626)	(0.0569)	(0.133)
Passed standard				0.366***	0.404***	0.262
				(0.0522)	(0.0598)	(0.158)
Passed standard x Female					-0.087	0.177
					-0.129	-0.291
Objective x Female			0.238			
			(0.149)			
Peer x Female			0.231*			0.143
			(0.124)			(0.196)
Expert x Female			0.191			0.0908
			(0.133)			(0.160)
Passed standard x Peer						0.207
						(0.227)
Passed standard x Expert						0.181
						(0.173)
Passed standard x Peer x Female						-0.386
						(0.401)
Passed standard x Expert x Female						-0.389
						(0.362)
Treatments	YES	YES	YES	YES	YES	YES
Academic characteristics	NO	YES	YES	YES	YES	YES
Personal characteristics	NO	YES	YES	YES	YES	YES
Observations	293	293	293	237	237	237
R-squared	0.329	0.335	0.344	0.421	0.423	0.430

Standard errors clustered at session level. *** p < 0.01, ** p < 0.05, * p < 0.1. Academic characteristics include subject of study, seniority, and grade in Principles of Economics. Personal characteristics include risk tolerance, age, and self-efficacy score. All regressions include treatment and COVD dummies.

Table A3: Mistakes in final payment choices and their cost by gender and treatment

	Control		Objective		Peers		Expert	
	F	M	F	M	F	M	F	M
Correct choice	33.3%	46.7%	83.3%	66.7%	66.7%	66.7%	25.0%	53.8%
Under benchmarking								
Should benchmark	13	14	18	10	11	13	11	18
Did not benchmark	**92.3%**	**50.0%**	72.2%	40.0%	63.6%	38.5%	**90.9%**	**61.1%**
Average cost (€)	29.3	26.6	30.2	33.0	31.3	30.0	32.7	31.1
Over benchmarking								
Should not benchmark	16	13	7	18	13	15	43	19
Did benchmark	**12.5%**	**61.5%**	14.3%	16.7%	15.4%	26.7%	14.0%	15.8%
Average cost (€)	5.3	5.5	3.0	5.7	4.0	6.3	6.0	4.7

Note: Pairs of percentages in bold are statistically different under a proportions test at the 95% confidence level.

Figure A1: Guessed performance quartile in Task 2 by gender.

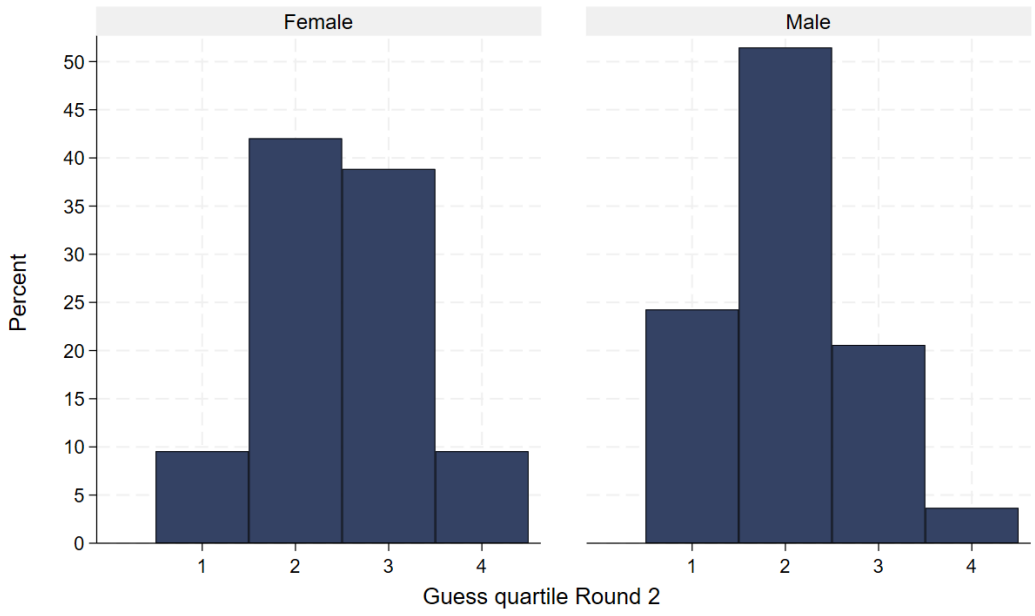


Figure A2: Confirmed payment scheme choices by treatment and gender.

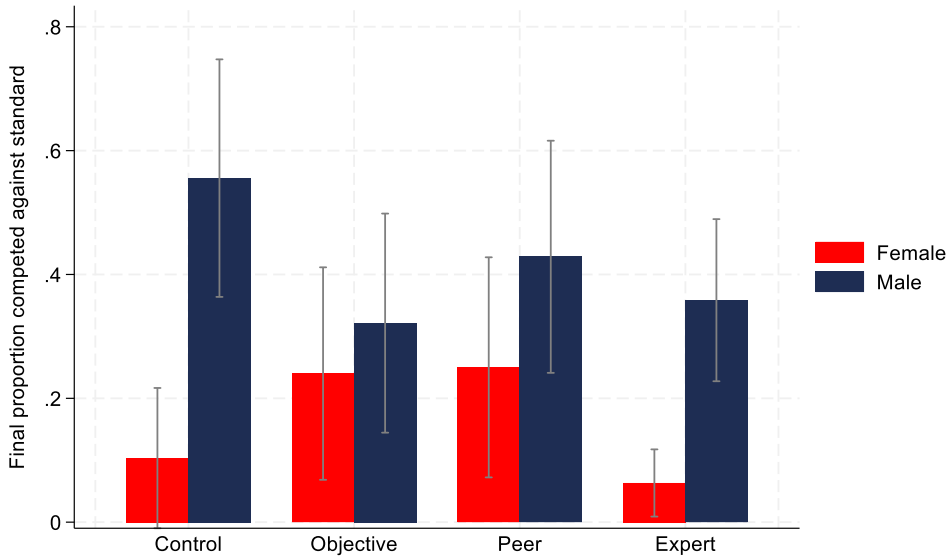
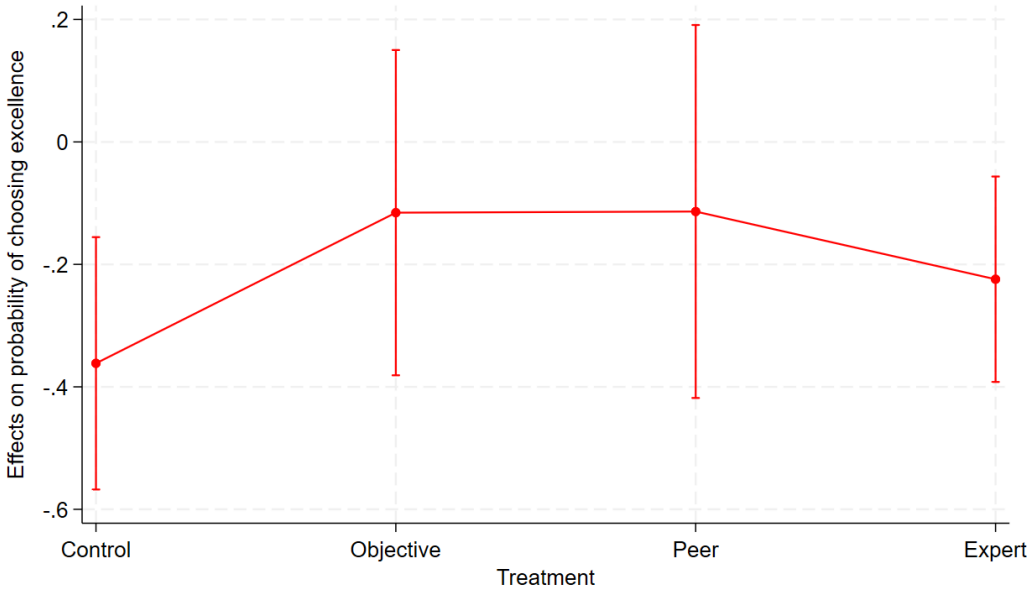


Figure A3: Marginal effect of being female on final payment scheme choice by treatment.



Appendix B: Theoretical model

Benchmark model

We consider a model of decision-making in which an agent must perform a task twice. After the first attempt, the agent observes its outcome, success or fail. Before the second attempt, the agent chooses how they want to be paid depending on its outcome. Then, they perform the task and are paid according to the payment scheme they chose.

The outcome of the task is a function of the agent's ability, denoted by $a \in \mathbb{R}$, and a standard of excellence, $s \in \mathbb{R}$. The agent is uncertain about both their own ability and the standard. The agent holds prior beliefs about a and s which are assumed to be independent random variables with log-Gamma distributions

$$a \sim \text{Log-Gamma}(\alpha_i, \theta), \quad s \sim \text{Log-Gamma}(\beta_i, \theta),$$

where $\alpha_i, \beta_i > 0$ are the shape parameter for the belief about the ability and the standard respectively and $\theta > 0$ is the rate parameter, assumed for exposition purposes to be the same for both variables. Note that a higher α_i reflects a more optimistic prior belief in the agent's ability to succeed whereas a higher β_i reflects a stronger prior belief that the task is difficult to pass.

The agent passes the task if $a \geq s$. However, in real-world settings, a variety of influencing factors make task outcomes less deterministic. Even if the agent's ability exceeds the standard of excellence, there may still be a non-zero chance of failure due to stress, distractions, or variations in effort. For that reason, we assume that the probability of passing the task depends on the agent's ability and the standard through a logit function. Specifically, the probability of success is defined as:

$$p = \frac{1}{1 + e^{-(a-s)}}$$

The probability of success in the task increases as the difference between the agent's ability and the standard of excellence becomes larger.

Whilst remaining realistic, this approximation to the cutoff rule implies that success in the task is a Bernoulli random variable, i.e. $y \sim \text{Bernoulli}(p)$, where $y = 1$ indicates success and $y = 0$ indicates failure. Moreover, because the log-Gamma distribution assumption implies that the exponentiated ability and exponentiated standard follow Gamma distributions, the agent's probability of success p_i follows a Beta distribution, i.e. $p_i \sim \text{Beta}(\alpha_i, \beta_i)$. This is due to the

conjugacy properties of the Gamma distribution as the sum and the ratio of two Gamma distributions is also a Gamma.

After the first attempt at the task, the agent observes the outcome $y_1 \in \{0, 1\}$ and updates their beliefs about p_i accordingly. If the agent passes the task, they gain evidence supporting their ability to succeed, while failure provides evidence on the contrary. Using the well-known conjugate updating rules for the Beta distribution, the updated shape parameters after observing outcome y_1 are simply

$$\alpha_i' = \alpha_i + y_1, \quad \beta_i' = \beta_i + (1 - y_1).$$

Before the agent makes a second attempt at the task, they must choose between:

1. Fixed scheme: The agent receives a reward R regardless of the outcome.
2. Excellence scheme: The agent receives a reward 4R if they pass the task ($y_2 = 1$), and no reward if they fail ($y_2 = 0$).

The agent evaluates each payment scheme based on their expected payoff given their updated belief about p_i. Under the fixed scheme, the payoff is just R. Under the excellence scheme, the expected payoff depends on the agent's expected value of p_i, which under the Beta posterior distribution is simply

$$E[p_i|y_1] = \frac{\alpha_i'}{\alpha_i' + \beta_i'}$$

Hence, the agent selects the excellence scheme if

$$4R \cdot \frac{\alpha_i'}{\alpha_i' + \beta_i'} \geq R \quad \Rightarrow \quad \alpha_i \geq \frac{\beta_i}{3} + (-1)^{y_1}$$

Otherwise, they select the fixed payment scheme. The choice rule above has two implications:

First, an agent is more likely to choose the excellence scheme if they passed the first attempt, i.e. $y_1 = 1$. Secondly, two agents with the same outcome in the first attempt may choose different payment schemes if they differ in their confidence about their own ability, as measured by α_i, or in their belief about the strictness of the standard, measured by β_i.

It will be useful to compare the above rule with the one in the case where the agent receives no assessment feedback after performing the task the first time. In that case, no belief updating takes place and the agent selects the excellent scheme if $E[p_i] \geq \frac{1}{4}$, that is,

$$\alpha_i \geq \frac{\beta_i}{3}$$

Hence, an agent who receives no performance feedback should be less (more) likely to select the excellence scheme than an agent who observes they passed (failed) the first attempt.

Deviating from Bayesian updating

There is substantive evidence of people deviating from standard Bayesian belief updating. Individuals may put too much or too little weight on the signals they receive or on their priors compared to the Bayesian benchmark. In addition, they may update differently in response to "good" or "bad" news.

We next incorporate these possibilities in the belief updating process that takes place in out set-up after agents observe the outcome of the first attempt at the task y_1. We will adopt the reduced form of biased updating introduced by Grether (1980). According to that, the agent's updated belief of passing the second attempt at the task is given by

$$Prob(p|y_1) = \frac{[\text{Prob}(\text{y}_1|\text{p})]^{\sigma}[\text{Prob}(p)]^{\delta}}{\int_0^1 [\text{Prob}(\text{y}_1|\text{z})]^{\sigma}[\text{Prob}(z)]^{\delta} dz}$$

where $c > 0$ measures the responsiveness of the update to the signal received and $\delta > 0$ measures the bias in the use of priors. When $\sigma > 1$ ($\sigma < 1$) the agent infers too much (little) information from the signal compared to the Bayesian benchmark. When $\delta > 1$ ($\delta < 1$) the agent considers priors to be more (less) informative than they are. The case $\sigma = \delta = 1$ corresponds to Bayesian updating. To minimize notational burden, we assume common parameters σ and δ across individuals and treatments.

A nice property of updating under Beta distributed beliefs is that they remain tractable. The updated shape parameters after observing outcome y_1 are now:

$$\alpha_i' = \delta \cdot \alpha_i + \sigma \cdot y_1, \;\; \beta_i' = \delta \cdot \beta_i + \sigma \cdot (1 - y_1),$$

and the agent's expected probability of succeeding in the second attempt at the task becomes

$$E[p_i|y_1] = \frac{\delta\alpha_i + \sigma \cdot y_1}{\sigma + \delta(\alpha_i + \beta_i)},$$

which is increasing (decreasing) in σ if the agent passed (failed) the first attempt. The effect of δ is the opposite. In words, the more weight an agent puts on the signal, or the less weight on

their priors, the more optimistic about their chances of passing the second attempt they become after receiving good news. The effect is the opposite is if they fail the first attempt.

Hence, the agent selects the excellence scheme if

$$4R \cdot \frac{\delta\alpha_i + \sigma \cdot y_1}{\sigma + \delta(\alpha_i + \beta_i)} \geq R \quad \Rightarrow \quad \alpha_i \geq \frac{\beta_i}{3} + (-1)^{y_1} \cdot \rho$$

where $\rho = \frac{\sigma}{\delta}$ represents the relative weight in belief updating assigned to the outcome of the first attempt compared to prior beliefs. A higher (lower) ρ indicates that the agents rely more (less) on the signal than on priors, resulting in less (more) conservative updating

There is substantial evidence of underinference from signals, i.e. $\sigma < 1$, and base-rate neglect, i.e. $\delta < 1$. Benjamin (2019) reports that experiments with unequal priors suggest that $\rho < 1$, which in turn would lead to conservatism, that is, agents adjusting their beliefs less in response to the first attempt outcome than under Bayesian updating.

However, there is also evidence of asymmetric and individual-specific inference biases. On the one hand, there is a well-identified confirmation bias that gives differential weight to signals depending on whether they confirm or disconfirm priors. On the other hand, signals can be interpreted differently depending on the task. Coffman et al. (2024b) observes that conservatism in belief updating is attenuated when men and women obtain good outcomes in a task they perceive as gender friendly.

This type in inference bias can be easily implemented in our model by allowing ρ to vary with the outcome of the first attempt. The choice rule can then be rewritten so that the agent selects the excellence scheme whenever

$$\alpha_i \geq \frac{\beta_i}{3} + (-1)^{y_1}\rho(y_1)$$

The literature suggests that ρ would be smaller when the outcome y_1 contradicts priors or conveys bad news, e.g. a fail in the task. This is associated with confirmation bias. However, under motivated reasoning (Kunda, 1990), the agent would assign a greater weight to their priors when the outcome y_1 confirms them, resulting too in a smaller ρ.