Gazes and numbers: Two experiments in strategic sophistication and gender bias

Maria Cubel and Santiago Sanchez-Pages

No. 78/21

BATH ECONOMICS RESEARCH PAPERS

Department of Economics

Department of Economics | UNIVERSITY OF BATH

Gazes and numbers: Two experiments in strategic sophistication and gender bias*

María Cubel† Santiago Sanchez-Pages‡

Abstract

We investigate whether gender differences in strategic behavior depend on the perceived gender bias of strategic interactions. We use 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 find that males display significantly higher strategic sophistication than females in the first game but not in the second one, which is perceived to be female biased. However, females are underrepresented among top performers in both games.

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

JEL codes: C72; C91; D91; J16.

*We thank Ayala Arad, Jordi Brandts, Christoph Bühren, Syngjoo Choi, Natassa Papadopoulou, Rosemarie Nagel 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 persistent 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; 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 this paper we explore whether the effect of perceived gender biases on tournament performance translates to 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 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). Chou (2018) 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 quite mixed. Camerer, Ho and Chong (2004), Ostling, Wang, Chou and Camerer (2011) and Arad and Rubinstein (2012) find slight differences in fa-

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

vor of men in the beauty contest, the LUPI game and the 11-20 game respectively. 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 however (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 behaviour and thus the relative performance 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, which then results 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 that men are significantly more likely to pick the weakly dominant strategy, but also that they are more likely to pick very high numbers. As a result, 1) the variance of expected payoffs for males (but not the mean) is significantly higher than for females and 2) there are fewer women among the top performers. This is consistent with the patterns in math PISA scores in OECD countries (Machin and Pekkarinen, 2008; OECD, 2020) and in American Math Competitions (Ellison and Swanson, 2010).

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 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 the top performers.

We explore the reasons behind these results by analyzing the responses to a post-experiment questionnaire. In the beauty contest, we observe that females display higher strategic awareness than males, i.e. they are more likely to recognize the situation as a game, but they seem to understand worse how combinations of strategies translate into outcomes. In both games we find that participants who believe that the other gender is better at the game display lower strategic sophistication. Taken together, this results would seem to corroborate that stereotypes and perceived gender biases influence strategic behavior 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

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

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 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 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

We 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,

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

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. Main statistics of Study 1 by gender.

The second main result in Table 1 is that the proportion of participants who chose the weakly dominant strategy differs by gender (Proportion test $p = 0.004$).[4] This would suggest that males display higher strategic sophistication than females in this game, where strategies are numbers and winning requires a mathematical operation. This conclusion is reinforced if we compare the cumulative distributions of responses by gender. Figure 1 below shows that the fraction of participants who chose very small numbers is larger for men than for women, each depicted with their stereotypical color. The Davidson and Duclos (2000) test of stochastic dominance corroborates this:[5] The test returns that the distribution of responses chosen by females first-order stochastically

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

[5] This test compares distributions at several points. A distribution is said to first stochastically dominate another 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.

dominates the distribution of responses by males (Table A1 in Appendix B). Dominance comes from responses between 0 and 9.

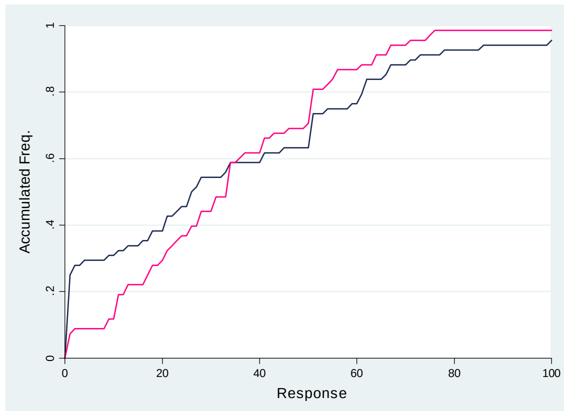

Figure 1: Cumulative distribution of responses by gender.

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

Table 1 also shows that there were no significant gender differences in average responses, median responses (despite the large gap) or in expected payoffs. The reason is that, as Figure 1 already suggests, more males than females selected very large numbers. This difference is not large enough to be picked up by the stochastic dominance test, but it is for expected payoffs as we show next.

The left panel in Figure 2 below depicts the Kernel density of expected payoffs by gender. It shows that the distribution for males has a larger variance than the one for females. The male-female variance ratio in expected payoff is 1.60, which is significantly different from one (Variance-comparison test, $p = 0.027$). This difference in variances is strikingly similar to the pattern observed in PISA scores for math and reading in most OECD countries, where boys display more extreme performances than girls (Machin and Pekkarinen, 2008). This pattern is also present in the most recent PISA 2018 (OECD, 2020).

The right panel in Figure 2 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).

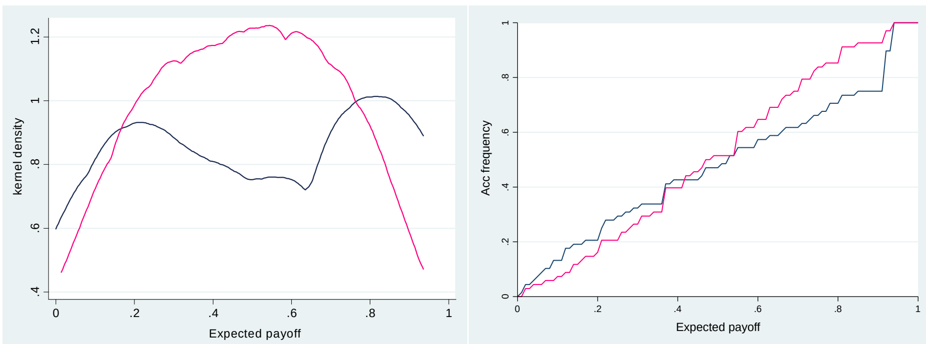

Figure 2: Density and cumulative distributions of expected payoffs by gender.

These results suggest that fewer women than men attain high expected payoffs in the two-player beauty contest. To investigate this gender difference in top performances 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 formed 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%.

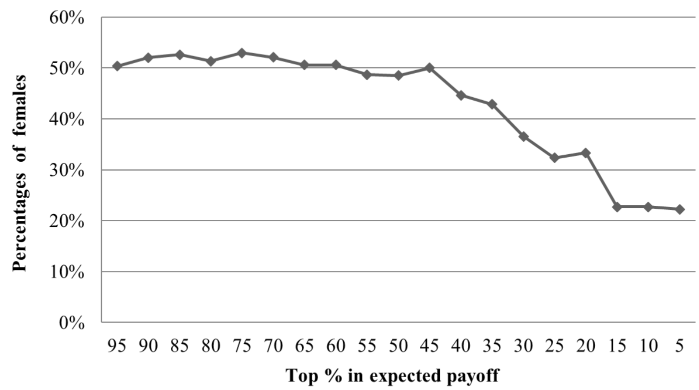

Figure 3: Percentage of females by top percentage of expected payoff.

Result 2 In the two-person beauty contest, the distribution of expected payoffs for males has a higher variance and first order stochastically dominates the distribution for females. As a result, women are significantly underrepresented among top performers.

This result is again eerily similar to the pattern of scores observed in math exams and competitions. Ellison and Swanson (2010) show male 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. The underrepresentation of girls among best scorers in the math component of PISA is also a common and stable feature for most OECD countries (Breda, Jouini and Napp, 2018).

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. Beliefs were not elicited in an incentive-compatible manner since the behavior of others plays no role and participants should not have acted upon their beliefs. However, these beliefs can be important if the level of game form recognition depends on the perceived relative strategic ability of the own gender.[6]

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 (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.

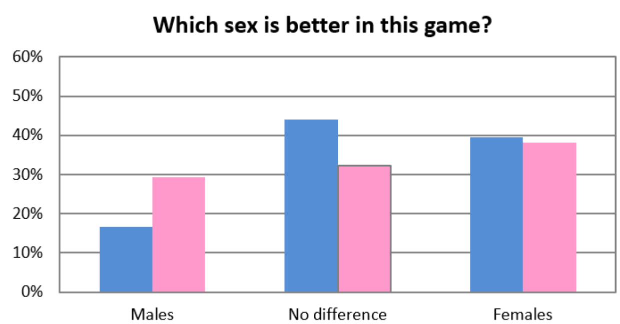

Figure 4: Responses by gender to questionnaire in Study 1.

Beliefs about which sex is better at the game correlate with strategic sophistication, as Table 2 below shows. Participants who thought the other sex

[6] We cannot rule out that participants used their responses in the questionnaire to rationalize their choices despite the absence of any interim feedback. For that reason, the relationships between beliefs and actions we dicuss in this subsection can only be correlations.

[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.

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%

Table 2. Choice of zero by gender and belief.

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.

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?"

Following Chou et al. (2009), we coded the responses to this question 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 ordered from a poorer to a better understanding of the game as illustrated in the selection of responses contained in Table 3 below.

Lack of attention
I chose randomly.
I chose the age I want to have a child.
I chose an important date.

Failure to recognize the situation as a game
The average is likely to be an even number.
The average should be between 30 and 50.
The average should end in 0 or 5.

Failure to understand how the winner is determined
If we both choose 0, the average will tend to infinity.
I did not understand the 2/3 thing.
I solved $\frac{x+y}{2} = \frac{2}{3}(x + y)$.
2/3 of the average is greater than the mean.

Used beliefs about the opponent
It is a race to the bottom but maybe not everybody noticed.
People tend to choose between 0 and 30.

Table 3. Illustration of failures in game form recognition: responses to the post-experiment questionnaire.

Although these categories are constructed from self-reports,[8] participants in these groups behaved differently. Figure 6 below shows that those who failed to select the dominant strategy because they used beliefs about the sophistication of the opponent picked significantly lower numbers and had significantly higher expected payoffs in average than participants who failed to pick zero for other reasons. This is corroborated by a comparison of their median response and expected payoff, which are significantly different from those of the rest of participants who selected weakly dominated strategies

[8] To minimise classification errors, answers were coded independently by the two authors. Disagreements were then discussed jointly. Those which could not be resolved were left unclassified together with too brief and incomplete answers (5 out of the 117 participants who did not pick zero).

(Median test, $p < 0.001$ and $p < 0.001$). This would indicate that participants who reported to have entertained beliefs about their opponent understood the game better than those who committed more basic errors.

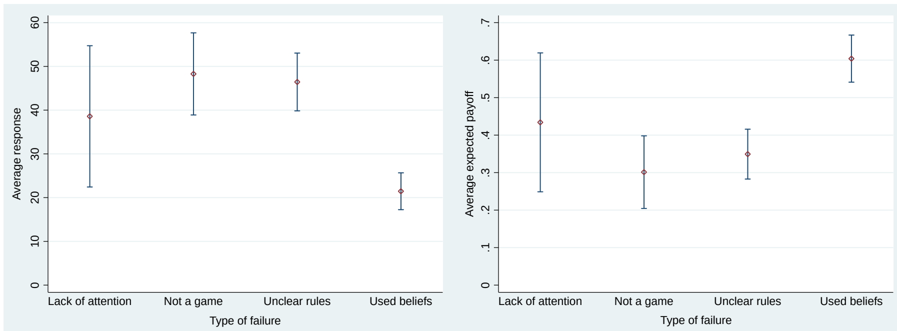

Figure 5: Average response and expected payoff by type of failure.

Let us now study gender differences in game-form recognition. Figure 6 contains the frequencies of type of failure by gender. The distributions of answers are marginally different (Chi-square test, $p = 0.072$). The most common type of failure in the sample as a whole is a lack of understanding of how combinations of strategies produce outcomes (36.7%). Females fall more frequently into this error than men. On the other hand, men are more likely than women to commit basic mistakes due to inattention and strategic unawareness. This is consistent with Result 2 above, showing that men are more likely to pick both the weakly dominant strategy and very high numbers.

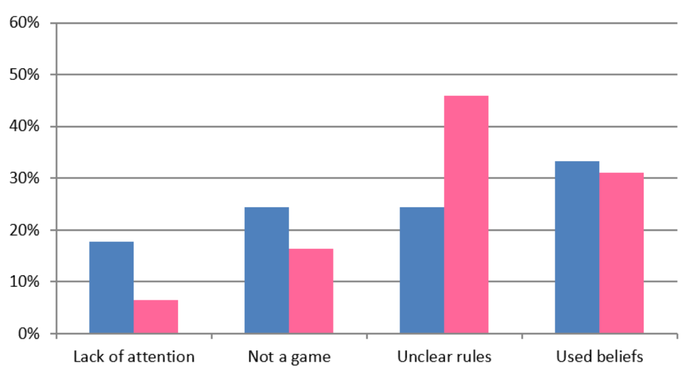

Figure 6: Failures in game form recognition by gender.

Result 3 Women show higher strategic awareness than men in the two-person beauty contest. However, they struggle more to understand how choices produce outcomes.

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 7 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.

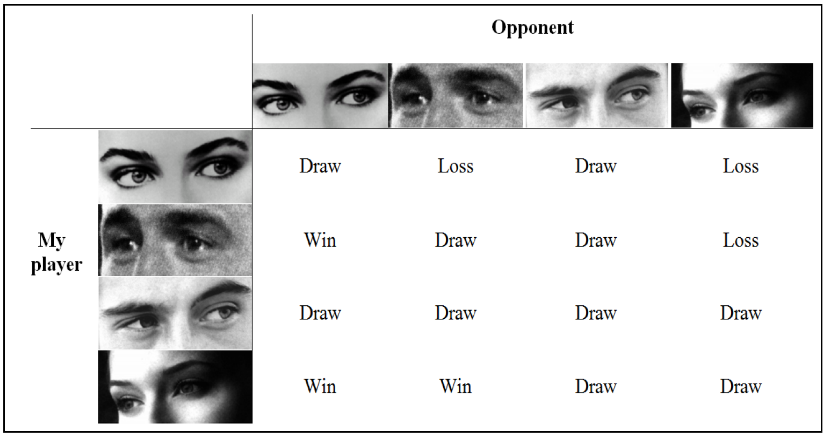

Figure 7. Outcome matrix for the gaze coach game.

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 Eyes test.[9] 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.

[9]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".

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 7 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 potentally 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 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

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 is based on the pair of players participants thought they were selecting as revealed by their answers in the identification task. We call the latter the players' "true" choices and their "true" expected payoff. The difference between both versions of the variables is that the one based on recorded chocies includes the error in emotion recognition, i.e. the error in matching words to photographs, whereas the "true" version contains only the error in game form recognition. 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 can 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 pick the weakly dominant strategy -one out of five participants- is very similar to the one we observed in Study 1. The proportion doubles but remains 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.[10]

	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.

The second result emerging from Table 4 is that there are no significant gender differences in the proportion of participants who selected the 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 4 There are no significant gender differences in aggregate behavior in the gaze coach game.

Taken together, Result 1 and 4 are consistent with (but not conclusive proof of) the idea that the perceived gender bias of competitive games has an impact on the relative strategic sophistication of men and women. We

[10] 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.

observed gender differences in favor of men in the two-person beauty contest where strategies are numbers and maths are important but no differences in the gaze coach game, which entails recognizing emotions in others correctly.

However, a more detailed look at the results in the gaze coach game complicates this picture. Figure 8 below depicts the cumulative distribution of expected and "true" expected payoffs by gender. Although they are not different in the former case, a familiar pattern emerges in the latter: The distribution of "true" expected payoffs for males first order stochastically dominates the one for females. This dominance, as in the two-person beauty contest, takes place for the highest expected payoffs, as the Davidson-Duclos test confirms (see Table A3 in Appendix B). This suggests that women are underrepresented among top performers in the gaze coach game too.

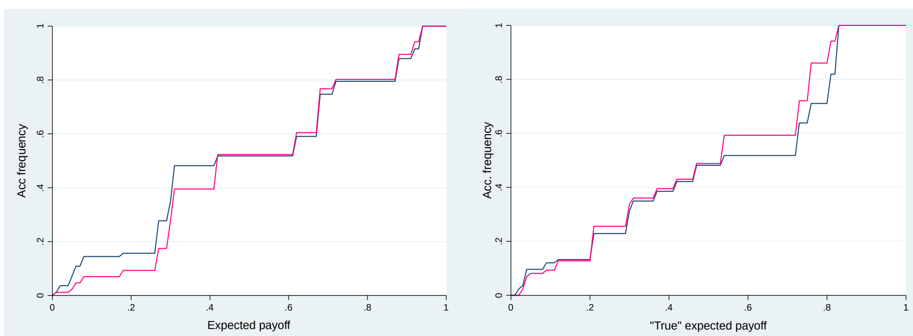

Figure 8. Cumulative distribution of expected and "true" expected payoffs by gender.

To corroborate this finding, we plot the percentage of females by top percentage of expected payoff in Figure 9. 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%.[11] However, for the "true" expected payoff, that is, once we remove the errors in

[11] 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.

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.

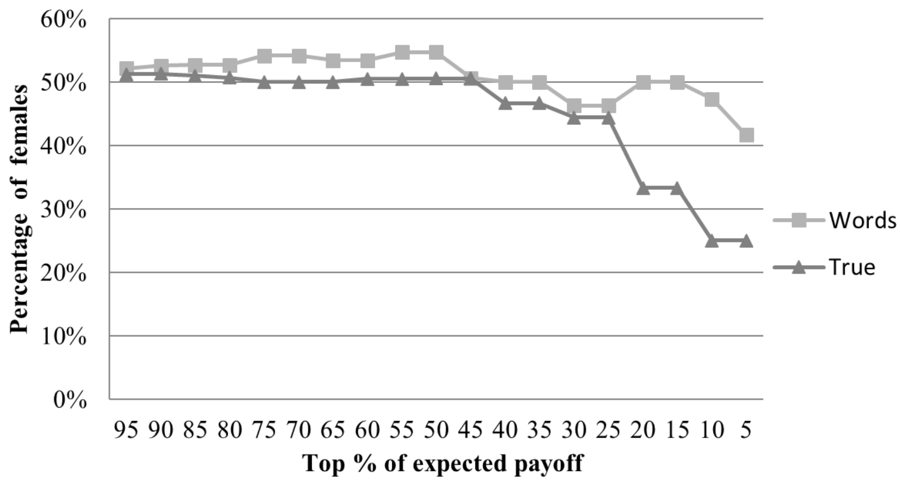

Figure 9. Percentage of females by top percentage of expected payoff.

Result 5 Women are also 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 pattern we observed in the two-player beauty contest 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-biased. We investigate this next.

3.3 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. This strong perceived female-bias in the identification task contrasts with performance in the task, where, as mentioned earlier, we did not find any significant gender difference.

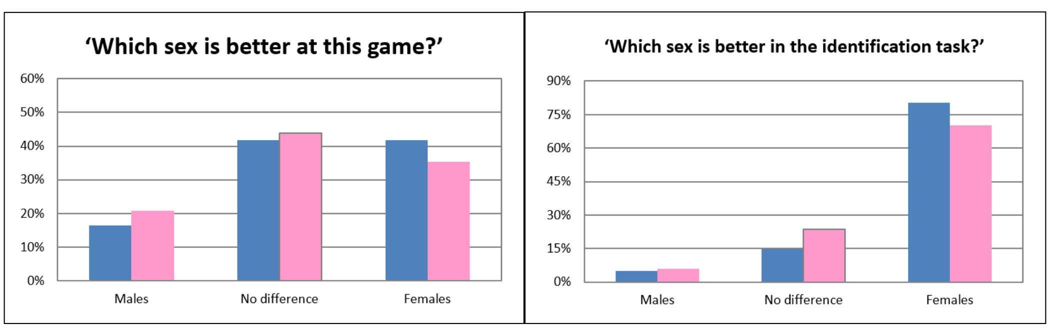

Figure 10. Responses by gender to questionnaire in Study 2.

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 belief.

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. Figure 11 shows the corresponding cumulative distributions by belief. 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 Table A4 in Appendix B). This would suggest instead that the perceived gender-bias affects game form recognition too.

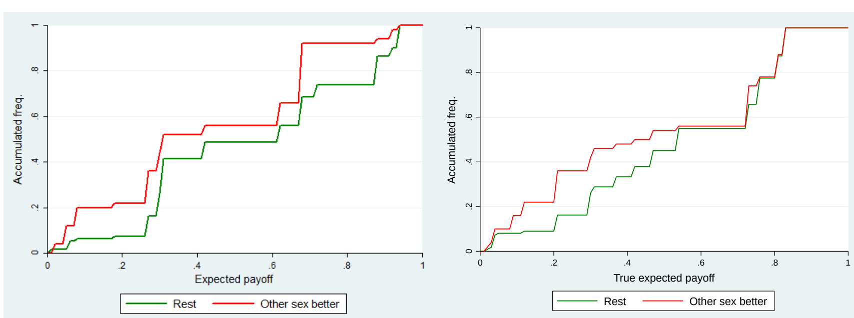

Figure 11. Cumulative distributions of expected payoffs by belief.

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 6 In both the two-player beauty contest and the gaze coach game, participants who believe the other sex is better in the game display lower strategic sophistication.

3.4 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 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. We coded these answers following the same procedure as in Study 1. 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, no participant explicitely reported to have seen the gaze coach game as a game of chance. Second, a new type of error emerges. A small but non-negligible fraction of participants 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. We classified this type of failure as lack of attention since these participants ignored completely the information provided in the outcome matrix.[12]

As in Study 1, the different types of failure in game form recognition correlate with participants' expected payoffs. Figure 12 below shows that the average "true" expected payoff is lower for participants who failed to select the weakly dominant strategy because of lack of attention. The median "true"

[12] 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".

expected payoff for the three types of failure are indeed statistically different (Median test, $p = 0.017$).

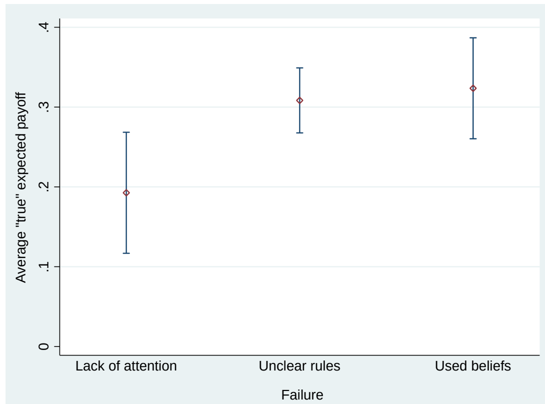

Figure 12: Average "true" expected payoff by type of failure.

Figure 12 depicts the distribution of types of failures by gender. As in Study 1, the most common type of failure for both men and women (63.8%) 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. Also in contrast with that game, we find no trace of significant gender differences in types of failure (Chi-square, $p = 0.609$).

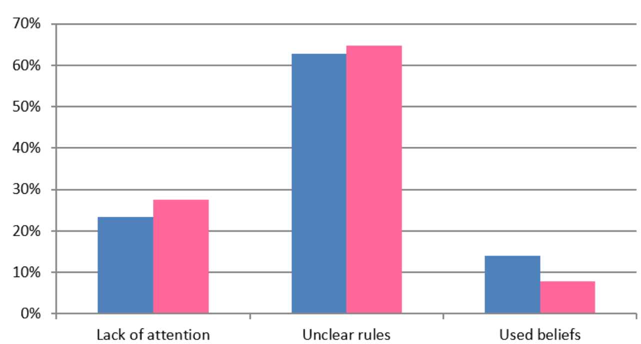

Figure 13: Failures in game recognition by gender.

Result 7 There are no significant gender differences in failures in game form recognition in the gaze coach game.

4 Conclusion

Our results show that the perceived gender bias of strategic interactions affects the behavior of men and women. Men displayed higher strategic sophistication 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 individuals who expected the other gender to be better at the game displayed lower levels of strategic thinking. Taken together, these findings suggest that women are less likely to display lower strategic sophistication than men when they perceive strategic interactions to be female-friendly. From a policy perspective, this would imply that as long as key strategic interactions in labour markets and in life are set and framed by men, it is natural to expect gender differences in outcomes.

Although the two games that we have studied are admittedly not isomorphic, our results indicate that the underperformance of women observed in tournaments (e.g. Gneezy et al., 2003; Gneezy and Rustichini, 2004) translates into a pervasive underrepresentation of women among top performers in competitive games. The nature of strategies and of the processes required to win seem to moderate the emergence of gender differences in strategic sophistication. But these factors do not seem to be powerful enough to fully eliminate or reverse these differences; women remain unsettlingly underrepresented among top performers even when the game entails 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 behaviour in non-competitive games such as coordination games. We leave this for future research.

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 real tournaments. 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] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] 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.

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

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

[17] 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.

[18] 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.

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

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

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

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

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

[24] 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.

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

[26] 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.

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

[28] 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.

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

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

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

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

[33] 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.

[34] Schiffer, B., Pawliczek, C., Muller, B.W., Gizewski, E.R., 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.

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

[36] 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.

A Appendix: Experimental instructions

Below we produce the translation of the experimental instructions originally written in Spanish. These instructions were provided in paper documents which were also read aloud by one of the experimenters (randomly chosen each session). We present first the general instructions with the variation for Study 2 in parentheses, then the instructions specific to each of the two studies and finally the post-experiment questionnaire with variations for Study 2 in parentheses.

GENERAL INSTRUCTIONS

Hello. Many thanks for taking part in this experiment.

The purpose of this session is to study how people make decisions in strategic settings. All data obtained in this session will be anonymous and will be used exclusively for scientific purposes.

After this introduction, we will describe you an scenario. You will have to take a decision in that scenario and write it down in the answer sheet we will provide you.

(Variation for Study 2: After this introduction, we will proceed in two independent phases. You will have to take a decision in each phase and write it down in the answer sheets we will provide you.)

In this scenario, you will be anonymously matched with another participant. In addition to the 5 euros show-up fee you will be able to obtain an additional 5 euros depending on the decisions you and the person you will be matched with take.

After collecting the answer sheets, we will give you some time to answer a short questionnaire. Once we collect the questionnaires, we will start calling you to the desk by your ID number. The amount you have earnt will then be paid to you in cash.

You must observe two rules:
1. Switch off your mobile phone.
2. Don't talk during the session.

If you do not follow these rules, you will be asked to leave the session with no compensation.

Finally, we ask you to:
1. Fill your ID number in your answer sheet and the questionnaire. If you don't do it, we won't be able to pay you.
2. Read carefully the instructions on the answer sheet. If you don't do it, the data collected will be useless.

STUDY 1

Read carefully the following instructions:

You have been randomly matched with another participant in this session.

You will not know who you will be matched with.

You and that person will play the following game:

Each one of you should choose a number between 0 and 100 (both included) with the objective of guessing 2/3 (two thirds) of the average of your two choices.

The one of you whose answer is closer to 2/3 of the average of the numbers you have chosen will receive 5 euros in addition to the show up fee. If you both pick the same number, that sum will be split between the two of you.

Take the time you need to answer. Once you do, write your choice in the answer sheet and wait in silence until the rest of participants have finished.

Which number do you choose? (remember, between 0 and 100, both included).

STUDY 2

Phase 1

Read carefully the following instructions:

You have been randomly matched with another participant in this session.

You will not know who you will be matched with.

You and that person will play the following game:

You are the coach of a team with 4 players. You and the person you have been matched with should select 2 of your players to participate in a tournament. In that tournament, each of your players will play with each of the players of the other team. Hence, a tournament has four matches. You and the other coach will select your two players without knowing the players chosen by the other coach.

Each of your four players belongs to a type. The coach you have been matched with has the same type of players available as you. Each type of player wins against some type of opponents and loses against other types. If two players of the same type play each other, they draw. The following table describes the matches a type of player wins, draws and loses depending on the type of the opponent player.

[Figure 7 here]

For each match one of your player wins you will receive 1 point, 0.5 if they draw and 0 if they lose. The coach winning the tournament is the one receiving most points. The winner will receive 5 euros in addition to the show up fee. In case of a draw, that sum will be split between the two of you.

Next you will find the 4 words which describe the 4 types of players and whose photographs appear in the table above. To select the 2 players that you want to participate in the tournament tick the box corresponding to the word that you think best describes the players you have selected (there is only one right word per photograph).

- Uneasy
- Cautious
- Anticipating
- Contemplative

Now, wait in silence until the rest of participants have finished.

Phase 2

Consider the following words:

- Uneasy
- Anticipating
- Cautious
- Contemplative

Next, we will present you 4 photographs. You must associate each photograph with the word that you think best describes what the person in the photograph is feeling or thinking. Just write the initial of the word (I, E, C and R in Spanish). It may seem that more than a word is applicable to one photograph, but there is only one correct answer. Before answering, make sure you have read the four words and examined the four photographs carefully. If you do not know the meaning of any of the words, raise your hand and we will clarify that to you in private.

For each correct association between words and photographs you will receive 20 euro cents.

POST-EXPERIMENT QUESTIONNAIRE

Please, answer the following questions in as much detail as you can.

1. How did you pick your number (the 2 players that you sent to the tournament)?
2. Who do you think is better in this game, women or men? Why?
3. In general, who do you think obtains better results in strategic situations, women or men? Why?
4. In average, who do you think picked higher numbers (obtained better results in the phase where you had to assign words to photographs), women or men?

B Appendix: Additional tables

Point	*Response*
0	2.073**
1	2.165**
5	2.381***
9	2.055**
15	1.149
22	0.396
33	0.541
44	0.175
50	0.000
66	0.000
83	-0.171
Positive (negative) t-statistics indicate that the accumulated frequency of female (male) responses is lower than the one for males (females).	

]

Table A1: *Female-male DD Test t-statistics for responses in Study 1.*

Point	*Expected payoff*
0.000	1.007
0.048	0.730
0.152	0.934
0.201	0.614
0.250	0.984
0.282	0.754
0.333	0.366
0.388	0.348
0.459	-0.344
0.500	-0.343
0.548	-0.694
0.597	-0.881
0.652	-1.461
0.701	-2.088**
0.758	-2.240**
0.805	-1.961**
0.846	-2.878***
0.916	-1.743*
Positive (negative) t-statistics indicate that the accumulated frequency of female (male) payoffs is higher than the one for males (females) .	

Table A2: *Female-male DD Test t-statistics for expected payoff in Study 1.*

Point	*Expected payoff*	Point	*"True" expected payoff*
0.009	-0.025	0.013	-1.431
0.013	-1.042	0.028	-0.492
0.040	-1.497	0.036	-0.057
0.057	-1.510	0.039	-0.626
0.072	-1.579	0.045	-0.342
0.171	-1.254	0.081	-0.577
0.263	-1.606	0.118	-0.089
0.297	-0.986	0.200	0.089
0.309	-1.138	0.202	0.408
0.414	0.067	0.290	0.332
0.618	0.189	0.306	0.150
0.671	0.819	0.364	0.130
0.675	0.310	0.414	0.112
0.714	0.115	0.460	0.083
0.872	0.325	0.536	0.982
0.914	0.661	0.729	1.151
0.932	0.872	0.757	2.404**
		0.802	2.491**
Positive (negative) t-statistics indicate that the accumulated frequency of female (male) payoffs is higher than the one for males (females).			

Table A3: *Female-male DD Test t-statistics for expected payoffs in Study 2.*

Point	*Expected payoff*	Point	*"True" expected payoff*
0.009	-1.427	0.013	0.505
0.013	0.893	0.028	0.721
0.040	2.738***	0.036	0.149
0.057	1.604	0.039	0.569
0.072	2.799***	0.045	0.380
0.171	2.902***	0.081	1.361
0.263	3.166***	0.118	2.011**
0.297	2.634***	0.200	1.671*
0.309	1.480	0.202	2.590***
0.414	1.027	0.290	1.952*
0.618	1.459	0.306	2.079**
0.671	1.857*	0.364	1.753
0.675	4.446***	0.414	1.441
0.714	3.558***	0.460	1.055
0.872	1.817*	0.536	0.123
0.914	2.461***	0.729	1.074
0.932	2.518***	0.757	0.073
		0.802	0.109
Positive (negative) t-statistics indicate that the accumulated frequency of payoffs for those who believe the other sex is better is higher (lower) than for the rest.			

Table A4: *DD Test t-statistics for expected payoffs in Study 2 by belief in the other sex is better at the game.*

C Appendix: Additional tables

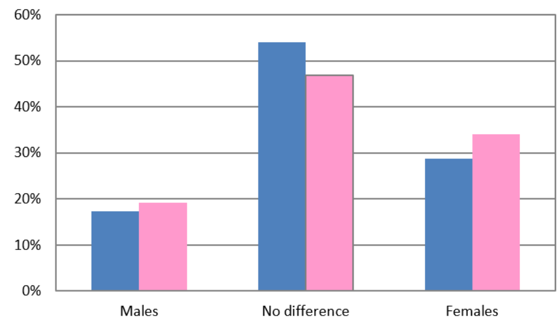

Figure A1: Responses in the out-experiment sample to "Which sex is better at this game?"