Luck or Skill: How Women and Men React to Noisy Feedback[*]

Gauri Kartini Shastry[†] — Wellesley College

Olga Shurchkov[‡] — Wellesley College

Lingjun "Lotus" Xia[§] — Massachusetts Institute of Technology

First Draft: August 2018

Current Draft: September 2019

Abstract

We design an experiment that shows that noisy feedback about relative standing elicits differential responses by gender and may exacerbate gender gaps in competitiveness and performance. Specifically, with limited feedback, women sort into competition based on ability; men only sort into competition on ability after receiving additional feedback on relative standing. Negative feedback eliminates this positive selection into tournament for women but not men. We document a possible mechanism: high-ability women attribute negative feedback to lack of ability, while high-ability men attribute it (correctly) to bad luck. Women are also less likely than men to take credit for positive feedback.

Key Words: gender differences, competition, attribution, feedback, economic experiments

JEL Classifications: C90, J16, J71

---

[*] We wish to thank Thomas Buser, Kristin Butcher, Jiafeng Chen, Catherine Eckel, Antonio Filippin, Eric Hilt, Casey Rothschild, and Jeremy Wilmer for helpful discussions and feedback, as well as the participants of the Wellesley College Economics Research Seminar, the University of Massachusetts Resource Economics Department Seminar, and the University of Tampere Institutions in Context Workshop on Gender Equality and Policy. Xia gratefully acknowledges the Jerome A. Schiff Fellowship for financial support. Shastry and Shurchkov gratefully acknowledge financial support from Wellesley College Faculty Awards. All remaining errors are our own.

[†] Department of Economics, Wellesley College, 106 Central St., Wellesley, MA, USA, gshastry@wellesley.edu

[‡] Corresponding author. Department of Economics, Wellesley College, 106 Central St., Wellesley, MA, USA, olga.shurchkov@wellesley.edu

[§] Research Assistant to Prof. N. Agarwal, MIT. lxia@wellesley.edu.

1. INTRODUCTION

Gender gaps in economic outcomes are pervasive and persistent. A simple comparison of annual wages for full-time working women and men reveals a gender gap of about 79 percent (Blau and Kahn 2017). Only 5 percent of Fortune 500 companies have female CEOs, with women occupying a meager 26.5 percent of senior-level managerial positions (Catalyst 2018). Leadership gaps are also prevalent in politics. In the US, women hold just around 20 percent of seats in the House of Representatives and Senate. The problem is not restricted to the US: in Canada, for example, just over one-quarter of members of the House of Commons are women (Catalyst 2018).

Explanations for these gender gaps fall into two broad categories. First, women might be discriminated against by employers, superiors, and/or coworkers, which may lead to fewer opportunities to achieve better economic outcomes even if we compare men and women within a given occupation or position (Sarsons 2017; Goldin 2014, for an overview). Second, women might self-select into lower-paying jobs and be less likely to undertake lucrative opportunities (Blau and Kahn 2017).[1] In this paper, we focus on gender differences in behavioral traits that may serve as underlying mechanisms behind this differential sorting (see Niederle 2016 and Shurchkov and Eckel 2018 for comprehensive surveys of the literature). In particular, a seminal study by Niederle and Vesterlund (2007) – hereafter NV – shows that men are twice as likely as women to enter a tournament, even when there is no significant gender gap in baseline ability. Subsequent research has investigated the provision of feedback on ability as a way to "nudge" women to be more competitive (Ertac and Szentes 2011; Brandt, Groenert, and Rott 2014). In many real-world settings, however, feedback depends partly on an individual's ability and partly on luck – a particular feature of feedback that has yet to be explored in the context of gender differences in tournament entry. Our experiment contributes to the literature by shedding light on how gender differences in signal extraction from noisy feedback about one's true ability can generate gender differences in selection into competition and, subsequently, earnings.

In our online experiment, participants first perform a task and report beliefs about their own performance. They receive a payment based partly on their own performance (ability component)

[1] Blau and Kahn (2017) show that about 50% of the gap is due to occupational choice, while 38% can be attributed to pure discrimination. However, choices could be made in anticipation of discrimination and may not reflect only different preferences.

and partly on whether their performance is higher or lower than that of a randomly chosen participant (luck component). They are then randomized into two main treatment conditions: control with no additional feedback and treatment with feedback about how their individual payment (inclusive of the luck component) relates to the average payment of a comparable group of participants. We refer to the control group as the limited feedback condition since the payment itself may reveal noisy information about one's ability.

In the second round, subjects perform the same task again, but first choose between a piece-rate payment scheme and a tournament-based payment scheme. We replicate NV's finding of the gender gap in competitiveness: in the control group, men enter the tournament more frequently than women, on average. However, we find that women know when to enter even with limited feedback: low-performers stay out of competition while high-performers enter. On the other hand, without additional feedback, men do not sort based on performance: low-performing men are as likely to enter as high-performing men. In fact, despite performing a task known to favor men, there is no gender gap in subsequent performance for those who enter the tournament because of this differential sorting. Positive sorting based on ability by women is consistent with similar findings that women act strategically in different contexts (see, for example, Exley, Niederle, and Vesterlund 2016 for bargaining).

Next, we find that additional feedback erases the gender gap in tournament entry, on average, by correcting the suboptimal entry pattern of low-performing men who now choose not to enter competition. At the same time, negative feedback eliminates the positive selection into the tournament for women – high ability women are no longer more likely to enter competition after receiving negative feedback, even though negative feedback for them is primarily due to bad luck. Consequently, feedback on relative standing *generates* a gender gap in performance and in earnings. Conditional on endogenous tournament entry, this gender gap reappears because fewer low-ability men enter the tournament in the feedback condition. This leads men to outperform the women who enter the tournament because our task favors men. In addition, some high ability women who receive negative feedback by chance opt out of competing for a larger payment, which enhances the gap in performance even when we do not condition on tournament entry. This result suggests that noisy feedback may exacerbate gender gaps in outcomes.

Our second set of results explores a potential explanation for the differential effect of feedback on sorting in and out of competition: the gender difference in assigning responsibility for particular outcomes, also called attribution of feedback. In connecting attribution of feedback to tournament selection, we address potential explanations for the persistent gender gaps in choosing to stay in certain educational or career tracks. In particular, if women blame themselves for negative feedback, while men are more likely to blame bad luck, then we would expect to see more men staying on competitive tracks and more women dropping out.

In our experiment, immediately before deciding whether to compete in the second round or choose the lower, piece-rate payment, participants are asked to assess the extent to which they believe the above or below average payment arose due to luck or ability (bad/low if the payment is below average or good/high if the payment is above average). We find that women consistently attribute negative feedback to low ability, regardless of their actual ability. This results in particularly distortionary consequences for tournament entry of high-ability women. These women receive negative feedback due to bad luck, but attribute it incorrectly to lack of ability. Consistently, these women are significantly less likely to compete relative to otherwise comparable women in the control group who do not receive additional feedback. On the other hand, men of high ability who receive negative feedback by chance correctly attribute it to luck and continue to enter the competition. Men attribute negative feedback to low ability only when the feedback confirms a pre-existing negative self-evaluation, potentially explaining the reduction in tournament entry for low-performing men.

This paper relates to several literatures. First, a large body of experimental work has tested the extent to which gender differences in behavioral traits, such as risk aversion, competitiveness, or social preferences, contribute to the gender gap in labor market outcomes (see Shurchkov and Eckel 2018 for a survey of the literature). Specifically, building on NV, Buser, Niederle, and Oosterbeek (2014) show that the gender difference in tournament entry explains a substantial portion of the gender difference in academic track choice. Furthermore, a number of studies show that the tournament entry gap increases with performance (NV; Buser, Peter and Wolter 2017; Buser, Ranehill and van Veldhuizen 2017). On the other hand, in our data, the tournament entry gap with limited feedback is driven by low-performing men who are disproportionately more likely to enter the tournament relative to low-performing women. With feedback on relative payment,

low-performing men update their beliefs and are significantly less likely to enter. This result is consistent with Mobius et al. (2011) who find that men who receive feedback are more likely to revise their beliefs about performance as compared to women who receive similar feedback. However, unlike in their study, we find that women who receive limited feedback sort correctly.

Second, we are the first to introduce a luck component into the information that men and women receive about their overall performance. In previous experiments (Ertac and Szentes 2011; Brandt, Groenert, and Rott 2014; and Berlin and Dargnies 2016), feedback perfectly reveals one's true ability. We show that gender differences in attribution of noisy feedback about relative standing to luck and ability produce gendered effects on sorting in tournaments, particularly for high-ability women.

The question of how people attribute successes and failures to ability or luck has been studied extensively in social psychology revealing that people exhibit "self-attribution bias:" the tendency to attribute success to own ability, but failures to some external forces, such as luck, in order to maintain self-esteem (Miller and Ross, 1975; Mezulis et al. 2004; see Eil and Rao 2011 for application in economics). Our study explores self-attribution in the context of reactions to feedback by asking the explicit attribution question and shedding light on how responses vary by gender. Our findings of gender differences in attribution of feedback provide a novel explanation for recent findings of gender differences in tournament entry even after receiving feedback. For example, Buser and Yuan (forthcoming) show that women are less likely than men to compete after losing in an earlier round. Along the same lines, Filippin and Gioia (2018) find that males become significantly more risk averse after losing a tournament than after randomly earning the same low payoff. Our findings are consistent with this observation, as men are more likely to blame the loss on bad luck instead of ability, increasing their risk-aversion.

The rest of the paper is organized as follows. Section 2 describes the experimental design and provides a first look at the data. Section 3 summarizes observed gender differences in preliminary outcomes, such as ability in the task, and behavioral traits such as confidence and risk attitudes. Section 4 presents our main findings on the effect of feedback on selection into competition and gender gaps in subsequent performance. Section 5 reports gender differences in attribution of feedback and links attribution to gender differences in tournament entry. Section 6 concludes.

2. The Online Experiment

Our controlled online survey experiment establishes an environment where we can observe the effect of feedback on sorting into competition and on how men and women differ in their attribution of that feedback.

The experiment was programmed in Qualtrics and conducted using the Amazon Mechanical Turk (AMT) platform. Workers on AMT have been shown to exhibit similar behavioral patterns and pay attention to the instructions to the same extent as traditional subjects (Paolacci, Chandler, and Ipeirotis, 2010; Germine et al. 2012). Rand (2012) reviews replication studies that indicate that AMT data are reliable. We used randomly placed attention checking questions in order to ensure full attention. A full set of screen shots with the consent form and experimental instructions is available in the online Appendix A.

2.1. Design of Round 1: Measuring Attribution across Treatments

In our experiment, participants began by completing round 1 of problem solving (see Section 2.2 below for a detailed description of the task). Participants had exactly 2.5 minutes to complete as many questions as possible out of 8. After completing the task, participants estimated their scores for that round of problem-solving. The resulting value measures each subject's *score confidence*.

Importantly, participants did not know their actual score from problem-solving at any point in the experiment. Instead, they received information about their payment. In particular, each participant learned that, in order to compute their payment, the computer randomly selected another participant who had previously taken the same test and compared their two scores.[2] If the participant's score was higher than or equal to that of her random match, she got 20 cents for each correct answer. However, if her score was lower than that of her random match, she got only 15 cents for each correct answer.[3] The participants then saw their payment (in US$). The exact

[2] The random match was drawn from the pool of participants in our pilot survey, wave 2. Subjects in this wave of the pilot were paid in an analogous manner to those in our main experiment. See online appendices B and C for detailed information on pilot waves 1 and 2.

[3] Given this design, savvy participants could have realized that they had 'won' the first round match-up if their payment was only divisible by 0.20 and not 0.15 (and vice-versa), as well as determining their actual score. This would eliminate some of the uncertainty about their own ability (it would reveal their score in the first round, but would not perfectly predict their score in a future round). We argue that participants did not respond to this divisibility further below.

wording of the payment information varied according to random assignment of participants into two treatments: match gender known condition or match gender unknown condition.

[UG] Match with unknown gender

In this treatment group, participants saw their payment without learning anything about the gender of their random match. In particular, they viewed the following on-screen text:

"*Our matching process has randomly matched you with a participant from the other group. Your score has been compared with his/hers, and your payment is shown below*."

We used a gender noncommittal pronoun "his/hers" to express that the random match might be of either gender.

[KG] Match with known gender

In this treatment group, the on-screen payment information text subtly revealed the gender of the randomly selected match. We kept this reference subtle in order to avoid alerting the participant to the fact that the research questions related to gender.

"*Our matching process has randomly matched you with a female (male) participant from the other group. Your score has been compared with hers (his), and your payment is shown below*."

The choice of "female" (hers) and "male" (his) depends on the actual gender of the random match. The gender specification is made implicit but repeated twice ("female participant," and "her," for example) to prevent participants from guessing the purpose of the experiment, and yet ensure they pick up on the cue.

Upon receiving payment information, subjects in both treatments were asked to assess whether they believed that their payment fell above or below the average payment earned by a previous group of participants (wave 2 of the pilot, see online appendices B and C for details). This measure allows us to gauge whether a given subject has a favorable or a negative view of their outcome. Specifically, we say that a subject who perceives her payment as below average has a "negative *self-evaluation*," while a subject who perceives her payment as above average has a "positive *self-evaluation*."

Next, participants were further randomized into three treatment conditions that varied the amount of feedback they would receive before answering a question on how they attribute their

payment outcome. To measure *attribution*, we asked all participants to use a slider in order to indicate the relative importance of luck, as opposed to their own ability, in determining their payment outcome (see Figure 1).

[FIGURE 1 ABOUT HERE]

[LF] Limited feedback condition

In this condition, participants did not receive any additional information before moving on to the attribution question. Thus, these participants assessed the contribution of luck and skill to their payment purely based on their perception of relative payment – their self-evaluation. For example, a participant with a positive self-evaluation assessed to what extent her presumed above-average payment was due to her own high ability, and to what extent it was due to good luck of being matched with a relatively weak participant from the other group. On the other hand, a participant with a negative self-evaluation estimated to what extent her presumed below-average payment was due to her own lack of ability, and to what extent it was due to bad luck of being matched with a relatively strong player from the other group. As noted above, the payment itself can be construed as a noisy signal about ability; we call this treatment the limited feedback condition, because participants did not receive any information with which to benchmark the payment.

[AF] Additional feedback condition

In this condition, participants learned whether their payment was *actually* above or below average. Therefore, there was no uncertainty about relative earnings, and these participants assessed the contribution of luck and skill to their payment based on actual feedback, rather than a self-evaluation. For example, a participant receiving positive feedback (payment above average) assessed to what extent her above-average payment was due to her own high ability, and to what extent it was due to good luck of being matched with a relatively weak player from the other group. On the other hand, a participant receiving negative feedback (payment below average) estimated to what extent her below-average payment was due to her own lack of ability, and to what extent it was due to bad luck of being matched with a relatively strong player from the other group.

Table 1 reviews the 2x2 factorial design in our experiment and provides the total number of subjects in each treatment condition.[4]

[TABLE 1 ABOUT HERE]

2.2. The Problem Solving Task

The choice of task in our experiment was based on three criteria. First, we looked for a task that cannot be easily "cracked" by the AMT subjects who, unlike lab subjects, have access to the internet and calculators. Second, we preferred a skill-based task to a menial task, such as the slider task (Gill and Prowse 2012) because of the more natural applications to real-world contexts where gender gaps are the greatest. Finally, our ideal task was either gender-neutral or stereotyped to suit men.[5]

In order to find a suitable task, we conducted two waves of pilot surveys that tested for gender differences in performance, confidence, and gender perceptions of three potential candidate tasks: Mental Rotation Task (MRT); "find the median" task; and a "pattern" task (also known as MPT or the matrix test). Detailed information on the three tasks, pilot survey protocol, and results can be found in the online Appendix, sections C and D.

Based on our pilot results, we settled on using the MRT in our main experiment. It is a test in which participants see a target three-dimensional shape made of 10 cubes and are asked to identify the rotated version of the target shape among the three choices. In our experiment, subjects received one point for each correct answer and zero points for each wrong or blank answer. Figure 2 shows a sample problem. [6]

[FIGURE 2 ABOUT HERE]

---

[4] We also ran a treatment condition where participants (101 in total) indicated their willingness to pay to receive feedback on relative payment (see Appendix B for details). However, very few participants were interested in receiving feedback on relative payment, which made it difficult to compare those who sorted into receiving feedback to those who did not. We therefore omit the optional feedback condition data from all subsequent analysis.

[5] Shurchkov (2012) points out that gender gaps in competitiveness are particularly pronounced in tasks which are perceived to favor men, such as the task we use in this study. In a separate pilot study, we experimented with a stereotypically female-favoring task, namely, the anagram task. However, we found that in many cases participants were able to use online search engines to achieve perfect scores.

[6] The MRT questions we used in our experiment were slightly different from the original MRT (Vandenberg and Kuse 1978). In the original MRT, there are four choices for each target shape. Exactly two of the choices are correct. Participants get 1 point for each correct choice and lose 1 point for each wrong choice. In order to reduce the difficulty level, we took out one of the correct choices for each target shape, and removed the penalty for incorrect choices.

Before completing the first round of problem-solving, each participant was given instructions for solving an MRT problem, as well as a practice example. They could not advance to the real problem-solving round until they correctly solved the practice question.

Within each problem-solving round, 8 MRT questions were presented on a single screen. The order of the questions was randomized for each participant.

2.3. Round 2: Tournament Entry and Questionnaire

In round 2 of the experiment, subjects once again had 2.5 minutes to solve as many MRT problems as possible. Prior to problem-solving, participants had to choose the type of payment scheme they wished to apply to subsequent performance. The two choices were:

- **Piece rate**: Participant gets 17.5 cents for each correct answer, regardless of anyone else's score.
- **Tournament**: We randomly match the participant with another participant from pilot wave 2. If the score is higher than the score of the random match, then the participant earns 25 cents for each correct answer. If the score is lower than the score of the match, then the participant only gets 10 cents for each correct answer.

Note that our experiment departs from the winner-take-all tournament design adopted in most previous studies on competitiveness, including NV. We believe that our design better approximates the nature of most competitions, where the loser still walks away with a prize, albeit a much more modest one. For example, the candidate who applies for but does not get the promotion is unlikely to be fired and can still go back to the original job. In Section 6, we further discuss the applications of this and other aspects of our design to real world environments where gender differences are particularly pronounced.

At the end of the experiment, each participant filled out a short questionnaire, which can be found in the online Appendix A. First, we elicited participants' risk preferences by asking them to self-report, on a scale of 1 to 10, how willing they are to take risks. Research has shown that the self-reported measure correlates significantly with experimental outcomes for risk-aversion and is widely accepted in the literature (Kagel and Roth 2016).

In order to confirm that our main experimental subjects hold similar gender stereotypes as the pilot subjects, the questionnaire asked the same question about gender perceptions associated with

MRT. Finally, the questionnaire also collected information on demographic characteristics, including age, gender, race, level of education, and income.

After participants finished the experiment, they were paid a base payment of $0.50 for completion. The final payment, including their bonus payment earned in the two problem-solving rounds, was transferred to their account within seven days. Including the base payment, the average payment was $2.1. The maximum payment was $4.1. The average duration of the experiment was about 11 minutes.

2.4. A First Look at the Data

We conducted four waves of data collection in the main experiment over the course of November 2017 – September 2018. In the first wave, we administered the full experimental design as described above. AMT recorded 308 valid responses with only 1 participant dropping due to failing to pass our attention check.

In the second wave, we focused on the additional feedback (AF) condition which we identified at the outset as our primary treatment of interest, collecting 88 additional valid responses. We also corrected a small coding error that affected the AF condition but did not preclude us from using the data from wave 1. Specifically, a coding error resulted in some participants receiving negative feedback when they should have received positive feedback, based on their payment relative to the average payment. Since the error only affected feedback received, we are able to use the data in our analysis: the error basically introduces an additional source of variation in the type of feedback subjects receive. Note that this was not by design, as we did not set out to deceive our subjects.

In the third wave, we collected more data in the LF and AF treatment conditions, but excluded the attribution question, in light of the potential concern that attribution might prime tournament entry. A total of 198 valid responses were collected in the third wave, limiting our power to test our results using just wave 3. However, we include wave fixed effects in all our regressions and confirm that tournament entry does not differ significantly in wave 3. The fourth wave was conducted in the summer of 2018 in order to collect more data and increase power and, like the first and second waves, included the attribution question.[7]

[7] The first three waves were conducted between November 2017 and February 2018; data were not analyzed between these waves. All of our results hold using just the data from these three waves (see Appendix E). We added the fourth

Table 2 reports the summary statistics of demographic characteristics and performs balance tests between treatment groups. The educational background of study participants is fairly similar to the national average in the United States, though skewing slightly towards being more educated (Ryan and Siebens 2016). The racial makeup in our study skews toward Whites and away from African-Americans and other minorities (US Census 2017). The income makeup in our study is generally similar to the national distribution, skewing towards the income range between $30,000 and $74,999 and away from the upper middle class (US Census 2017). Only 2 out of 71 comparisons across treatments are significant at the 0.05 level. All the procedures are randomized in the survey, but we conduct robustness checks and find that our results are robust to the inclusion of demographic controls (see online Appendix G).

[TABLE 2 ABOUT HERE]

In terms of the gender difference in demographic variables, women are slightly older than men in our sample (mean age for men is 37 and mean age for women is 41; $p < 0.0001$) and men are less likely than women to have attended but not graduated from college (21 percent of men and 27 percent of women; $p = 0.03$). Men and women are statistically indistinguishable on most other demographic characteristics (see online Appendix Table G1).

3. Gender Differences in Ability and Behavioral Traits

Table 3 reports summary statistics of key experimental variables by gender. All of these variables are from the round 1 task, collected before subjects were randomized into feedback treatment conditions, except for self-reported risk preference, which is collected after the round 2 task. We pool the data across the four waves and across treatments.[8] The order of variables reported in the table follows the chronological order of the design of the experiment.

[TABLE 3 ABOUT HERE]

wave with the hope of getting more 'positive surprises' – low-performing individuals getting positive feedback – but realized ex-post that there was insufficient noise in our design to generate many surprises. Most of the negative surprises were generated due to the randomness of being in the first wave with the coding error. There is no reason to think that being in the first wave is related to any unobservable characteristics; regardless, we still include wave fixed effects, relying on the variation within each wave. We also confirm that our results hold, qualitatively, if we just use waves 1 and 4 (see Appendix F), but we prefer to use all the data to maximize power.

[8] These gender differences are robust to the inclusion of wave fixed effects.

We find that men outperform women in the MRT task and consequently earn a higher bonus in round 1. Indeed, there is a consensus in psychology that MRT consistently elicits gender differences in performance (Masters and Sanders 1993). See online Appendix Figure G1 for the distribution of performance by gender.

Women in our experiment are significantly less confident than men, both in terms of their prediction of round 1 score (score confidence) and in terms of whether they believe their payment to be above or below average (confidence about relative payment).[9] A simple comparison of actual and predicted score for men and women reveals that women are directionally under-confident in our experiment (mean confidence of 3.61 relative to actual average score of 3.77 but the t-test p-value is 0.130). Men on the other hand correctly predict their scores, on average. Figure 3 plots the relationship between participants' expected and actual scores for the entire distribution of ability. We make three observations: 1) Men are systematically more score-confident ($p < 0.01$) than women, conditional on getting the same score. 2) Men with median performance (solved 4 out of 8 questions) on average correctly estimate their scores. Women at the median, on the other hand, underestimate their score (p-value of 0.0004). 3) Participants of both genders with higher-than-median performance tend to underestimate their score. Participants of both genders with lower-than-median performance tend to overestimate their score. Gender differences in confidence are a robust finding in the experimental literature (see for example, Beyer 1990), although men are typically found to be universally overconfident in lab experiments (see for example, NV).

[FIGURE 3 ABOUT HERE]

Table 4 uses ordinary least squares with wave fixed effects to confirm that the gender gap in both confidence measures decreases somewhat but remains significant if we control for actual performance (Columns 2 and 4). The finding is robust to including demographic controls (see online Appendix Table G2).

[TABLE 4 ABOUT HERE]

[9] Our experiment also replicates the gender gap in risk preferences found in the literature, with women reporting significantly lower risk attitudes (see Shurchkov and Eckel 2018 for a comprehensive review; but see Crosetto and Filippin 2017 for an important caveat that gender gaps in risk-taking vanish in absence of a clear safe option).

4. Effects of Feedback on Behavior

4.1. Tournament entry on average

We begin the discussion of our main results with a look at tournament entry patterns in our data. First, we consider the limited feedback treatment when facing a male opponent or an opponent of unspecific gender, the conditions most similar to those in NV's seminal paper. Figure 4 shows that the gender gap in tournament entry is significant and economically large under these conditions, replicating the NV result (dark grey bars, showing the tournament entry share of 0.49 for men and 0.28 for women, p-value of 0.002). Note that the result replicates even though tournament entry in our experiment is less risky than in NV and other similar studies, with the loser still receiving some payment.

[FIGURE 4 ABOUT HERE]

Next, we highlight the two changes that help to eliminate the gap in competitiveness in our setting. First, we continue with the limited feedback environment, and observe that facing a female match (i.e., opponent) significantly increases selection into tournament by women, 49 percent of whom compete in this condition (p-value of 0.023). On the other hand, the knowledge that the opponent is female actually decreases entry by men to 37 percent, although the reduction in competitiveness for men is not statistically significant (p-value of 0.250). The small sample of female-to-female matches in the limited feedback condition warrants caution when interpreting the results, and indeed the reverse gender gap in that condition is not statistically significant (the comparison of 49 percent entry for women to 37 percent entry for men produces a p-value of 0.331). Second, we find that additional feedback eliminates the gender gap in tournament entry, by reducing tournament entry by men, without changing tournament entry by women, on average (right panel of Figure 4). [10]

[10] As mentioned in Footnote 3 above, savvy participants could have figured out their first round score and whether they had 'won' or 'lost' the first round match-up if their payment was divisible by 0.20 or 0.15 but not both. This would eliminate some of the uncertainty about their own ability, but there would still be noise in the payment. We find no evidence that the participants responded to whether or not their payment was divisible by 0.15. Appendix Figures G3 and G4 present the fraction who entered the tournament and the fraction with a negative self-evaluation for each possible payment from the first round for those in the limited feedback treatment. Consider those who earned \$0.30 (got 2 questions right and lost the match-up) and those who earned \$0.40 (got 2 questions right, but won the matchup). Those who lost the tournament are slightly *less* likely to consider their payment below average and while they are also less likely to enter the tournament, the difference is very small. Similarly, consider those who earned \$0.75 and \$1, all of whom got 5 questions right, but some won the match-up and others lost it. Tournament entry and

4.2. SORTING INTO TOURNAMENT BY ABILITY

We next consider how feedback affects sorting into tournament with respect to ability. Figure 5 plots the relationship between score in round 1 and the probability of tournament entry for men and women in the limited feedback condition and in the additional feedback condition. In the absence of feedback on relative payment (Panel A), women are generally more likely to enter the tournament as their score in round 1 increases. However, for men, there is a slight U-shaped relationship, if any, between tournament entry and score: low-performing men are equally likely to compete as high-performing men. When feedback on relative payment is provided (Panel B), men react to it, while women appear not to react much. In particular, feedback on relative payment corrects the suboptimal entry pattern of low-performing men, but does not substantially alter the behavior of high or low-performing women.

[FIGURE 5 ABOUT HERE]

Table 5 verifies that the pattern of tournament selection by gender holds when we apply a linear probability model with wave fixed effects and controls for gender of the match.[11] Panel A accounts for round 1 score linearly. Note that we control for both genders interacted with round 1 score, leaving out the main effect, for ease of interpretation. In the limited feedback condition, women with the lowest scores are marginally less likely to enter the tournament than similar men (Column 1, row 1 of Panel A). Scoring one point higher in round 1 is associated with an approximately five percentage point increase in the probability of entering tournament ($p < 0.05$) for women. However, for men, scoring one point higher is not associated with a substantial change in the probability of entering the tournament (although the interaction terms for men and women are not statistically different from each other). Column 2 shows that risk preferences and confidence do not fully explain the pattern of tournament entry. These results are consistent with the finding of

self-evaluation actually go in the opposite direction – those who lost the match-up are more likely to have a positive self-evaluation and enter the tournament. Attribution tells a similar story, but the sample size gets very small since we have to separate the sample by their self-evaluation (the attribution question wording depends on whether they perceived their payment to be above or below average). Furthermore, looking at the participants in the additional feedback group, participants respond more strongly to the type of feedback than to whether or not their payment was divisible by \$0.15, but the numbers in each cell are too small to infer much. We confirm these results with regressions documenting that tournament entry, self-evaluation, and attribution are not affected by whether the payment is divisible by 0.15, once we control for either the score or the bonus from the first round.

[11] Results in Table 5 are qualitatively robust to including demographic controls (see online Appendix Table G3) or using a logit specification (see online Appendix Table G4).

women acting strategically in other contexts. For example, Exley, Niederle, and Vesterlund (2016) observe that women positively select into negotiations, whereas men do not.

[TABLE 5 ABOUT HERE]

Columns 3 and 4 of Table 5, Panel A, confirm that feedback about relative payment changes the behavior of men who now positively sort into competition in a manner similar to women. The coefficients on the male interaction terms differ significantly between Columns 1 and 3 (p-value of 0.08) and between Columns 2 and 4 (p-value of 0.03). Feedback on relative standing appears to erase the gender gap in tournament entry; the gender gap is significantly different in the additional feedback relative to the limited feedback condition. We test this in a few ways. First we note that the coefficients on female and female interacted with score are almost jointly significant in Column 1 (p-value of 0.112) but not in Column 3 (p-value of 0.646). But we also test the equality of these two coefficients across the limited feedback and additional feedback conditions and find p-values of 0.047 (comparing Columns 1 and 3) and 0.02 (comparing Columns 2 and 4).

Panel B of Table 5 breaks up the analysis by ability bins (with the lowest bin, those scoring 0 or 1 out of 8 as the omitted category, the second bin including those scoring below average, either 2 or 3 out of 8, the third bin including those who scored approximately average, either 4 or 5 out of 8, and the highest bin including those score 6 or above out of 8).[12] Once again, we observe that, in the limited feedback condition, women with the lowest scores are marginally less likely to enter the tournament than similar men (Column 1, Panel B, row 1). The interactions reveal that selection into tournament increases with ability for women. Relative to the bottom group, women in the second bin are 15 percentage points more likely to choose tournament, while women in the two highest groups are 22 and 39 percentage points more likely to compete, respectively. These interactions are jointly significant, as seen in the p-values at the bottom of Panel B (p-values around 0.05). The pattern for men, on the other hand, is slightly U-shaped, as we saw in Figure 5. In fact, men in the second and third groups are directionally less likely to compete as compared to those at the bottom and the coefficients on all interaction terms are not jointly significant. As in Panel A, the difference in sorting between men and women is not statistically significant – the three

[12] Results using these score bins are robust to using quartiles instead, with individuals scoring below average (4 out of 8) in the bottom quartile, those scoring exactly average in the second, those scoring 5 out of 8 in the third quartile and those scoring 6 or higher in the top quartile (see online Appendix Table G5).

interactions for men are not jointly different from the three interactions for women – but the interaction terms for men and women do differ marginally for the top two bins (p-values 0.102 and 0.072, respectively, in Column 2).

Columns 3 and 4 of Panel B confirm that additional feedback changes the sorting pattern for men from the slight U-shape to an increasing trend. The test of joint significance of interactions relative to the omitted lowest scorers for men now produces p-values of 0.04 and 0.10 (with controls).

Columns 3 and 4 of both panels of Table 5 suggest that feedback does not affect tournament entry behavior for women since they continue to sort correctly. However, these results mask the heterogeneity in women's response with respect to the type of feedback they receive. Specifically, equally capable women sort differently after receiving negative feedback than after receiving positive feedback. Figure 6 demonstrates this by plotting the relationship between score in round 1 and the probability of tournament entry for men and women in the additional feedback condition (Panel B of Figure 5) separated by the type of feedback they received. Panel A shows differential responses by women. Women who receive positive feedback exhibit positive sorting based on score, but women who receive negative feedback no longer increase tournament entry as score increases. Panel B displays the relationships for men, whose tournament entry probability increases with score regardless of whether the feedback they receive is negative or positive.[13]

[FIGURE 6 ABOUT HERE]

Table 6 further supports this finding by reporting estimates from linear regressions with wave and match gender fixed effects that split up the sample by gender and compare subjects in the additional feedback condition to those who did not receive any feedback on relative payment. All columns include controls for risk and confidence, although the results are robust to omitting these controls. Columns 1 and 4 of Table 6 include all participants of the relevant gender and replicate our results from Table 5. Women sort correctly into tournament with or without additional feedback on relative payment (Column 1). On the other hand, men who do not receive feedback

[13] Appendix Figure G5 replicates Figure 6 but using men and women in the limited feedback condition, separating by the type of feedback they would have received if they had been in the additional feedback condition. The plots confirm that there is no reason to think that women and men who scored in the 4-6 range would have acted differently based on the type of feedback they would have gotten (the two curves converge in that range). Panel A also suggests that women who would have gotten negative feedback (but did not) still sort correctly into tournament.

on relative standing do not sort positively based on score (row 2 of Column 4), while men who do receive this additional feedback exhibit a strong positive sorting relationship (row 3 of Column 4).

[TABLE 6 ABOUT HERE]

Restricting the sample to the women who receive or would have received negative feedback on relative standing reveals an interesting result (Columns 2 and 3). Women who would have received negative feedback, yet did not because they were in the limited feedback group (row 2 of Column 2), still sort based on score. As compared to Column 1, the coefficient is larger in magnitude, although statistically significant only at 10% as the sample size shrinks. More importantly, comparable women in the additional feedback group no longer sort correctly (row 3 of Column 2). Controlling for score in bins, instead of linearly, Column 3 demonstrates that this is driven by women at the top of the score distribution, although we see marginally significant differences in tournament entry along the entire distribution: higher-performing women in the limited feedback group are significantly more likely to enter, while higher-performing women who receive negative feedback are not. The p-value at the bottom of the table, testing equality of the three interactions for the two treatment conditions (that is, testing three hypotheses where the first is Limited feedback X Score Bin 2 = Additional feedback X Score Bin 2 and the second and third are the corresponding hypotheses for bins 3 and 4), confirms that women sort into tournament entry differently when they receive negative feedback than when they receive no additional feedback.

We replicate this analysis for men (Columns 5-6). As with all feedback in Column 4, men who receive negative feedback seem to respond by correcting their tournament sorting. Even though we lose power, row 3 of Column 5 shows that even negative feedback produces a directionally positive sorting based on score. When we break up the effect into ability bins for men, we observe once again that entry increases with score for men who actually receive the negative feedback, although the F-test of equality of the three interaction terms does not reject equality. Online Appendix Tables G6, G7 and G8 confirm that the results in Table 6 are qualitatively the same if we include demographic controls, use a logit specification, or use score quartiles instead of bins, respectively.

4.3. PERFORMANCE AND PAYMENT IN THE SECOND ROUND

It is worth noting that, while feedback on relative payment erases one type of gender gap in our experiment, it generates another gender gap due to differential sorting. This is true whether or not we condition on tournament entry. Figure 7 plots average scores for men and women in the round 2 task, along with confidence intervals. We replicate this figure for the round 2 bonus payment in Appendix Figure G2; the results are very similar, unsurprisingly, since the bonus payment is closely linked to performance. Panel A of Figure 7 demonstrates that, among those who do not receive additional feedback, women and men who enter the tournament perform equally well in round 2 (mean score of 4.2 for men and 4.6 for women; p-value of 0.37). Recall that the MRT favors men on average, but a large share of low-performing men entering the tournament causes the average score for men to be relatively low. On the other hand, feedback on relative standing causes low-performing men to drop out of competition, which leads to a significant gender gap in tournament performance in the additional feedback condition (mean score of 5.40 for men and 4.52 for women; p-value of 0.006).

Panel A conditions on tournament entry which is endogenous by design. However, the primary concern about the gender gap in tournament entry is that women lose out on positive returns by not entering tournaments in the first place. In addition, the emergence of the gender gap could be exacerbated by women's response to negative feedback since high ability women opt out of competition. In Panels B and C of Figure 7, we include all participants, regardless of tournament entry, but separate the sample by performance in the round 1 task. In the limited feedback treatment, there is no gender gap in tournament entry among the subjects in the top ability bin (those scoring 6 or higher out of 8 in round 1): mean score in round 2 of 5.97 for men and 6.43 for women, p-value of 0.37. However, receiving additional feedback creates a gender gap in performance among these high performers: mean score of 6.57 for men and 5.66 for women (p-value of 0.002). The difference in the gender gap is statistically significant in the additional feedback condition relative to the limited feedback condition (p-value of 0.02), suggesting that noisy feedback exacerbates the gender gap in outcomes for high ability women. Panel C suggests that there is a gender gap in performance for lower ability women too, and it is larger in the additional feedback treatment, though the difference is not statistically significant.

In summary, feedback about relative standing – both positive and negative – benefits men in terms of improving their tournament entry decision-making (and their payment in the second round, since it is based on their performance). However, negative feedback distorts women's entry decision: relative to otherwise comparable women who would have received negative feedback but do not by virtue of being in the control group, women who do receive negative feedback no longer sort based on ability. These responses to noisy feedback result in a gender gap in performance and payment in the second round. The next section of the paper investigates whether attribution of feedback to luck and ability by men and women can explain these gender differences.

5. Gender Differences in Attribution of Feedback

This section investigates a potential mechanism behind the finding that negative feedback causes equally capable women to reduce tournament entry relative to those who do not receive additional feedback. In particular, we ask whether the same type of feedback might be perceived differently by the two genders. In order to correctly measure attribution, we define four types of feedback outcomes: (1) positive feedback following a positive self-evaluation (i.e., positive reinforcement); (2) positive feedback following a negative self-evaluation (i.e., positive surprise); (3) negative feedback following a negative self-evaluation (i.e., negative reinforcement); (4) negative feedback following a negative self-evaluation (i.e., negative surprise). Furthermore, we restrict our attention to the additional feedback condition, where subjects assess the attribution of actual relative payment information to luck and to own ability (in the limited feedback condition, subjects assess the attribution of their self-evaluation of relative payment). Finally, we also restrict our attention to the participants in waves 1, 2 and 4, as the attribution question was omitted from wave 3. The attribution measure ranges from 0 (attributing relative payment outcome feedback completely to luck) to 100 (attributing payment outcome feedback completely to self).

Across all types of feedback, men and women do not significantly differ in their attribution, with men attributing 66.9 percent of the outcome to their own ability and women attributing 68.5 percent of the outcome to their own ability, on average. However, attribution varies with the type of feedback subjects receive. Table 7 reports average attribution by gender and type of feedback. The top two rows demonstrate that men and women who hold a negative self-evaluation (i.e., originally believed they had a below-average payment) act similarly in terms of attributing feedback. Both men and women who receive reinforcing negative feedback of their payment

actually being below average attribute this adverse outcome to their own relatively low ability more than to luck (about 70 percent to ability and 30 percent to luck). Both men and women who are surprised to learn that their payment was actually above average still attribute much of that outcome to own ability, albeit less than when they received reinforcing negative feedback.

[TABLE 7 ABOUT HERE]

The bottom two rows of Table 7 reveal that men and women react to feedback substantially differently when they originally hold a positive self-evaluation (i.e., originally believed they had an above-average payment). Men who receive negative feedback are significantly more likely than women to attribute it to bad luck. However, when feedback reinforces the self-evaluation as positive, men are significantly more likely than women to take credit for it as being due to their high ability.

Table 8 presents the estimates from a linear regression model with wave and match gender fixed effects that provides further insight into the gender gap in attribution of feedback (Panel A) and connects these results to the gender differences in how feedback affects the tournament entry decision (Panel B). The omitted category in all specifications is a man who receives positive feedback. The first two columns of Table 8 include all subjects in the additional feedback condition, regardless of their original self-evaluation, while Columns 3 and 4 restrict the sample to those who initially believed their payment was below average and Columns 5 and 6 restrict the sample to those who initially believed their payment was above average. The first row of Panel A reveals that, on average, women are more likely than men to attribute positive feedback to good luck rather than their own ability. Directionally, the effect is present in all specifications, but is driven by women who initially hold an above-average (positive) self-evaluation. Compared to women who receive positive feedback, women who receive negative feedback are significantly more likely to attribute this negative feedback to (lack of) ability (see p-values at the bottom of the panel). This is true regardless of original self-evaluation, directionally, but only significant when the negative feedback comes as a surprise. On the other hand, the second row shows that men only attribute the negative feedback to their own ability when it confirms their original negative self-evaluation (Columns 3 and 4). When the negative feedback is unexpected, men are significantly more likely to attribute it to bad luck (Columns 5 and 6). The p-values at the bottom of the panel

confirm that men and women respond differently to negative feedback overall and when they initially have positive self-evaluations.

[TABLE 8 ABOUT HERE]

Gender differences in attribution can help us explain the gender differences in the effect of feedback on relative standing on the tournament selection decision (Panel B of Table 8). In particular, when we consider the entire sample in Columns 1 and 2, we see that men and women are on average less likely to compete having received negative feedback. For those participants who initially held a negative self-evaluation (Columns 3 and 4), the effect is symmetric for both genders – receiving negative feedback decreases the likelihood of entry. Attribution can explain this: both men and women attribute reinforcing negative feedback to ability. However, the effect of negative feedback is not symmetric by gender when we consider the subsample of participants who initially held a positive self-evaluation (Columns 5 and 6). In particular, women who were surprised to receive negative feedback are significantly less likely to enter the tournament, which could be explained by the fact that they were more likely to attribute this news to lack of ability. On the other hand, men do not change their tournament entry decision, consistent with the fact that they attribute the negative surprise to bad luck. Online Appendix Tables G9 and G10 confirm that the results in Table 8 are robust to including demographic controls and using a logit specification for Panel B, respectively.

Table 8 relates gender differences in tournament entry to attribution but does not address the differential sorting with respect to ability that we documented in Tables 5 and 6. In Table 9, we revisit the analysis in Table 5, focusing on participants who would have received negative feedback. That is, we compare the subjects who receive negative feedback in the additional feedback treatment group and those who would have received negative feedback in the limited feedback group. The first four columns use tournament entry as the dependent variable, as in Table 5, while Columns 5 and 6 use attribution. Panel A shows the linear specifications, directionally replicating the results in Table 5 but losing power due to a smaller number of observations. Recall that women sort correctly into the tournament without receiving additional feedback (Columns 1 and 2), but not when they receive negative feedback on relative standing (Columns 3 and 4). On the other hand, men do not sort correctly into tournament with limited feedback (Columns 1 and 2) but do when they receive additional feedback (Columns 3 and 4). Columns 5 and 6 show that

men with higher scores are more likely to attribute the negative feedback to luck, while women with higher scores are more likely to attribute the negative feedback to lack of ability, although neither coefficient is statistically significant and they are not significantly different from each other (p-value at the bottom of the panel).

[TABLE 9 ABOUT HERE]

Panel B of Table 9 breaks up the score distribution into bins, as in Panel B of Table 5. Columns 1 and 2, using tournament entry as the dependent variable and focusing on the limited feedback condition, replicate the result that, even among those women who would have received negative feedback, the positive sorting relationship with respect to score remains, and women with higher ability are more likely to compete. However, women in the additional feedback condition (Columns 3 and 4) no longer correctly sort into tournament based on their score upon receiving negative feedback. The p-values at the bottom of the table confirm that the interaction terms are jointly significant in the limited feedback condition but not in the additional feedback condition. The sorting behavior of men does not differ substantially by feedback condition and the male interaction terms are jointly insignificant in both groups, although the interaction terms in the additional feedback condition are positive, relative to the lowest bin, and monotonically increasing (comparing Columns 1 and 2 to Columns 3 and 4).

Comparing men to women in Columns 1 to 4, we see that negative feedback discourages women, but not men, in the higher bins from entering the tournament. The p-value at the bottom of the table rejects equality of the interactions (men vs. women) in the limited feedback condition.

Columns 5 and 6 of Panel B, Table 9, shed light on whether attribution can once again explain the gender differences. Men in the higher score bins attribute their negative payment feedback more to bad luck, which explains why they do not alter their tournament entry behavior (the p-value at the bottom of the table indicates that the male interaction terms are jointly significant). On the other hand, whether women attribute negative feedback to ability or luck does not differ by their actual ability – in fact, women in the higher bins attribute the negative feedback more to ability than those in the bottom bin, although the differences are not statistically significant. This disproportionately changes the entry behavior of high-ability women, who receive negative feedback primarily due to bad luck, either because they were paired with someone who scored higher, or because they, by chance, happened to be in wave 1. Online Appendix Tables G11, G12

and G13 confirm that the results in Table 9 are qualitatively similar when we include demographic controls, use a logit specification for tournament entry, or use score quartiles instead of bins, respectively.

To summarize, we find a connection between the gender difference in attribution of feedback and the subsequent decision to compete. Women attribute negative feedback to lack of ability, regardless of whether the feedback is consistent with their original self-evaluation. This results in particularly distortionary consequences for tournament entry of high-ability women. These women receive negative feedback mostly due to bad luck, but attribute it incorrectly to lack of ability, which causes them to be significantly less likely to compete relative to otherwise comparable women in the control group who do not receive additional feedback. On the other hand, men attribute negative feedback to lack of ability only when the feedback confirms their pre-existing negative self-evaluation. When men expect a positive outcome, so that negative feedback undermines their positive self-evaluation, they attribute this feedback to bad luck. Thus, men of high ability who receive negative feedback mostly by chance correctly attribute it to chance and continue to enter the competition.

6. Discussion and Directions for Future Research

This paper describes an online experiment designed to test the extent to which gender differences in attribution of noisy feedback to luck and ability can explain tournament entry patterns. We find that, on average, men are significantly more likely than women to enter competition in the absence of feedback on relative standing. The gender gap is particularly strong when women face a male opponent or when the gender of the opponent is unknown. On average, feedback on relative standing eliminates the observed gender gap in tournament entry. A closer look at sorting patterns reveals that, in the absence of this feedback, women positively self-select into tournament based on score, whereas men do not – a finding that is consistent with Exley, Niederle, and Vesterlund (2016) who tell a similar story in the context of bargaining.

In our experiment, feedback about relative payment decreases the rate of tournament entry among low-performing men, erasing the gender gap in competitiveness. Additional feedback does not affect the rate of tournament entry for women, on average, but we find that women who receive negative feedback no longer sort into competition based on score, even though the negative

feedback is likely due to bad luck. Brandt, Groenert and Rott (2014) also find that feedback reduces entry for low-ability men, but, unlike in our study, feedback improves tournament entry of high-ability women. Our novel design that introduces a luck component into the information about one's performance reconciles this seeming discrepancy. Specifically, feedback in previous studies provides subjects with a precise signal of ability, so that there is no room for uncertainty about its attribution. On the other hand, we allow our subjects to vary in terms of the extent to which they attribute the feedback they receive to ability as opposed to luck. We find that women attribute negative feedback to lack of ability, regardless of whether it is consistent with their self-evaluation, possibly explaining why high-ability women opt out of competition after receiving negative feedback. Women are also more likely than men to attribute surprising positive feedback to luck, which could explain why high-performing women do not increase tournament entry even after receiving positive feedback. On the other hand, men attribute negative feedback to lack of ability only when they hold a negative self-evaluation initially. Upon receiving a negative signal confirming that belief, these men, who tend to be the low-performers, are then more likely to drop out of the tournament. On the other hand, when negative feedback comes as a surprise, men correctly attribute this bad news to luck and continue to enter the tournament.

These observations have potentially important implications for the labor market. In particularly risky environments, where luck is a large component in determining outcomes, such as in financial markets, women receiving negative feedback may benefit from a reminder that luck is a big factor. Otherwise, unlucky female investors, for example, may misattribute losses to low ability rather than luck and may be deterred from making future investments. An investigation of the effect of attribution of feedback on more real-world outcomes, such as investing behavior or educational track choice, and over a longer time horizon, is a fruitful direction for future research.

Note that, in the stylized context of our experiment, where the benefits of entering the tournament are completely driven by performance, sorting based on actual or perceived ability is the main mechanism for efficient selection. We find that women are particularly effective at sorting based on ability even with limited feedback. In this sense, our experiment is most directly applicable to environments where competition only serves one purpose, and that is to succeed at the particular task, such as attaining publication at a highly selective journal. In this setting, the act of submitting a paper to a high-ranking journal does not produce many further benefits other than

trying to publish there. Our experimental results suggest that negative feedback would be interpreted differently by men and women in such environments, potentially exacerbating gender gaps.

We further hypothesize that attribution of feedback can be important in real-world competitions that have benefits above and beyond those directly tied to one's performance. For example, the act of competing itself may serve as a signal of future performance or other attributes correlated with performance, such as ambition and confidence. Moreover, persistence in challenging or competitive fields despite negative feedback can also lead to an improvement in ability over time, as individuals learn and accumulate human capital. This is particularly relevant to the gender gap in major and career choice (Goldin 2013, Buser, Niederle and Oosterbeek 2014, Kugler, Tinsley and Ukhaneva 2017). Gender differences in attribution of feedback are more likely to play a role here, because negative feedback attributed to bad luck could lead to persistence in the competitive track (men), while negative feedback attributed to lack of ability could lead to dropping the career track (women). Thus, in future work, we plan to investigate the consequences of gender differences in attribution for economic outcomes where dropping out due to perceived low ability may not be the optimal course of action in the long run.

REFERENCES

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

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

Blau, F. D., & Kahn L. M. (2017). The gender wage gap: extent, trends, and explanations. *Journal of Economic Literature*, 55(3), 789–865.

Brandts, J., Groenert, V., & Rott, C. (2014). The impact of advice on women's and men's selection into competition. *Management Science,* 61(5), 1018–1035.

Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, 129(3), 1409–1447.

Buser, T., Peter, N., & Wolter, S. C. (2017). Gender, competitiveness, and study choices in high school: evidence from Switzerland. *American Economic Review*, 107(5), 125–130.

Buser, T., Ranehill, E., & van Veldhuizen, R. (2017). Gender differences in willingness to compete: the role of public observability. University of Zurich, Department of Economics, Working Paper No. 257.

Buser, T. & Yuan, H. (2019). Do women give up competing more easily? Evidence from the lab and the Dutch Math Olympiad. *American Economic Journal: Applied Economics*, 11(3), 225–252.

Catalyst. (2018). Knowledge center: women in S&P 500 companies. http://www.catalyst.org.

——, 2018. Knowledge center: women in government. http://www.catalyst.org.

Crosetto, P. & Filippin, A. (2017). Safe options induce gender differences in risk attitudes. IZA DP No. 10793.

Eil, D. & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics,* 3(2), 114–138.

Ertac, S. & Szentes, B. (2011). The effect of performance feedback on gender differences in competitiveness: experimental evidence. Koç University-TUSIAD Economic Research Forum Working Papers 1104.

Exley, C. L., Niederle, M., & Vesterlund, L. (2016). Knowing when to ask: the cost of leaning in. NBER Working Paper w22961, National Bureau of Economic Research.

Filippin, A. & Gioia, F. (2018). Competition and subsequent risk-taking behaviour: heterogeneity across gender and outcomes. *Journal of Behavioral and Experimental Economics*, 75, 84–94.

Germine, L., Nakayama, K., Duchaine, B., Chabris, C. F., Chatterjee, G. & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review,* 19(5), 847–857.

Gill, D. & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102, 469–503.

Goldin, C. (2014). A grand gender convergence: its last chapter. *American Economic Review* 104(4), 1091–1119.

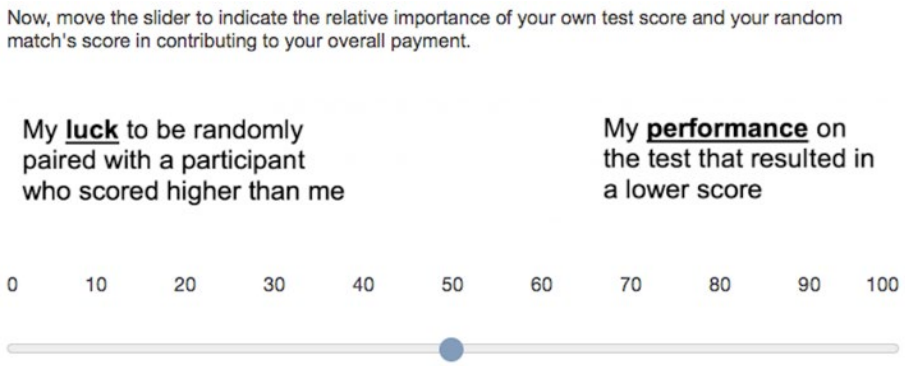
Goldin, C. (2013). Can 'Yellen Effect' attract young women to economics? *Bloomberg View*, Oct 14, 2013. https://www.bloomberg.com/view/articles/2013-10-14/can-yellen-effect-attract-young-women-to-economics, Accessed Mar 1, 2017.

Kagel, J. H. & Roth, A. E. eds. (2016). *The Handbook of Experimental Economics, Volume 2*. (Princeton University Press, Princeton, NJ).

Kugler, A. D., Tinsley C. H., & Ukhaneva, O. (2017). Choice of majors: are women really different from men? Working Paper 23735, National Bureau of Economic Research.
Masters, M. S. & Sanders, B. (1993). Is the gender difference in mental rotation disappearing? *Behavior Genetics,* 23(4), 337.
Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin* 130 (5): 711.
Miller, D. T. & Ross, M. (1975). Self-serving biases in the attribution of causality: fact or fiction? *Psychological Bulletin,* 82(2): 213.
Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). Managing self-confidence: theory and experimental evidence. Working Paper 17014, National Bureau of Economic Research.
Niederle, M. & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics,* 122(3), 1067–1101.
Niederle, M. (2016). Gender. Chapter 8 in *The Handbook of Experimental Economics, Volume 2*, John Kagel and Alvin E. Roth, eds. (Princeton University Press, Princeton, NJ).
Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making,* 5(5), 411–419.
Rand, D. (2012). "The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology,* 299(21), 172–179.
Ryan, C. & Siebens, J. (2016). Educational attainment in the United States: 2015. U.S. Census Bureau. Retrieved December 22, 2017.
Sarsons, H. (2017). Gender differences in recognition for group work: gender differences in academia. *American Economic Review: Papers and Proceedings*, 107(5), 141–145.
Shurchkov, O. (2012). Under pressure: gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association,* 10(5): 1189–1213.
Shurchkov, O. & Eckel, C. C. (2018). Gender differences in behavioral traits and labor market outcomes. in *The Oxford Handbook of Women and the Economy*, Susan L. Averett, Laura M. Argys, and Saul D. Hoffman, eds. (Oxford University Press, New York, NY.)
United States Census Bureau Quick facts. (2017). Available at: https://www.census.gov/quickfacts/fact/table/US/PST045217 (accessed 24 Nov 2018).
United States Census Bureau Current Population Survey. (2017). HINC-01. Selected characteristics of households by total money income, all races. Retrieved from: https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-01.html (accessed 24 Nov 2018).

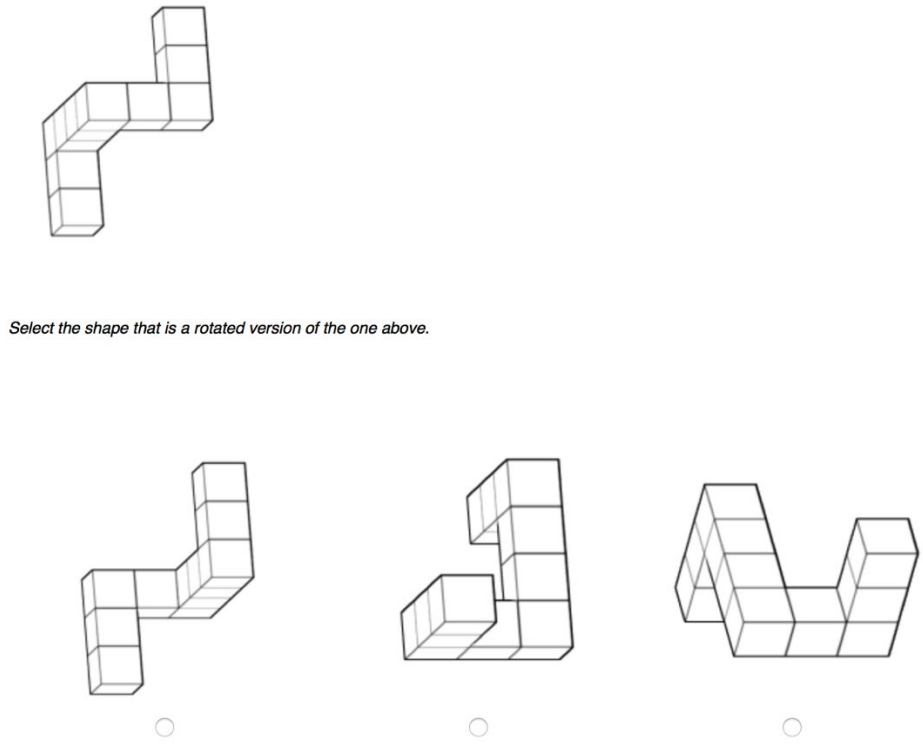
FIGURES

Figure 1: The Attribution Question for a Participant with Negative Information



Note: The measure of attribution is scaled from 0 to 100. A value of 100 means attributing one's payment entirely to one's own performance on the test; a value of 0 means attributing one's payment entirely to luck, that is, being paired with a specific match.

Figure 2: Sample MRT Problem



Note: The correct answer is the third choice.

Figure 3: The Relationship between Expected Score (Score Confidence) and Actual Score in Round 1, by Gender

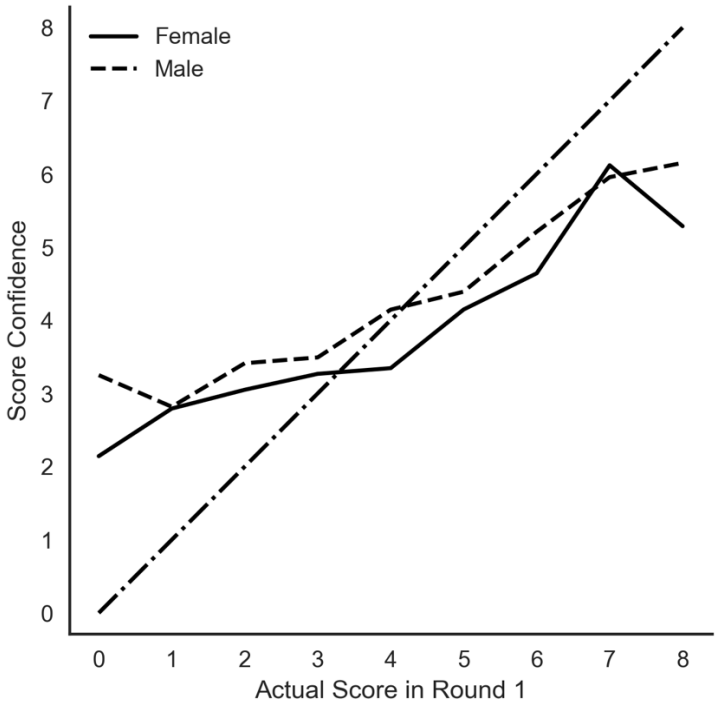


Figure 4: Gender Differences in Tournament Entry in the Limited Feedback and Additional Feedback Treatments, on Average and by Information Condition about the Gender of Random Match

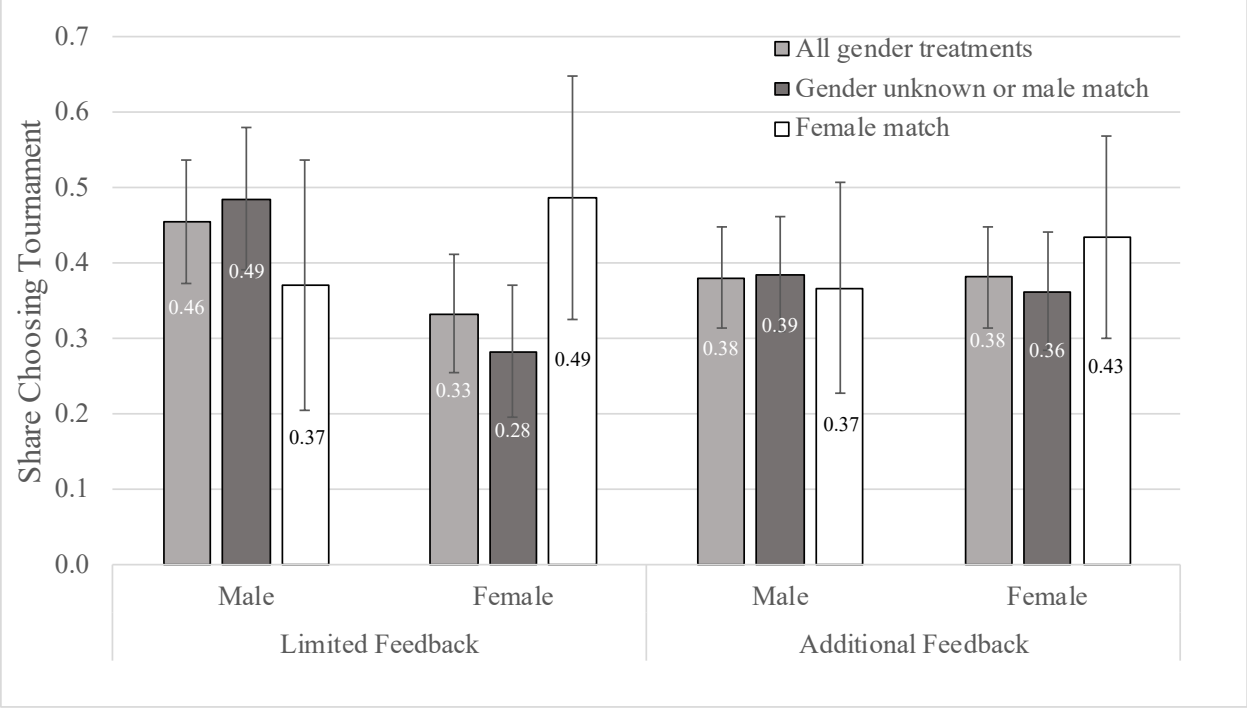


Figure 5: Local Linear Regression of Tournament Entry on Score in Round 1 by Gender and Feedback Treatment Condition

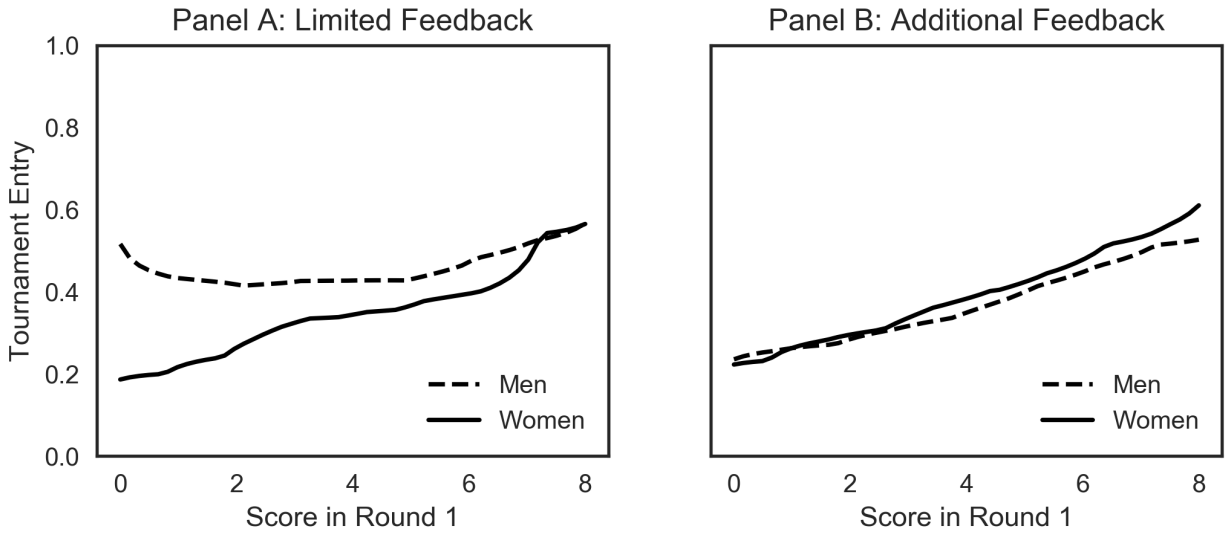


Figure 6: Local Linear Regression of Tournament Entry on Score in Round 1 by Gender and Type of Additional Feedback Received

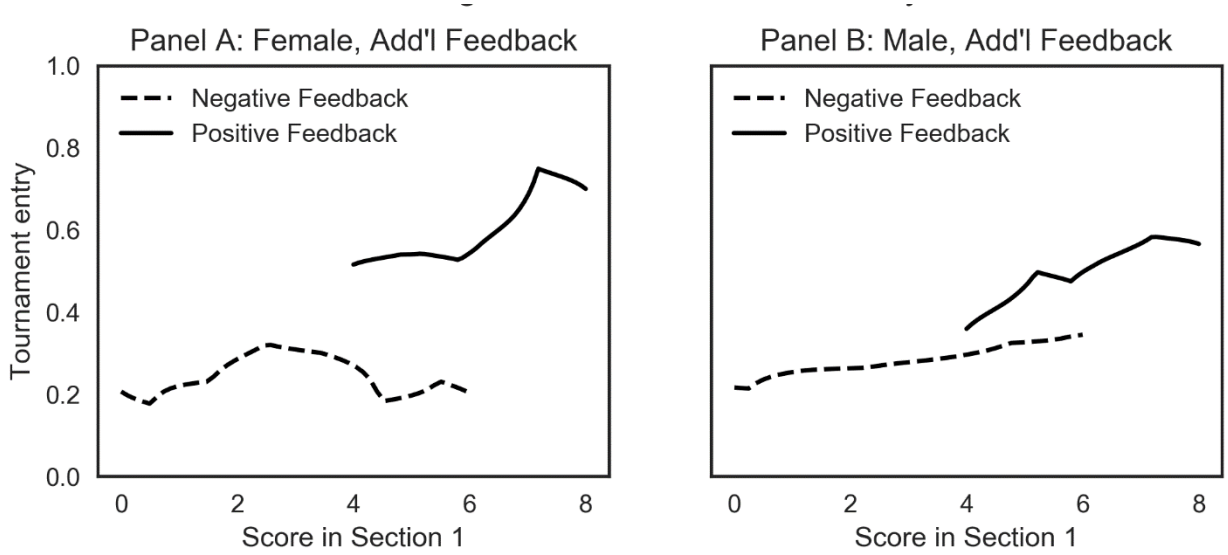
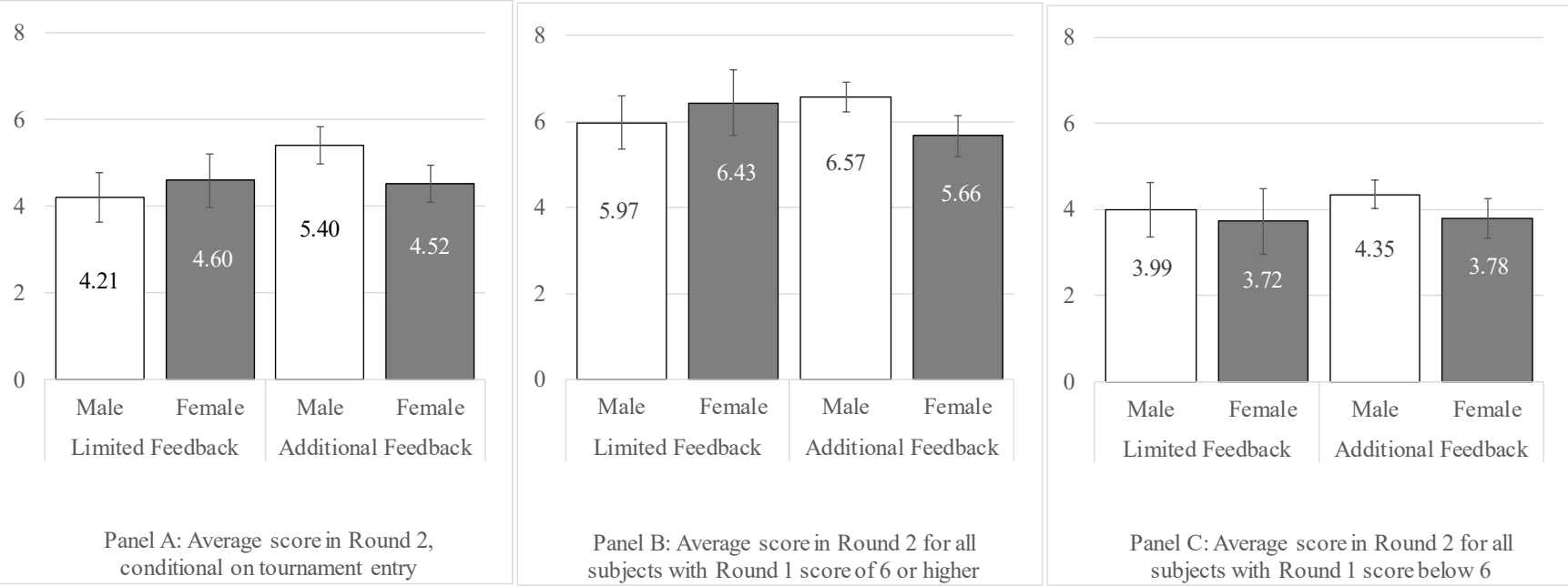


Figure 7: Gender Differences in Performance in Tournament (Round 2) in the Limited Feedback and Additional Feedback Treatments, on Average and by Round 1 Performance



Panel A: Average score in Round 2, conditional on tournament entry

Panel B: Average score in Round 2 for all subjects with Round 1 score of 6 or higher

Panel C: Average score in Round 2 for all subjects with Round 1 score below 6

Table 1: Treatment Summary

	Gender of Match Unknown (UG)	Gender of Match Known (KG)	Total Subjects
Limited Feedback (LF)	140	144	284
Additional Feedback (FF)	205	202	407
Total Subjects	345	346	691

Table 2: Summary of Demographic Characteristics and Balance Tests

	Wave 1			Wave 2	Wave 3			Wave 4		
Variables	Limited Feedback	Additional Feedback	LF vs AF p-value	Additional Feedback	Limited Feedback	Additional Feedback	LF vs. AF p-value	Limited Feedback	Additional Feedback	LF vs. AF p-value
age	42.23	40.13	0.15	38.2	36.85	36.06	0.64	38.84	39.40	0.69
female	0.59	0.55	0.59	0.5	0.48	0.46	0.77	0.51	0.49	0.73
education										
Less than high school	0.00%	0.00%		0.00%	1.00%	0.00%	0.32	0.00%	0.00%	
High school or GED	13.46%	7.92%	0.21	21.59%	8.00%	8.16%	0.97	10.00%	13.56%	0.45
Some College	25.96%	15.84%	0.08	26.14%	24.00%	22.45%	0.80	23.75%	29.66%	0.36
2-year college degree	9.62%	14.85%	0.25	12.50%	7.00%	18.37%	0.02	10.00%	11.86%	0.68
4-year college degree	35.58%	46.53%	0.11	29.55%	47.00%	40.82%	0.38	45.00%	34.75%	0.15
Master's degree	10.58%	11.88%	0.74	7.95%	10.00%	7.14%	0.48	10.00%	7.63%	0.56
Professional degree	3.85%	2.97%	0.68	1.14%	3.00%	1.02%	0.32	1.25%	1.69%	0.80
Doctoral degree	0.96%	-1.21E-17	0.40	1.14%	0.00%	2.04%	0.15	0.00%	0.85%	0.41
income										
Less than $10,000	3.85%	4.95%	0.74	4.55%	6.00%	9.18%	0.40	2.50%	5.08%	0.37
$10,000 - $19,999	11.54%	7.92%	0.36	6.82%	8.00%	10.20%	0.59	8.75%	5.08%	0.31
$20,000 - $29,999	15.38%	7.92%	0.11	18.18%	10.00%	12.24%	0.62	6.25%	13.56%	0.10
$30,000 - $39,999	6.73%	13.86%	0.11	15.91%	13.00%	16.33%	0.51	18.75%	11.86%	0.18
$40,000 - $49,999	16.35%	7.92%	0.07	11.36%	10.00%	8.16%	0.66	5.00%	11.86%	0.10
$50,000 - $74,999	22.12%	18.81%	0.56	21.59%	27.00%	23.47%	0.57	35.00%	31.36%	0.59
$75,000 - $99,999	13.46%	10.89%	0.57	9.09%	17.00%	10.20%	0.17	13.75%	12.71%	0.83
$100,000 - $149,999	6.73%	21.78%	0.00	7.95%	7.00%	8.16%	0.76	8.75%	5.08%	0.31
$150,000 - $249,999	3.85%	5.94%	0.40	4.55%	2.00%	2.04%	0.98	1.25%	3.39%	0.35
$250,000 - $499,999	0.00%	0.00%	1.00	0.00%	0.00%	0.00%		0.00%	0.00%	
race										
Asian	6.73%	11.88%	0.14	6.82%	7.00%	10.20%	0.42	7.50%	8.47%	0.81
Black or African American	4.81%	7.92%	0.35	3.41%	8.00%	7.14%	0.82	5.00%	8.47%	0.35
Native American	0.96%	0.00%	0.23	0.00%	2.00%	2.04%	0.98	1.25%	0.85%	0.78
White	84.62%	78.22%	0.20	89.77%	83.00%	77.55%	0.34	83.75%	81.36%	0.67
Other/ Do not wish to disclose	2.88%	1.98%	0.67	0.00%	0.00%	3.06%	0.08	2.50%	0.85%	0.35

Table 3: Mean Comparisons of Gender Differences in Behavioral Traits

Variable	Male	Female	Diff
Average score in Round 1	4.32	3.77	0.549***
Average bonus in Round 1	0.83	0.71	0.122***
Score Confidence	4.32	3.61	0.716***
Proportion self-evaluating below average	0.46	0.63	-0.173***
Self-Reported Risk Preference	5.68	4.57	1.111***
Number of obs.	337	352	

Notes: Significance levels *** p<0.01, ** p<0.05, * p<0.1. Summary statistics are based on data from no feedback and forced feedback conditions. Risk elicitation occurred as part of the post-experiment questionnaire and therefore took place post-treatment.

Table 4: Ordinary Least Squares Estimates of Gender Gaps in Behavioral Traits

Dependent variable	Score Confidence (Self-Reported Score in Round 1)		Self-evaluation of Payment to be Below Average	
	(1)	(2)	(3)	(4)
Female	-0.714***	-0.447***	0.168***	0.116***
	(0.149)	(0.133)	(0.0375)	(0.0360)
Score in Round 1		0.484***		-0.0946***
		(0.0368)		(0.00844)
Dependent variable mean	3.960	3.960	0.548	0.548
Observations	691	691	691	691
R-squared	0.0367	0.238	0.0408	0.161

Notes: Robust standard errors in parentheses. All specifications include wave fixed effects and controls for gender of the match. In Columns 1 and 2, score-confidence is measured on a scale of 0 to 10 (the highest possible score in Round 1 is 8). Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Determinants of Tournament Entry Decision by Treatment

Samples	Limited Feedback		Additional Feedback	
	(1)	(2)	(3)	(4)
Panel A				
Female	-0.244*	-0.213	0.0224	0.0703
	(0.138)	(0.134)	(0.111)	(0.112)
Male x Score in Round 1	0.0112	-0.00376	0.0617***	0.0592***
	(0.0228)	(0.0226)	(0.0177)	(0.0186)
Female x Score in Round 1	0.0471**	0.0381*	0.0670***	0.0653***
	(0.0230)	(0.0228)	(0.0185)	(0.0198)
Dependent variable mean	0.391	0.391	0.381	0.383
Observations	284	284	407	405
R-squared	0.0358	0.0781	0.0733	0.137
Panel B				
Female	-0.360*	-0.340*	-0.0953	-0.0383
	(0.202)	(0.196)	(0.196)	(0.210)
Male x Score Bin 2	-0.0630	-0.0599	0.00740	0.000290
	(0.198)	(0.195)	(0.178)	(0.188)
Male x Score Bin 3	-0.0860	-0.138	0.123	0.120
	(0.194)	(0.190)	(0.178)	(0.188)
Male x Score Bin 4	0.0123	-0.0523	0.248	0.214
	(0.201)	(0.198)	(0.181)	(0.196)
Female x Score Bin 2	0.151	0.173	0.154	0.159
	(0.109)	(0.106)	(0.115)	(0.123)
Female x Score Bin 3	0.218**	0.217**	0.256**	0.227*
	(0.109)	(0.106)	(0.117)	(0.126)
Female x Score Bin 4	0.386***	0.372***	0.377***	0.349**
	(0.145)	(0.137)	(0.130)	(0.142)
Dependent variable mean	0.391	0.391	0.381	0.383
F-test of male interactions (p-value)	0.835	0.787	0.0436	0.103
F-test of female interactions (p-value)	0.0520	0.0503	0.0177	0.0769
Observations	284	284	407	405
R-squared	0.0458	0.0903	0.0574	0.122

Notes: Robust standard errors in parentheses. All specifications include wave fixed effects and controls for gender of the match. Even numbered columns control for risk and confidence. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Determinants of Tournament Entry Decision by Type of Feedback Received

Samples	Women			Men		
Type of Potential Feedback	All	Negative		All	Negative	
	(1)	(2)	(3)	(4)	(5)	(6)
Additional Feedback	-0.0335	0.138	0.0480	-0.281**	-0.320*	-0.226
	(0.115)	(0.144)	(0.137)	(0.130)	(0.179)	(0.264)
Limited Feedback x Score in Round 1	0.0470**	0.0618*		-0.00132	-0.0189	
	(0.0228)	(0.0359)		(0.0221)	(0.0403)	
Add'l Feedback x Score in Round 1	0.0603***	0.0103		0.0430**	0.0405	
	(0.0199)	(0.0327)		(0.0185)	(0.0399)	
Limited Feedback x Score Bin 2			0.183*			-0.0399
			(0.103)			(0.201)
Limited Feedback x Score Bin 3			0.209*			-0.146
			(0.121)			(0.219)
Limited Feedback x Score Bin 4			0.959***			0.00845
			(0.100)			(0.270)
Add'l Feedback x Score Bin 2			0.181			0.0293
			(0.124)			(0.202)
Add'l Feedback x Score Bin 3			0.0634			0.0498
			(0.157)			(0.226)
Add'l Feedback x Score Bin 4			0.0190			0.0764
			(0.183)			(0.304)
Dependent variable mean	0.361	0.269	0.269	0.412	0.350	0.350
F-test of equality of interactions (p-value)	0.646	0.263	$p < 0.01$	0.106	0.257	0.879
Observations	352	223	223	337	177	177
R-squared	0.106	0.0526	0.0903	0.138	0.0719	0.0709

Notes: Robust standard errors in parentheses. All specifications include wave fixed effects and controls for risk, confidence and gender of the match. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Average Attribution to Own Ability by Gender and Type of Feedback Outcome

Type of Feedback	Male	Female	Difference
Negative reinforcement (Self-Evaluation of Payment Below Average and Payment Actually Below)	71.7	72.7	-1.04
Positive surprise (Self-Evaluation of Payment Below Average but Payment Actually Above)	59.4	57.5	1.88
Negative surprise (Self-Evaluation of Payment Above Average, but Payment Actually Below)	54.5	72.9	-18.39***
Positive reinforcement (Self-Evaluation of Payment above Average and Payment Actually Above)	74.7	64.7	10.05**

Notes: Summary statistics are based on data from waves 1, 2 and 4 of the experiment, additional feedback condition only (wave 3 did not contain the attribution question). Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 8: Gender Differences in the Effect of Receiving Negative Feedback on Attribution and Tournament Entry

Sample	All		Negative Self-Evaluation		Positive Self-Evaluation	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Dep Var: Attribution						
Female	-9.107**	-7.773*	-1.008	-0.206	-11.22**	-9.883*
	(4.173)	(4.329)	(7.511)	(7.697)	(4.832)	(5.126)
Male x Negative Feedback	-6.063	-2.586	13.08**	16.53*	-25.30***	-21.39***
	(3.991)	(6.050)	(6.498)	(8.941)	(5.258)	(7.721)
Female x Negative Feedback	11.74***	15.55***	15.05**	19.06**	6.772	10.51
	(4.331)	(5.263)	(6.465)	(7.882)	(6.237)	(7.479)
Score in Round 1		1.036		1.173		1.202
		(1.296)		(2.066)		(1.653)
Dependent variable mean	67.68	67.62	68.04	67.95	67.24	67.24
F-test of equality of interactions (p-value)	0.00208	0.00191	0.828	0.780	0.0000416	0.0000690
F-test of female = female X neg fbk (p-value)	0.00704	0.00281	0.206	0.129	0.0679	0.0595
Observations	309	307	169	167	140	140
R-squared	0.0391	0.0473	0.0690	0.0772	0.175	0.185
Panel B: Dep Var: Tournament Entry						
Female	0.0925	0.130*	0.125	0.127	0.0886	0.141
	(0.0749)	(0.0705)	(0.121)	(0.117)	(0.0983)	(0.0933)
Male x Negative Feedback	-0.195***	-0.0902	-0.202*	-0.143	-0.105	-0.0388
	(0.0692)	(0.0875)	(0.103)	(0.126)	(0.110)	(0.133)
Female x Negative Feedback	-0.311***	-0.161*	-0.246**	-0.145	-0.384***	-0.244*
	(0.0722)	(0.0937)	(0.102)	(0.132)	(0.112)	(0.141)
Score in Round 1		0.0388*		0.0286		0.0326
		(0.0202)		(0.0304)		(0.0303)
Dependent variable mean	0.381	0.383	0.326	0.329	0.446	0.446
F-test of equality of interactions (p-value)	0.232	0.452	0.754	0.987	0.0599	0.155
F-test of female = female X neg fbk (p-value)	p < 0.01	0.0360	0.0631	0.199	0.0108	0.0532
Observations	407	405	221	219	186	186
R-squared	0.0730	0.144	0.0765	0.120	0.0710	0.171

Notes: Robust standard errors in parentheses. All specifications include wave fixed effects and controls for gender of the match. The sample is restricted to participants who received negative feedback in the addiitonal feedback treatment and participants who would have received negative feedback in the limited feedback treatment. Columns 5 and 6 omit wave 3 since it did not contain the attribution question. Even numbered columns control for risk and confidence.

Table 9: Gender Differences in Sorting into Tournament Entry in Response to Negative Feedback

Dep Var:	Tournament Entry				Attribution	
Samples	Limited Feedback		Forced Feedback		Forced Feedback	
	(1)	(2)	(3)	(4)	(3)	(4)
Panel A						
Female	-0.335**	-0.335**	0.0991	0.126	-2.243	0.101
	(0.168)	(0.166)	(0.149)	(0.151)	(9.555)	(9.072)
Male x Score in Round 1	-0.0157	-0.0169	0.0423	0.0478	-2.928	-1.493
	(0.0400)	(0.0394)	(0.0389)	(0.0391)	(2.348)	(2.281)
Female x Score in Round 1	0.0491	0.0573	0.00419	0.0104	0.519	0.940
	(0.0365)	(0.0363)	(0.0330)	(0.0326)	(1.991)	(1.948)
Dependent variable mean	0.341	0.341	0.274	0.277	68.89	68.81
F-test of equality of interactions (p-value)	0.211	0.143	0.416	0.427	0.246	0.378
Observations	176	176	226	224	178	176
R-squared	0.0649	0.0834	0.0213	0.0475	0.0455	0.0747
Panel B						
Female	-0.346*	-0.350*	-0.141	-0.110	-11.34	-10.18
	(0.202)	(0.199)	(0.208)	(0.217)	(12.56)	(10.96)
Male x Score Bin 2	-0.0460	-0.0407	-0.00725	0.00999	-9.609	-7.725
	(0.198)	(0.199)	(0.190)	(0.197)	(9.420)	(9.047)
Male x Score Bin 3	-0.110	-0.118	0.0177	0.0169	-6.559	-2.929
	(0.214)	(0.213)	(0.216)	(0.221)	(10.17)	(9.912)
Male x Score Bin 4	0.0393	0.0319	0.0187	0.0534	-44.17***	-36.55***
	(0.267)	(0.263)	(0.291)	(0.299)	(14.10)	(13.92)
Female x Score Bin 2	0.152	0.182*	0.182	0.194	11.63	12.39
	(0.111)	(0.104)	(0.116)	(0.124)	(9.609)	(8.730)
Female x Score Bin 3	0.179	0.222*	0.0370	0.0398	9.267	10.21
	(0.127)	(0.123)	(0.153)	(0.158)	(10.58)	(9.844)
Female x Score Bin 4	0.930***	0.943***	-0.0155	-0.00541	9.347	10.86
	(0.111)	(0.104)	(0.183)	(0.181)	(13.35)	(12.53)
Dependent variable mean	0.341	0.341	0.274	0.277	68.89	68.81
F-test of male interactions (p-value)	0.890	0.874	0.997	0.998	0.00882	0.0268
F-test of female interactions (p-value)	p < 0.01	p < 0.01	0.248	0.233	0.684	0.569
F-test of equality of interactions (p-value)	p < 0.01	p < 0.01	0.557	0.571	0.0412	0.0526
Observations	176	176	226	224	178	176
R-squared	0.0952	0.112	0.0331	0.0597	0.112	0.132

Notes: Robust standard errors in parentheses. All specifications include wave fixed effects and controls for gender of the match. The sample is restricted to participants who received negative feedback in the addiitonal feedback treatment and participants who would have received negative feedback in the limited feedback treatment. Columns 5 and 6 omit wave 3 since it did not contain the attribution question. Even numbered columns control for risk and confidence. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.