DOI:
10.1016/j.jebo.2021.12.014

Document Version
Peer reviewed version

Gender differences in equilibrium play and strategic sophistication variability*

María Cubel† Santiago Sanchez-Pages‡

Abstract

We investigate the existence of gender differences in strategic sophistication in two weakly dominance solvable games where a prize is at stake. The first one is the two-person beauty contest, where strategies are numbers and players must perform mathematical operations. The second is the novel "gaze coach game", where strategies are photographs of the eye region and the two players must assign emotional states to these images. We observe that females follow equilibrium play less often in the former game but not in the latter. Males display greater strategic sophistication variability. As a result, females are underrepresented among top performers in both games.

Keywords: gender differences, strategic sophistication, competition, gender bias, variability.

JEL codes: C72; C91; D91; J16.

*We thank the editor Lionel Page, two anonymous referees, Ayala Arad, Jordi Brandts, Christoph Bühren, Syngjoo Choi, Patricia Esteve-Gonzalez, Natassa Papadopoulou, Rosemarie Nagel, Lise Vesterlund and audiences at IMEBESS Florence, M-BEES Maastricht, SAET Faro, the TIBER Symposium and at the universities of the Balearic Islands, Granada, Kent, Lancaster, Middlesex and Rotterdam for their useful comments and suggestions. All remaining errors are completely ours. Both authors acknowledge financial support from the Spanish Ministry for Science and Innovation (grant ECO2015-66281-P).

†Department of Economics, University of Bath. E-mail: maria.cubel@gmail.com

‡Department of Political Economy, King's College London. E-mail: sanchez.pages@gmail.com. URL: http://www.sanchezpages.com/.

1 Introduction

In its investigation of the reasons driving the persistence of gender differences in labor market outcomes, one strand of the literature has highlighted the role of gender differences in competitive performance. Evidence from experiments and observational data suggests that women perform worse than men in tournaments (e.g. Gneezy, Niederle and Rustichini, 2003; Gneezy and Rustichini, 2004; Backus et al. 2016). However, this effect seems to be mediated by the perceived gender-bias of the task at hand, that is, whether males or females are expected to perform better. Females perform as well as males when the task is perceived to be female-friendly (Guenther, Arslan, Schwieren and Strobel, 2010; Shurchkov, 2012; Iriberri and Rey-Biel, 2017).[1]

In addition, the literature has shown that women tend to be underrepresented among top performers in a variety of tasks. For instance, males outnumber females among best performing students in standardized maths tests (Ellison and Swanson, 2010; Breda, Jouini and Napp, 2018). This observation has been taken as evidence in favor of the Greater Variability Hypothesis (GVH), which states that males display greater variance in cognitive ability than females. Studies finding that males are overrepresented in the upper and lower tails of the distribution of scores in maths tests (Machin and Pekkarinen, 2008) and creative tasks (He and Wong, 2011) lend support to this hypothesis. However, the literature also suggests that gender differences in performance variability are not universal across domains, countries and tasks (Hyde and Mertz, 2009; Taylor and Barbot, 2021).

In this paper, we study whether perceived gender biases affect equilibrium play and outcome variability in competitive strategic interactions. By competitive strategic interactions we mean games where a prize is at stake and where being strategically sophisticated enhances one's chances of victory.

Why is this important? Strategic thinking is crucial in many human interactions. Success in social, educational and workplace interactions rests on understanding that friends, competitors, employers and co-workers adjust their

[1] See Niederle and Vesterlund (2011) and Niederle (2016) for reviews of this literature.

behavior to incentives and the behavior of others. Individuals with "theory of mind," that is, the ability to assess others' thoughts, emotions and intentions (Baron-Cohen, 1991), are better at making and maintaining social relations and obtain better educational results (Fe, Gill and Prowse, 2019). Choi, Kim and Lim (2019) show that strategic sophistication is positively related to labor and household income. In addition, they observe differential effects of strategic sophistication on personal income and in the marriage market by gender. Hence, understanding whether the perceived gender bias of strategic interactions affects the behavior of women and men can contribute to explain the prevalence of gender differences in labor market and life outcomes.

The available evidence on gender differences in strategic sophistication is mixed. Camerer, Ho and Chong (2004), Ostling, Wang, Chou and Camerer (2011) and Arad and Rubinstein (2012) find slight differences in favor of men in the beauty contest, the LUPI game and the 11-20 game respectively. Gauriot, Page and Wooders (2020) find that female professional tennis players play less in line with equilibrium predictions than male players. Cubel and Sanchez-Pages (2017) find differences in favor of women in the beauty contest when gender is primed and in favor of men under no monetary incentives. Others (e.g. Burnham et al., 2009; Brañas-Garza, Garcia-Muñoz and Hernan, 2012) find no gender differences. However, so far, no paper has studied whether the perceived gender-bias of strategic interactions affects the relative performance and outcome variability of women and men in competitive games.

Our basic hypothesis is that the perceived gender-bias of a strategic interaction might affect *game form recognition* (Chou, McConnell, Nagel and Plott, 2009). This might be driven by differences in attention and engagement, which may lead to an incomplete understanding of how combinations of strategies produce outcomes. The perceived gender bias of the interaction may also lead to different levels of *strategic awareness*, defined as the realization that playing requires reasoning about others (Fehr and Huck, 2016). Differences in strategic awareness and in game form recognition are related to Selten's (1978) "3-level theory", which broke down strategic reasoning into three levels of increasing complexity: routine, imagination and rationality. Our hypothesis is that the

perceived gender bias of the strategic interaction might affect differentially the willingness of men and women to engage in these levels of understanding, resulting in differences in strategic sophistication.

It is well-known that changes in the game form can affect behavior dramatically (e.g. Cox and James, 2012). We focus instead on changes in the nature of the strategy set and what it takes to win the game. To that end we employ two two-person competitive games. The first one is the two-person beauty contest (Grosskopf and Nagel, 2008). In this game, available strategies are numbers and winning requires a mathematical computation (approaching two thirds of the average response of the two players). This game thus aims to exploit the negative stereotype associating women to math (e.g. Nosek, Banaji and Greenwald, 2002).

The second game is the novel "gaze coach game", where participants must select a subset of imaginary players to participate in a tournament against the team selected by another participant. These made-up players are presented by photographs of their eye region. Winning requires associating emotional states to these images correctly. The design employs the Eyes test (Baron-Cohen et al., 2001), which measures the ability to attribute and recognize mental states in others. The gaze coach game thus exploits the commonly held stereotype suggesting that women are more empathic and have a greater capacity to recognize emotions. This stereotype is borne out by large studies using the Eyes test; women score slightly but significantly higher than men (Schiffer et al., 2013; Baron-Cohen et al., 2015).

Both games have a weakly dominant strategy so a player sophisticated enough to identify that strategy never loses. That these games are weakly dominance solvable also implies that beliefs about the strategic sophistication of the opponent are irrelevant. This is important for two reasons. First, because game form recognition may affect belief formation (Bosch-Rosa and Meissner, 2020). Second, because there may be gender differences in beliefs about the sophistication of other players (Cubel and Sanchez-Pages, 2017) and on how men and women act upon these beliefs (Huberman and Rubinstein, 2001).

In the two-person beauty contest, we find no gender differences in aggregate behavior or outcomes. However, we observe greater male variability in strategic sophistication among males; men are significantly more likely to pick both the weakly dominant strategy and very high numbers. As a result, there are fewer women among the bottom and top performers in the game.

The gaze coach game is perceived as more female-friendly both by participants and by an out-experiment sample. In this game, we find no significant gender differences either in the proportion of participants who pick the weakly dominant strategy nor in expected payoffs. However, once we correct for the mistakes participants made when assigning emotions to the images, we observe that women are again underrepresented among top performers.

We explore the reasons behind these results by analyzing the responses to a non-incentivized post-experiment questionnaire. We observe no significant gender differences in the type of mistakes men and women make in these games. We do find however that participants who believe that the other gender is better at the game display lower strategic sophistication than the rest. This would seem to corroborate that stereotypes and perceived gender biases influence strategic sophistication in competitive games.

The remainder of the paper is organized as follows. Section 2 describes the experimental design and results of Study 1, the two-person beauty contest. Section 3 does the same for Study 2, the gaze coach game. Section 4 concludes. The appendix contains the experimental instructions, the post-experiment questionnaire and additional tables and figures.

2 Study 1: The two-person beauty contest

2.1 Experimental design and procedure

The two-person beauty contest game was originally proposed by Grosskopf and Nagel (2008). Two players choose an integer between 0 and 100 aiming to guess two-thirds of the average number chosen in the pair. This game has a unique Nash equilibrium were both players respond zero. Choosing zero is

also a weakly dominant strategy. The game is isomorphic to one where the player who chooses the smallest number in the pair wins; hence, the lower the number a participant picks the larger their probability of winning. Contrary to the n-player beauty contest, beliefs about the strategic sophistication of others are irrelevant.[2] Any gender differences in the choice of the weakly dominant strategy can thus be safely attributed to a failure in game form recognition rather than to the gender differences in beliefs observed by Rubinstein and Huberman (2001) and Cubel and Sanchez-Pages (2017).

This study was conducted at the School of Economics and Business of the University of Barcelona in 2016. The School is large and hosts students in Economics, Business, Statistics and Sociology. Participants were recruited through posters, leaflets and class presentations and had no previous training on game theory. In total, 136 people participated in the study, 50.0% of them female. Sessions lasted between 30 and 40 minutes. Participants received a five euros show up fee and five additional euros for winning the prize in their pair (2.50 euros if both participants picked the same number). The average payment was, by construction, 7.50 euros.

The experiment was conducted in a large lecture theatre using pen and paper. We ran two sessions with 62 and 64 participants each. After arriving, participants were seated with plenty of space in between. They were asked to read the instructions (see Appendix A) along with one of the experimenters, who did so aloud. Participants played anonymously against a randomly chosen person in the session.[3] They took their decision and recorded it in their reporting sheet. When they all finished and reporting sheets were collected, participants were asked to fill up a brief non-incentivized questionnaire designed to measure beliefs, explicit gender stereotypes and types of failures in game form recognition (see Appendix A). Experimenters answered privately any questions participants had and provided no feedback at any time. At the end of the session, participants were called one by one to the main desk by

[2]For a similar approach, see the one-player beauty contest studied by Bosch-Rosa and Meissner (2020).

[3]During the session, one of the experimenters generated a random matching of participants into pairs using https://www.randomlists.com/team-generator.

their participant number. There, they were informed about the response of the participant they had been randomly matched with and were paid accordingly. Then, they signed their receipt and left the room.

2.2 Results

2.2.1 Choices and expected payoffs

Let us compare responses by gender along four dimensions: The fraction of subjects who chose the weakly dominant strategy, average response, median response and average expected payoff. To compute the latter we followed Nagel, Buehren and Frank (2017): We combined the response of each participant with the choice of each of the other participants in their session and took the average of all the outcomes by giving 1 to each win, 0.5 to draws and zero to losses.

The results of these comparisons are contained in Table 1 below. The first noticeable result is that only about one in six participants picked zero. This proportion is similar to the one Grosskopf and Nagel (2008) observed in their study (10%) and falls in between the two samples studied in Chou et al. (2009), who report that 0% of students in a community college and 46% of Caltech students chose zero.

	Choices of zero	Average	Median	Expected payoff
Males	25.00%	32.79	25.5	0.517
Females	7.35%	33.30	33	0.482
All	16.17%	33.05	30	0.5

Table 1. Aggregate results by gender and total.

Another result emerging from Table 1 is that, whereas there are no gender differences in average choices and expected payoffs, the proportion of participants who chose the weakly dominant strategy differs by gender (Proportion test, $p = 0.004$).[4]

[4] All tests reported are two-tailed unless explicitly stated.

Result 1 In the two-person beauty contest, men chose the weakly dominant strategy more often than women.

2.2.2 Variability in strategic sophistication

Result 1 would suggest that males display higher strategic sophistication than females in this game where strategies are numbers and winning requires a mathematical operation. But the absence of significant gender differences in average responses, median responses (despite the large gap) and expected payoffs indicates that males should also be overrepresented at the upper tail of the distribution.

Figure 1 below shows the kernel density and cumulative distribution of responses. Both panels show that the fraction of participants who chose extreme numbers is larger for men than for women. The variance of male responses is 1.89 times larger than the one for females. The variance-comparison test shows that the male/female variance ratio is indeed higher than one ($p = 0.005$). The greater variability in male responses can also be seen in the cumulative distribution of responses for males, which crosses the one for females from above. The Davidson and Duclos (2000) test of stochastic dominance corroborates that the differences between the two curves are significant in opposite directions for very low and very high numbers (Table A1 in Appendix B).[5]

[5]The null hypothesis of this test is nondominance. It proceeds by comparing the two distributions at several points. The nondominance null is rejected, and one distribution first stochastically dominates the other, if for all comparison points for which differences between the two distributions are statistically significant the sign of these differences is identical. We report the results of all comparisons employing this test in Appendix B.

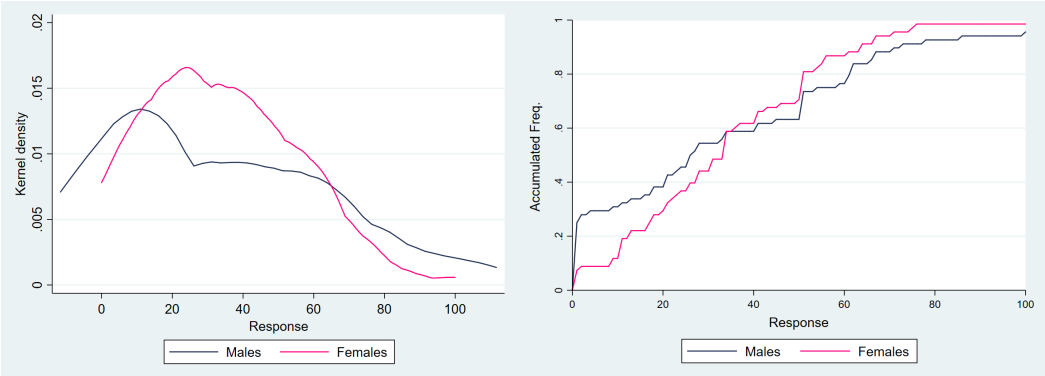

Figure 1: Cumulative distribution of responses by gender.

Figure 2 performs the same exercise for expected payoffs. The left panel shows that the distribution for males is bimodal, a result of these participants more frequently picking very low and very high numbers. Importantly, their expected payoffs also have a larger variance than the one for females. The male-female variance ratio in expected payoff is 1.60, which is again significantly larger than one (Variance-comparison test, $p = 0.027$). The right panel shows another significant difference: The cumulative distribution of expected payoffs obtained by males first order stochastically dominates the one for females. The Davidson-Duclos test confirms that dominance takes place at the high payoffs range (see Table A2 in Appendix B). This implies that women are underrepresented among top performers in the two-player beauty contest game.

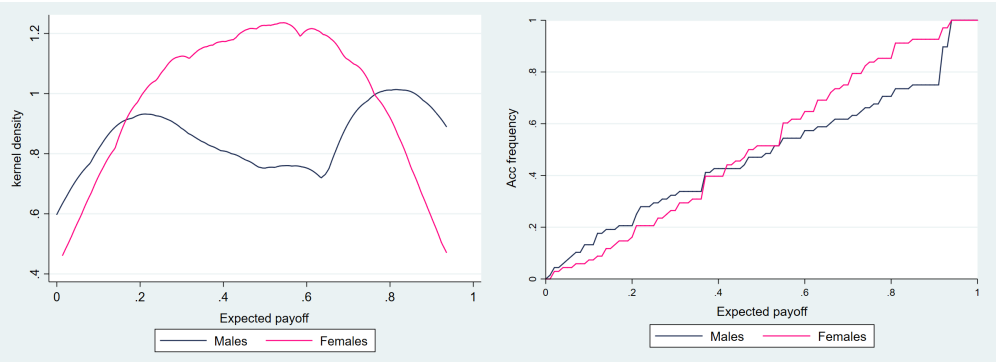

Figure 2: Density and cumulative distributions of expected payoffs by gender.

To investigate this further, we next plot the percentage of females at several top percentage thresholds. Perfect equality in performance by gender would require a 50% of females at all top percentages in the distribution (as they comprised 50% of the sample). However, Figure 3 shows that the proportion of women is around 50% for top percentages up to the top 45%. It declines sharply after that and becomes significantly lower than 50% for the top 15% (one-sample proportion test, $p = 0.010$) and beyond. Despite half of participants in our study were females, they only represented a quarter of those who performed in the top 15%.

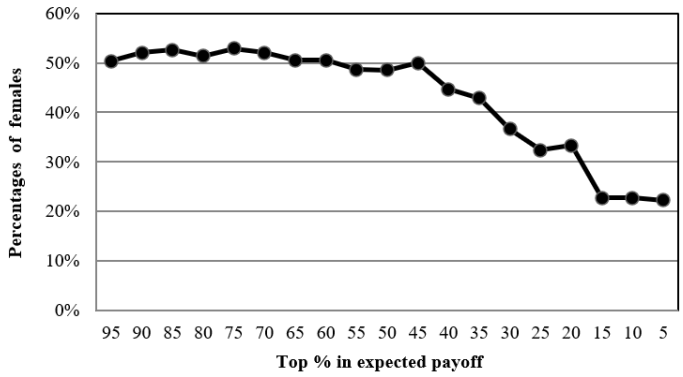

Figure 3: Percentage of females by top percentage of expected payoff.

Result 2 In the two-person beauty contest, the distribution of expected payoffs for males has a greater variability than the distribution for females. As a result, women are significantly underrepresented among top performers.

The patterns we observe in the two person beauty contest game are strikingly similar to those in PISA scores for math and reading in most OECD countries, including the most recent wave, where boys display more extreme performances than girls (Machin and Pekkarinen, 2008; OECD, 2020). The underrepresentation of females among top performers in the game is also highly similar to the pattern of scores observed in math exams and competitions. Ellison and Swanson (2010) show that males outnumber females by a ratio of

two to one among students scoring 800 in the SAT Math. This representation gap is even more acute at the top 1% of performers in American Math Competitions, where the male-female ratio exceeds ten to one.

2.2.3 Beliefs and gender bias

To delve into these results, we next analyze the responses to the questionnaire we administered at the end of the experiment. The questions aimed to understand how participants made their choices and to elicit beliefs about the perceived gender bias of the game. These beliefs were elicited directly similarly to Coffman (2014) and Bordalo et al. (2019). Non incentive-compatible belief elicitation is not uncommon either in experiments with dominance solvable games (e.g., Ryvdal et al., 2009). Our aim was not to study whether our subjects acted on their beliefs about the expected behavior of their opponent. We were interested instead in the presence of a perceived gender bias (if any) in our two games, and in its potential association with game form recognition and the willingness to engage in strategic thinking. For that reason, we devote special attention below to subjects' adherence to equilibrium play conditional on their views on the relative strategic ability of the other gender.

Previous research has indeed shown that gender-based beliefs are related to differential behavior in tournaments and strategic interactions.[6] We cannot rule out that participants used their responses to our non-incentivised questionnaire to rationalize their choices despite the absence of interim feedback. For that reason, the relationship between beliefs and actions we discuss in this subsection are only correlational.

Contrary to our expectations, the answers to the question "Which sex is better at this game?" do not suggest that participants perceived the two-person beauty contest to be male-biased. Figure 4 below breaks down their responses by gender. The distribution displays no significant gender differences

[6] Cubel and Sanchez-Pages (2017) found that women who believed females are better in the beauty contest game displayed higher depth of reasoning in the game when gender was made salient. Bordalo et al. (2019) showed that gender stereotypes correlate with differential performance in a teamwork task. Finally, Hernandez-Arenaz (2020) observed that beliefs about the gender bias of a task affect self-selection into a high-paying tournament scheme.

(Chi-square, $p = 0.170$). More participants believe that females are better at this game than in the other way around. A very similar picture emerges in an out-experiment sample coming from the same population ($n = 134$), who responded to this question online (see Figure A1 in Appendix C).[7] This would seem at odds with the gender differences we observe in behavior. Note however that more women than men thought that men are better at the game.

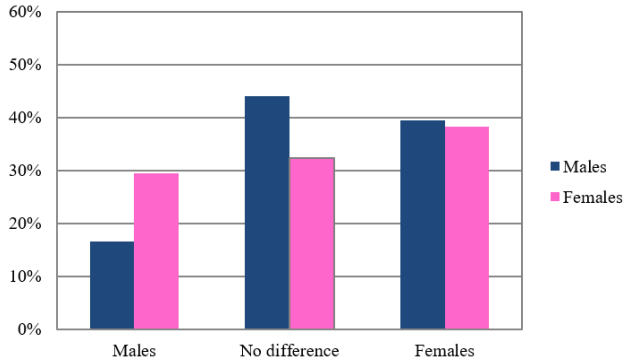

Figure 4: Responses to "Which sex is better in this game?" by gender.

Beliefs about which sex is better at the game correlate with strategic sophistication, as Table 2 below shows. Participants who thought the other sex is better at this game chose the weakly dominant strategy significantly less often than the rest of participants (one-tailed proportion test, $p = 0.041$). This difference is marginally significant by gender ($p = 0.061$ for males; $p = 0.068$ for females). In the case of females, the difference is most striking since no woman who believed that men are better at the game picked zero.

	Other sex better	Rest
Male	15.4%	32.5%
Female	0%	10.4%
All	8.7%	20.5%

[7] The questionnaire was sent to people in our subject pool who had not taken part in the experiment. They answered questions 2 to 4 in the questionnaire at the end of Appendix B. Four 10 euro prizes were randomly awarded to those who participated in the survey.

Table 2. Choices of zero by gender and response to "Which sex is better in this game?".

This result corroborates that players' beliefs about the relative superiority of their group in strategic interactions relate to their strategic sophistication. Cubel and Sanchez-Pages (2017) observed in the n-player beauty contest that women who believed their gender to be better at the game displayed greater depth of reasoning when gender was primed. We observe a similar association for all participants and in the absence of gender priming. The results of Study 2 below offer further evidence of this relationship. For that reason, we postpone formally stating this as a result until Section 3**,** where we also discuss the issue of a desirability bias in the responses to the questionnaire**.**

2.2.4 Failures in game form recognition

Participants who did not select the weakly dominant strategy failed to recognize the form of the game. This failure might range from a lack of strategic awareness to a misunderstanding of the relationship between choices and outcomes. To analyze this, we look next at the responses to the question in the post-experiment questionnaire "How did you choose your answer?"

One researcher outside the team of authors and not affiliated to their institutions coded the responses to this question.[8] The coder had no access to the choice of the subjects in the experiment. Following Chou et al. (2009), responses were classified into four type of failures: Lack of attention, strategic unawareness, unclear rules about how the winner is determined, and use of beliefs about the behavior of the opponent. These categories are illustrated in the selection of responses contained in Table 3 below.

Table 3. Examples of failures in game form recognition from the post-experiment questionnaire.

[8] The coder left too brief or vague answers unclassified (9 out of the 117 participants who did not pick zero).

Although these categories are constructed from self-reports, participants in these groups behaved differently. Responses and expected payoffs for the four types of error are not drawn from distributions with the same median (Median test, $p = 0.014$ and $p = 0.004$). Participants whose responses to the questionnaire indicate that they entertained beliefs about the sophistication of the opponent picked lower numbers (median 26) and had the highest average expected payoffs. This would suggest that these participants performed better in the game than those who failed to pick zero for other reasons.[9]

Let us now study gender differences in game-form recognition. Figure 5 contains the frequencies of type of failure by gender. The most common type of failure in the sample as a whole is the use of beliefs about the behavior of others (29.0%) followed very closely by the unclear understanding of the rules of the game (28.0%). Females seem to fall more frequently into this error than men. On the other hand, men are slightly more likely than women to commit errors due to inattention. This would seem to be consistent with the greater male variability in male choices and payoffs. That said, the two distributions of answers are not statistically different (Chi-square test, $p = 0.428$). We will observe the same in Study 2 so, for that reason, we postpone stating this result to Section 3.

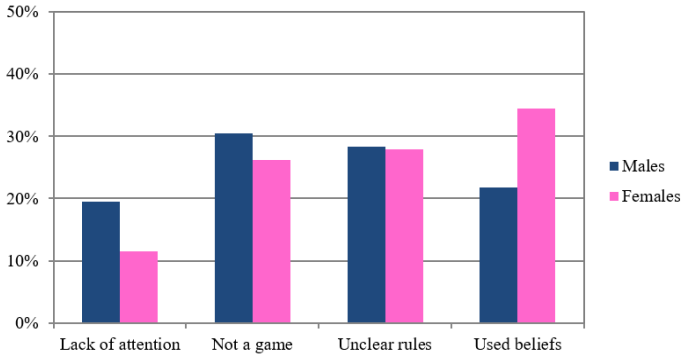

Figure 5: Failures in game form recognition by gender.

[9] The distributions of responses and expected payoffs for these participants is statistically different from those for the other participants who selected weakly dominated strategies (Mann-Whitney, $p = 0.001$ and $p = 0.002$, respectively).

3 Study 2: The gaze coach game

3.1 Experimental design and procedure

In the gaze coach game, participants played the role of coaches who must select two out of a set of four imaginary players to participate in a tournament against the team selected by another participant. The game is inspired by Arad (2012) but, unlike in her case, there is no explicit ranking of players and their order is irrelevant since each player in a team plays a match against each of the two players of the opponent's team. Thus, a tournament between a pair of participants/coaches is composed by four matches. A win is worth one point; a draw, 0.5 points and a loss earns zero points. The winner of the tournament is the participant whose pair of players obtains most points.

Figure 6 below shows the outcome matrix for the game. This matrix details the result of all possible matches between any pair of players. Participants/coaches were presented with photographs of the eye regions of the four players available. Among the six possible strategies, the weakly dominant one is to choose the two players at the bottom of the matrix. Therefore, like the two-person beauty contest, the gaze coach game is a weakly dominance solvable game where beliefs about the sophistication of the opponent play no role.

Figure 6. Outcome matrix for the gaze coach game.

However, rather than choosing photographs, participants must choose the word that they think best describes what the player they want to select is thinking or feeling. The set of four words they were given was "Uneasy", "Cautious", "Anticipating", and "Contemplative". Participants were told that each photograph corresponded to only one of the words. Hence, to select a set of players, participants had to recognize correctly their emotional states as portrayed in the images.

Both the images and their associated words came from the Spanish version of the Eyes test.[10] The original Eyes test (Baron-Cohen, Jollife, Mortimore and Robertson, 1997) was designed as an adult test for autism and Asperger Syndrome. This test has become a popular measure of theory of mind and mentalizing capabilities since it requires the attribution and recognition of mental states in others. The complete Eyes test entails matching each of 36 photographs of the eye region to the word within a set of four that best describes the mental or emotional state of the person in the image. The set of words changes across items.

We designed the gaze coach game to be perceived as a female-biased interaction. This expectation was based on 1) the stereotype who sees women as more empathic and better at recognizing emotions in others; 2) the results of the Eyes test, where neurotypical women outperform men slightly but significantly (Schiffer et al. 2013; Baron-Cohen et al., 2015); and 3) the results of a preliminary survey we ran with an out-sample population (n=86), where only 1.2% of participants believed males are better at the test.

To select the photographs of the four players we employed in the experiment, we administered the Eyes test to the aforementioned out-sample group. We did not find any gender difference in average scores, but we did find differences in the percentage of correct responses in some items. We discarded these. From the remaining ones, we selected the photographs of two men and two women whose associated words mixed emotions with positive and negative connotations: "Anticipating" is the top male player in Figure 6 whilst "Uneasy" is the bottom one; "Cautious" is the top female player and "Contemplative" the bottom one. The weakly dominant strategy, which was to select {Uneasy, Contemplative} ({U,C} henceforth), also mixed qualities with potentially opposite connotations.

Note however, that a failure to select the weakly dominant strategy in the experiment may come from a failure in game form recognition or/and a

[10] Available at https://www.autismresearchcentre.com/tests/eyes-test-adult/. The four words in Spanish corresponding to the images were "Inquieto", "Cauto", "Expectante" and "Reflexivo".

failure in correctly associating photographs to words. To disentangle these two types of error, we asked participants to assign each of the four words to one of the four images. Subjects did this after they had selected their two preferred players for their team. The identification task was incentivized with 20 euro cents per correct answer. Answers in this task allow us to infer the pair of players participants believed they were selecting.

The experiment was ran at the School of Economics of the University of Barcelona in 2017. A total of 168 participants with no training in game theory, 51.8% of them females, took part in the experiment. No subject who participated in Study 1 took part in Study 2. We conducted a total of four sessions with 36-56 participants each. Sessions lasted between 40 and 50 minutes. Participants received a flat fee of five euros and five additional euros if they were the winner in their pair (2.50 euros if they drew). The average payment was 7.90 euros, 7.50 euros from the main experiment (by construction) plus 40 euro cents in average for the identification task. The rest of procedures for Study 2 were identical to those in Study 1.

3.2 Results

3.2.1 Choices and expected payoffs

Next, we compare responses by gender along two dimensions: The fraction of participants who chose the weakly dominant strategy and participants' expected payoff. The latter is again computed by averaging the scores a participant would have obtained by combining their strategy with each of the strategies chosen by the rest of participants in their session.

We consider two versions of these two variables. One is based on the choices recorded by participants in their reporting sheets. The other version removes the error in the identification task. Recall that after they recorded their choices, participants were asked to match the image of each player with one of the four words we had given them. With this we can obtain their "true" choice, that is, the pair of players participants thought they were selecting. We then use this "true" choice to compute their "true" expected payoff. To

reiterate, the difference between both versions, recorded vs. "true", is that the "true" version contains only the error in game form recognition whereas the one based on recorded choices also includes the errors in emotion recognition, i.e. the error in matching words to photographs. To clarify, when constructing the "true" expected payoff, we used the "true" choices of all participants in the session. Alternatively, we could have used only the "true" choice of the player in question whilst keeping the rest of players at their recorded choices. Although interesting, that counterfactual presents the problem that players with the same "true" choice in the same session could be assigned different expected payoffs.

Table 4 below shows average results by gender. The first main result emerging from that table is that the proportion of participants who picked the weakly dominant strategy -one out of five participants- was very similar to the one we observed in Study 1. The proportion doubles but remained relatively low for "true" choices, i.e. after we correct the error participants committed when matching words to photographs. The difference in the proportion of participants who believed they picked {U,C} from those who actually picked it is statistically different for the whole sample and by gender, indicating that the errors in emotion recognition were indeed frequent (McNemar test, $p < 0.001$ overall and for males, $p = 0.002$ for females). Still, it is surprising that less than half of participants managed to find the weakly dominant strategy in a game with context and with a relatively small strategy space.[11]

	Females	Males	All
Choices of {U,C}	19.8%	20.5%	20.1%
"True" choices of {U,C}	40.7%	48.2%	44.4%
Expected payoff	0.511	0.487	0.5
"True" expected payoff	0.487	0.515	0.5

Table 4: Aggregate results by gender and total.

[11] As a point of comparison, Chou et al. (2009) report that when they contextualized the two-player beauty contest as a battle between two generals, only 46% of the community college students sampled picked zero.

The second result emerging from Table 4 is that there are no significant gender differences in the proportion of participants who selected the weakly dominant strategy (proportion test, $p = 0.907$ for recorded choices; $p = 0.326$ for "true" choices) nor in expected payoffs (Mann-Whitney test, $p = 0.581$ for expected payoffs based on recorded choices; $p = 0.279$ for "true" expected payoffs). We observe no gender differences in the identification task either (Chi-square, $p = 0.374$), although this may be attributed to the small number of items participants had to identify.

Result 3 There are no significant gender differences in choices and expected payoffs in the gaze coach game.

Taken together, results 1 and 3 are consistent with (but not conclusive proof of) the idea that the gender bias of competitive games has an impact on the relative strategic sophistication of men and women. In the two-person beauty contest, where strategies were numbers and maths were important, we observed that men adjusted better than females to equilibrium play. In contrast, in the gaze coach game, where players must recognize emotions in others correctly, we found no gender differences in the frequency of equilibrium play.

3.2.2 Variability in strategic sophistication

There are no gender differences in the distribution of expected payoffs based on recorded choices. Both distributions are bimodal. This is due to the weakly dominant strategy being the modal choice (44.4%) and to a significant fraction of participants (19%) picking the "cautious" player, the worst of the four. We explore this failure further in section 3.2.4.

The male/female variance ratio for "true" expected payoffs is 1.15, which is not significantly larger than one ($p = 0.255$). But it is again the case that a look beyond aggregate results reveals differential variability in strategic sophistication by gender: As in the two-person beauty contest, the upper tail of the distribution for males is fatter (see left panel of Figure 7). According to the Davidson-Duclos test, the distribution of "true" payoffs for males first

order stochastically dominates the one for females (see right panel of Figure 7 and Table A3 in Appendix B). The dominance takes place at the range of higher payoffs. This suggests that women were underrepresented among top performers in the gaze coach game too.

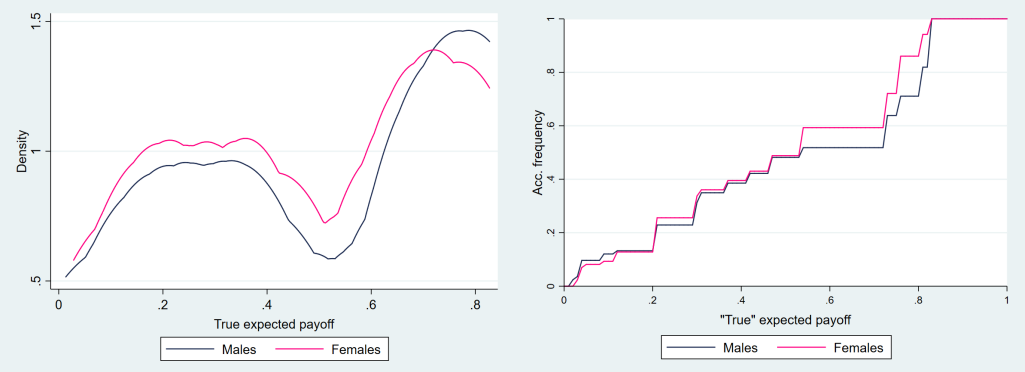

Figure 7. Density and cumulative distributions of "true" expected payoff by gender.

To corroborate this finding, we next plot the percentage of females by top percentage of expected payoff similarly to what we did in Figure 3. For expected payoffs based on recorded choices, which include the error in emotion recognition, the percentage of women across all top percentages is never significantly different from 50%.[12] However, for the "true" expected payoff, that is, once we remove the errors in emotion recognition, the proportion of women at top percentages starts declining again from the top 45% until becoming significantly lower than 50% at the top 20% (Proportion test, $p = 0.045$). As in the two-person beauty contest, only a meagre quarter of performers in the top 10% are women.

[12] Recall that the proportion of women in this sample was 51.8%. By setting the threshold at 50% we are being conservative when estimating a potential representation gap against women.

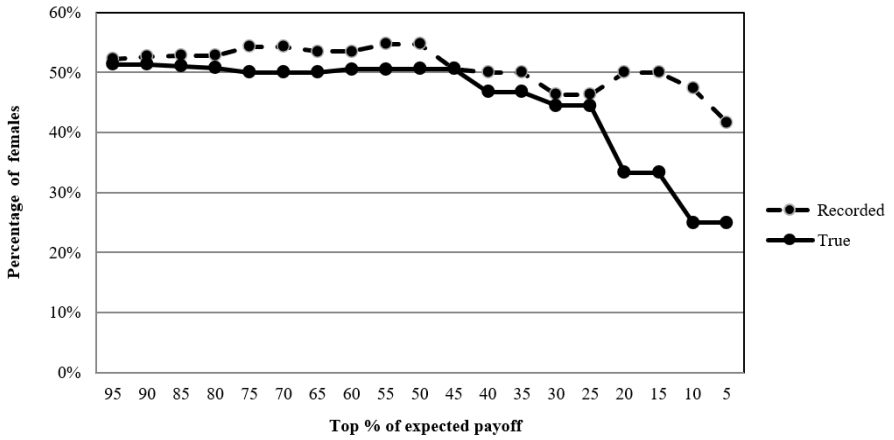

Figure 8. Percentage of females by top percentage of expected payoff.

Result 4 Women are significantly underrepresented among the top performers in the gaze coach game.

This result suggests that the absence of gender differences in the distribution of expected payoffs based on recorded choices might be due to the relative advantage of women when associating the players' photographs with their emotional state. Once we remove errors in emotion recognition, the underrepresentation of women among top performers reappears. One reason for this might be that the basic interaction underlying the gaze coach game, encapsulated in the outcome matrix in Figure 7, might have not been perceived as female-friendly. We investigate this next.

3.2.3 Beliefs and gender bias

The post-experiment questionnaire for Study 2 contained the same questions as the one for Study 1 with the exception of the last item, which was replaced by the question "Which sex is better in the identification task?" Figure 10 shows the distribution of answers to that question and to "Which sex is better at this game?"

A minority of participants believed that men are better in the gaze coach game. Actually, the aggregate distribution of responses does not differ from the

one for the analogous question in Study 1 (Chi-square, $p = 0.567$). There are no significant gender differences in the distribution of answers to the question ($p = 0.648$). This is in sharp contrast with answers to the question about the identification task, which confirms the stereotype associating women to better emotion recognition. As in the out-experiment sample, a large majority of participants (75.2%) believed that women are better at associating emotional states to the images provided.

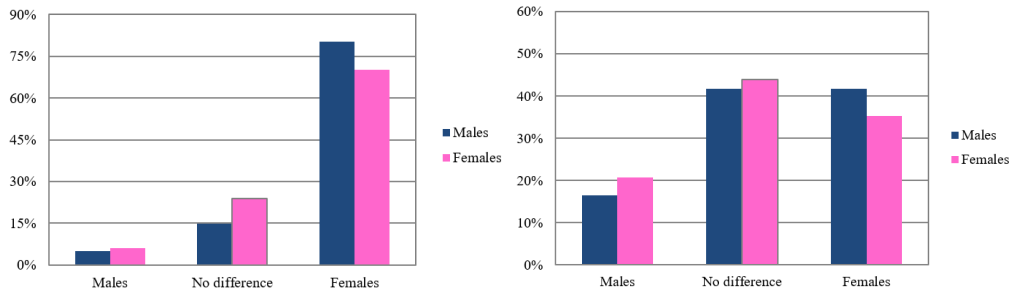

Figure 9. Responses to "Which gender is better in the identification task?" (left) and "Which sex is better in this game?" (right) by gender.

The next step is to explore whether beliefs about which sex is better at this game correlate with strategic sophistication. Table 5 below shows that participants who believed that the other sex is better at the game selected the weakly dominant pair of players significantly less often than the rest of participants (one-tailed proportion test, $p = 0.004$). This difference is also significant for men ($p = 0.011$) but only marginally for females ($p = 0.056$). However, when we look at participants' "true" choices, we find no differences by belief ($p = 0.452$). This would suggest that believing that the other sex is better at this game correlates with difficulties in emotion recognition rather than in game form recognition.

	Other sex better	Rest
Males	9.09%	30.43%
Females	5.88%	23.07%
All	8.00%	26.12%
All "True"	45.04%	44.00%

Table 5. Choice of {U,C} by gender and response to "Which sex is better in this game?"

However, the distribution of correct associations in the identification task does not differ by response to this question (Mann-Whitney, $p = 0.555$). In addition, the belief that the other sex is better at the game correlates with lower expected payoffs based on recorded choices *and* with lower "true" expected payoffs. In both cases, the distribution of expected payoffs of participants who believe the other sex is better is first-order stochastically dominated by the distribution for the rest of participants, as the Davidson-Duclos test confirms (see Figure A2 and Table A4 in Appendix B). This would suggest instead that the perceived gender-bias affected game form recognition too.

In any case, these observations together with the analogous ones in Section 2.3 for Study 1 lead us to state the following result on the relationship between strategic sophistication and the perceived gender bias of strategic interactions.

Result 5 In both the two-player beauty contest and the gaze coach game, participants who believe the other sex is better in the game choose the weakly dominant strategy less often.

Some words of caution are in order here. We cannot discard the presence of a socially desirability bias in the responses to the question about which sex was better in each game. A social desirability bias could have led subjects believing their own gender was better to respond "No difference" instead, or even that the other gender was better. This bias could be behind the lack of a clear perceived gender bias in our two games. Still, we believe this issue does not significantly affect our Result 5. The reason is that it is plausible

to assume that if a participant saw answering that the own gender was better in the game as a socially undesirable answer, they would have opted for "No difference" rather than for picking the other gender, as the distance between the former option and their true belief was smaller. Under that assumption, one could remain relatively confident that subjects who stated that the other gender was better at the game truly believed so.

3.2.4 Failures in game form recognition

In the last part of our analysis, we explore the pattern of failures in game form recognition and whether they exhibit any gender differences. To this end, we study the answers to the open-ended question "How did you choose your answer?" in the post-experiment questionnaire.

We only considered the responses of participants whose "true" choice was a dominated strategy. This is because, unlike recorded choices, "true" choices only contain error in game form recognition. As in Study 1, these answers were coded by an external researcher. Two broad differences emerge in the type of failures we observe. First, due to the simpler nature of the gaze coach game compared to the two-player beauty contest, no participant explicitly reported to have seen the gaze coach game as a game of chance. Second, a new type of error emerges. A significant fraction of participants ignored the outcome matrix and chose their players based on the qualities they associated to each of the four words. They selected the words they thought would create a good team.[13]

As in Study 1, the different types of failure in game form recognition are associated with different expected payoffs. "True" expected payoffs for the three types of failure are not drawn from distributions with the same median (Median test, $p = 0.031$). The lowest expected payoffs were obtained in average by participants who followed the words rather than the outcome matrix.[14]

[13]Some examples that illustrate this type of failure are: "Cautious players contribute calmness to teams at critical points in a tournament"; "I selected the two adjectives which sounded more risk averse"; "Being cautious and contemplative are winning qualities in sports".

[14]The distribution of "true" expected payoffs for these participants is statistically different

Figure 10 depicts the distribution of types of failures by gender. The most common type of failure for both men and women (52.8%) is a lack of understanding of the rules of the game, i.e. how choices produce outcomes. In contrast with the two-person beauty contest, very few participants seemed to have entertained any belief about the behavior of opponents. This is likely due to the lower strategic complexity of the game, e.g., its much smaller strategy space. Again, we observe no significant gender differences in types of failure (Chi-square, $p = 0.296$).

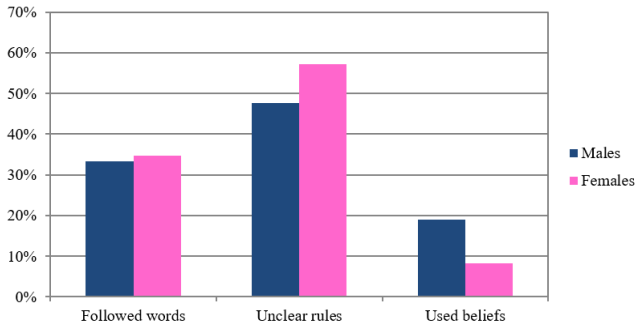

Figure 10: Failures in game form recognition by gender.

Result 6 There are no significant gender differences in failures in game form recognition in either of the two games.

4 Conclusion

Our results show that the perceived gender bias of strategic interactions affects the behavior of men and women. Men picked more often the weakly dominant strategy than women in the beauty contest, where strategies are numbers and players must perform a mathematical operation, but not in the gaze coach problem, where winning requires assigning emotions to images correctly. In line with previous results (Cubel and Sanchez-Pages, 2017), we observe that

from that for the other participants who failed to pick the weakly dominant strategy (Mann-Whitney, $p = 0.002$).

individuals who expected the other gender to be better at the game displayed lower strategic sophistication. Taken together, these findings suggest that women are more likely to play according to the equilibrium when they perceive the strategic interaction to be more female-friendly. However, it is important to bear in mind that a significant fraction of our participants had wrong beliefs about the actual gender bias of the games.

It is also worth noting that the issues with our post-experiment questionnaire that we have discussed throughout the paper could have been mitigated had we incentivized it. This was a conscious choice. We thought that the question "Which sex is better at this game?", even if vague, could elicit subjects' stereotypes in a more direct way than an incentivized question targeting a specific statistic. We acknowledge that the inference we can reasonably draw from the responses elicited this way is limited.

Although the two games that we have studied in this paper are admittedly not isomorphic, our results indicate that the underrepresentation of women among top performers in a variety of tasks across countries translates into a pervasive underrepresentation of women among top performers in competitive games. A more female-friendly interaction -in terms of the nature of strategies and the processes required to win- did not seem to be powerful enough to fully eliminate or reverse these differences. Women remained unsettlingly underrepresented among top performers even when the game entailed a task overwhelmingly perceived as female-biased. Based on this evidence, one logic next step would be to analyze the existence of gender differences in strategic behavior in non-competitive games; Thöni, Volk and Cortina (2021) already observed greater male variability in contributions in a public good game. Another open avenue for future research is whether changes in the (observable) gender composition of the set of players has an impact on the gender differences in variability of strategic sophistication we have found in this study.

References

[1] Arad, A. 2012. The tennis coach problem: A game-theoretic and experimental study, *The B.E. Journal of Theoretical Economics*, 12(1), article 10.

[2] Arad, A., and Rubinstein, A. 2012. The 11-20 money request game: A level-k reasoning study, *American Economic Review*, 102(7):3561-3573.

[3] Backus, P., Cubel, M., Guid, M., Sanchez-Pages, S., and Mañas, E. 2016. Gender, competition and performance: Evidence from expert chess players. Unpublished manuscript.

[4] Baron-Cohen, S. 1991. Precursors to a theory of mind: Understanding attention in others. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading.* Oxford: Basil Blackwell.

[5] Baron-Cohen, S., Jollife, T., Mortimore, C., and Robertson, M. 1997. Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome, *Journal of Child Psychology and Psychiatry*, 38(7):813-822.

[6] Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. 2001. The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism, *Journal of Child Psychology and Psychiatry*, 42(2):241-251.

[7] Baron-Cohen, S., Bowen, D.C., Holt, R.J., Allison, C., Auyeung, B., Lombardo, M.V., Smith, P., and Lai, M-C. 2015. The "reading the mind in the eyes" test: Complete absence of typical sex difference in ~400 men and women with autism, *PLoS ONE*, 10(8):e0136521.

[8] Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. 2019. Beliefs about gender, *American Economic Review*, 109(3):739-73.

[9] Bosch-Rosa, C., and Meissner, T. 2020. The one player guessing game: a diagnosis on the relationship between equilibrium play, beliefs, and best responses, *Experimental Economics*, 23:1129–1147.

[10] Brañas-Garza, P., Garcıa-Muñoz, T., and Hernan, R. 2012. Cognitive effort in the beauty contest game, *Journal of Economic Behavior and Organization*, 83(2):254–260.

[11] Breda, T., Jouini, E., and Napp, C. 2018. Societal inequalities amplify gender gaps in math, *Science*, 359 (6381):1219-1220.

[12] Burnham, T., Cesarini, D., Johannesson, M., Lichtenstein, P., and Wallace, B. 2009. Higher cognitive ability is associated with lower entries in a p-beauty contest, *Journal of Economic Behavior and Organization*, 72(1):171–175.

[13] Camerer, C.F., Ho, T-H., and Chong, J.K. 2004. A cognitive hierarchy model of games, *Quarterly Journal of Economics*, 119(3):861-898.

[14] Choi, S., Kim, S., and Lim, W. 2019. Strategic thinking skills and their economic importance. Unpublished manuscript.

[15] Chou, E., McConnell, M., Nagel, R., and Plott, C.R. 2009. The control of game form regognition in experimetns: understanding dominant strategy failures in a simple two person "guessing" game, *Experimental Economics*, 12:159-179.

[16] Coffman, K.B. 2014. Evidence on self-stereotyping and the contribution of ideas, *Quarterly Journal of Economics*, 129(4):1625–60.

[17] Cox, J.C., and James, D. 2012. Clocks and trees: Isomorphic Dutch auctions and centipede games, *Econometrica*, 80(2): 883–903.

[18] Cubel, M., and Sanchez-Pages, S. 2017.Gender differences and stereotypes in strategic reasoning, *The Economic Journal*, 127(601):728–756.

[19] Davidson, R., and Duclos, J-Y. 2000. Statistical inference for stochastic dominance and for the measurement of poverty and inequality, *Econometrica*, 68(6):1435-1464.

[20] Ellison, G., and Swanson, A. 2010. The gender gap in secondary school mathematics at high achievement levels: Evidence from the American mathematics competitions, *Journal of Economic Perspectives*, 24(2):109-128.

[21] Fe, E., Gill, D., and Prowse, V. 2019. Cognitive skills, strategic sophistication, and life outcomes. Unpublished manuscript.

[22] Fehr, D., and Huck, S. 2016. Who knows it is a game? On strategic awareness and cognitive ability, *Experimental Economics*, 19:713–726.

[23] Gauriot, R., Page, L., and Wooders, J. 2020. Expertise, gender, and equilibrium play. Unpublished manuscript.

[24] Gneezy, U., Niederle, M., and Rustichini A. 2003. Performance in competitive environments: gender differences, *Quarterly Journal of Economics*, 118:1049–74.

[25] Gneezy, U., and Rustichini, A. 2004. Gender and competition at a young age, *American Economic Review*, 94:377–81.

[26] Grosskopf, B. and Nagel, R. 2008. The two-person beauty contest, *Games and Economic Behavior*, 62(1):93-99.

[27] Guenther, C., Arslan, N., Schwieren, C., and Strobel, M. 2010. 'Women can't jump' – An experiment on competitive attitudes and stereotype threat', *Journal of Economic Behavior and Organization*, 75:395-401.

[28] He, W., and Wong, W. 2011. Gender differences in creative thinking revisited: Findings from analysis of variability, *Personality and Individual Differences*, 51(7):807–811.

[29] Hernandez-Arenaz, I. 2020. Stereotypes and tournament self-selection: A theoretical and experimental approach, *European Economic Review*, 126:103448.

[30] Huberman, G., and Rubinstein, A. 2001. Correct belief, wrong action and a puzzling gender difference. Unpublished manuscript.

[31] Hyde, J.S., and Mertz, J.E. 2009. Gender, culture, and mathematics performance, *Proceedings of the National Academy of Sciences*, 106(22):8801-8807.

[32] Iriberri, N., and Rey-Biel, P. 2017. Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision, *Journal of Economic Behavior and Organization*, 135: 99-111.

[33] Machin, S., and Pekkarinen, T. 2008. Global sex differences in test score variability, *Science*, 322(5906):1331-1332.

[34] Nagel, R., Buehren, C., and Frank, B. 2017. Inspired and Inspiring: Hervé Moulin and the discovery of the beauty contest game, *Mathematical Social Science,* 90:191-207.

[35] Niederle, M. 2016. Gender. In *Handbook of Experimental Economics, volume 2*, J. Kagel and A.E. Roth (Eds.), Princeton, NJ: Princeton University Press.

[36] Niederle, M., and Vesterlund, L. 2011. Gender and Competition, *Annual Review of Economics*, 3: 601-630.

[37] Nosek, B.A., Banaji, M.R., and Greenwald, A.G. 2002. Math = male, me = female, therefore math $\neq$ me, *Journal of Personality and Social Psychology*, 83(1):44–59.

[38] OECD. 2020. Girls' and boys' performance in PISA. In *PISA 2018 results, volume 2: Where all students can succeed*, Paris: OECD Publishing.

[39] Östling, R., Wang, J.T., Chou, E.Y., and Camerer, C.F. 2011. Testing game theory in the field: Swedish LUPI lottery games, *American Economic Journal: Microeconomics*, 3(3):1-33.

[40] Rydval, O., Ortmann, A., and Ostatnicky, M. 2009. Three very simple games and what it takes to solve them, *Journal of Economic Behavior an Organization*, 72(1): 589-601.

[41] Schiffer, B., Pawliczek, C., Muller, B.W., Gizewski, E.R., and Walter, H. 2013. Why don't men understand women? Altered neural networks for reading the language of male and female eyes, *PLoS One*, 8(4):e60278.

[42] Selten, R. 1978. The chain store paradox, *Theory and Decision,* 9(2): 127–159.

[43] 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.

[44] Taylor, C.L., and Barbot, B. 2021. Gender differences in creativity: Examining the greater male variability hypothesis in different domains and tasks, *Personality and Individual Differences*, 174:110661.

[45] Thöni, C., Volk, S., and Cortina, J.M. 2020. Greater male variability in cooperation: Meta-analytic evidence for an evolutionary perspective, *Psychological Science*, 32(1):50-63.