Gender differences and stereotypes in strategic reasoning[1]

María Cubel[2] Santiago Sanchez-Pages[3]

January 2017

[1]We thank comments and suggestions made by two anonymous referees and the editor in charge, Andrea Galeotti, and also by Larbi Alaoui, Ayala Arad, Colin Camerer, Nagore Iriberri, Tatiana Kornienko, John Morgan, Muriel Niederle, Lise Vesterlund and audiences at Alicante, Barcelona, Edinburgh, Pittsburgh, the University of New South Wales, the COSME-FEDEA workshop and the RES 2016 conference. We are also grateful to Efi Adamopoulou, Peter Backus, Dirk Foremny, Ana Nuevo-Chiquero and Amedeo Piolatto for their assistance during the experimentalsessions. We are particularly indebted to Ariel Rubinstein for retrieving part of the data featured here and for his constructive criticisms. All remaining errors are ours. Both authors acknowledge financial support from the Spanish Ministry for Science and Innovation research grant ECO2012-33243 and from the Generalitat de Catalunya grant 2009SGR1051.

[2]Brunel University, University of Barcelona and IEB. E-mail: cubel@ub.edu. URL: https://sites.google.com/site/mariacubel

[3]King's College London and University of Barcelona. E-mail: sanchez.pages@gmail.com. URL: http://www.homepages.ed.ac.uk/ssanchez/.

Abstract

Recent literature has emphasized that individuals display varying levels of strategic reasoning. This paper explores the existence and endogeneity of gender differences in strategic behaviour. We report results from two experimental studies employing the beauty contest game, one in the laboratory and one in the classroom. We observe robust and significant gender differences in observed depth of strategic reasoning in favour of men in zero-stake situations. These differences disappear when a monetary prize is awarded. We also find that depth of strategic reasoning varies with gender priming. Females display engage in more rounds of reasoning than males when gender and stereotypes are made salient. This effect of priming is driven by females who believe women are superior in the game.

Keywords: guessing game, depth of reasoning, gender, beliefs, stereotype threat.
JEL codes: C72, C91, D81, J16.

1 Introduction

The experimental literature has established the existence of strong individual heterogeneity in strategic sophistication. Observed individual behaviour departs drastically from the predictions derived under the assumption of publicly-known unbounded cognitive capabilities (Nagel, 1995; Stahl and Wilson, 1995; Ho et al., 1998; Costa-Gomes et al., 2001; Bosch-Domenech et al., 2002). This heterogeneity reflects differences in the extent to which individuals engage in mentalising processes or "theory of mind," that is, the activity of thinking about others' thoughts, emotions and intentions (Baron-Cohen, 1991). Models of level-k thinking have tried to account for these differences by acknowledging that individuals vary in their cognitive level and have non-equilibrium beliefs about the sophistication of others.[1]

However, these models remain silent on the extent to which the distribution of observed strategic sophistication represents an underlying distribution of ability. Observed depth of strategic reasoning is likely to be endogenous to beliefs about the sophistication of others and the incentives provided. An individual may be willing to engage in additional mentalising effort if he/she perceives the rest of players to be relatively sophisticated. On the other hand, a person may be reluctant to engage in further levels of reasoning if stakes are low. Behaviour in those cases would be a reflection of the motivation of individuals to engage in the reasoning process as much as of their cognitive ability.[2]

In this paper, we experimentally study the heterogeneity and endogeneity of strategic sophistication in the context of gender. We analyze the existence of gender differences in observed depth of reasoning, beliefs and in the sensitivity to financial incentives and gender priming in games. To the best of our knowledge, we are the first ones to do so.

The first question we address is whether there exist gender differences in depth of strategic reasoning. In the psychology literature, superior mentalising ability is typically ascribed to women (Baron-Cohen, 2002). But no study has explored whether this perceived superiority affects behaviour. The second question is whether such gender differences (if any) are mediated by

[1] Level-k models of thinking were introduced by Nagel (1995) and Stahl and Wilson (1994, 1995). Later, Camerer et al. (2004) proposed the cognitive hierarchy model. Both models are anchored on the existence of non-strategic individuals, labelled level-0, but differ on how individuals respond to the presence of less sophisticated ones. See Crawford et al. (2013) for a survey and Strzalecki (2014) for a recent development in this literature.

[2] See Choi (2012) and Alaoui and Penta (2016) for recent attempts to develop theoretical models accounting for the endogeneity of strategic sophistication.

beliefs about the relative strategic sophistication of men and women. Gender stereotypes affect daily behaviour in a pervasive manner. Perceptions about the gender-bias of tasks have been shown to have an impact on gender differences in performance (Guenther et al., 2010; Shurchkov; 2012). There is also evidence showing that stereotypes about men's superiority in maths produce gender differences in test performance when gender is primed (Danaher and Crandall, 2008; Nguyen and Ryan, 2008). Therefore, stereotypes might also influence strategic behaviour. We investigate whether gender salience and changes in the gender composition of the group of players alter their observed depth of reasoning. Our third question relates to the endogeneity of depth of reasoning to monetary incentives. Higher monetary incentives are likely to induce deeper strategic reasoning. But monetary stakes, or the absence of them, might also frame the interaction in a different light for men and women, intensify or crowd out intrinsic motivation, and create gender differences in strategic behaviour.

We explore these questions by means of a laboratory experiment where we manipulate financial incentives, gender priming and the gender composition of the set of participants. The design employs the p-beauty contest (Nagel, 1995). This game is well suited for our purposes for three reasons. First, incentives are easy to adjust by changing its monetary value. Second, beliefs about the relative sophistication of opponents are extremely important. The guessing game involves a calculation task. Gender priming may trigger negative stereotypes about women's mathematical skills. Because it is a strategic interaction, it may also activate beliefs about females' advantage in mentalising. Finally, it is a competitive game; players must outguess others in order to win a prize. The literature (see below) has shown that the perceived gender-bias of a task affects the performance of men and women in competitions; gender priming can activate such perceptions.

Participants first play several non-incentivised rounds of the guessing game with different values of p and no feedback. One of them is the standard $\frac{2}{3}$-beauty contest. In the control treatment, subjects then play an incentivised round of the $\frac{2}{3}$-beauty contest. In the *Priming* treatment, gender and stereotypes about the relative ability of males and females in the game are made salient. So after the non-incentivised phase, subjects play two incentivised rounds of the guessing game, one in a mixed gender group and another in a same gender group. The comparison between the incentivised and the non-incentivised rounds of the control treatment allow us to study the existence of gender differences in depth of strategic reasoning. The comparison between the primed and the gender-neutral incentivised rounds, and between mixed gender and same gender rounds allows us to

study whether stereotypes about the relative strategic sophistication of men and women affect behaviour.

We find that females display less depth of reasoning than males when monetary incentives are absent. No gender differences exist when a monetary prize is at stake in the control treatment. Gender differences re-emerge but in the opposite direction in the *Priming* treatment. Females perform more rounds of reasoning than males. Changes in gender composition only affect a subset of males who play according to more rounds of reasoning when playing in mixed gender groups compared to single gender groups.

We explore the reasons behind the effect of priming and changes in gender composition by analyzing the responses to a questionnaire administered to participants and non-participants of similar characteristics. We find that beliefs about which gender is better in the game have a significant effect on behaviour. Males who believe females are superior display less depth of reasoning than those who believe the opposite. This difference is in line with the concept of stereotype threat (Steele, 1997) by which members of negatively stereotyped groups perform worse in fear of confirming the stereotype. On the other hand, our gender manipulation has the effect of inducing women who believe females are better in the game to perform further rounds of strategic reasoning. We conclude that this positive effect of gender priming on the observed depth of reasoning of womenis due to females perceiving themselves as superior in the game.

This study does not allow us however to analyze the effect of the introduction of financial incentives on the observed depth of reasoning because its within-subject design can induce feedback-free learning (Weber, 2003). This is important because monetary incentives can constitute a contextual cue and frame the interaction in a different light. We report results from a second experiment run on first-responses. Students in different cohorts played the beauty contest with and without a monetary prize. We find that female subjects play according to less rounds of reasoning than males under zero-stakes. No gender differences exist when a monetary prize is awarded. These results coincide with those we obtained in the laboratory study. In addition, we observe a significant increase in the rounds of strategic reasoning of females when monetary incentives are present. Males do not react significantly to the presence of a monetary reward.

The rests of the paper proceeds as follows: Section 2 reviews the related literature. Sections 3 presents the design and results of our laboratory experiment. We analyze the responses to questionnaires and the results of the classroom experiment in Section 4. In Section 5, we conclude and discuss further the relevance of our results.

2 Related literature

Very few experimental studies report gender differences in strategic behaviour. Camerer et al. (2004) show results from a beauty contest game in single sex groups in their Table 2, but they only report summary statistics. Burnham et al. (2009) find no gender differences in choices in the beauty contest. This is consistent with our results when gender is not primed and there are monetary incentives. Östling et al. (2011) and Arad and Rubinstein (2012)[3] report that females display slightly lower strategic sophistication in the Lowest Unique Positive Integer (LUPI) game and in the Colonel Blotto games respectively. The main goal of these studies was not to investigate the existence of gender differences in strategic sophistication.

Several studies have explored the existence of other types of individual differences in the beauty contest. Burnham et al. (2009) and Gill and Prowse (2015) show that there is a significant correlation between higher cognitive ability and lower entries. behaviour in the beauty contest is similar across subject pools, although some differences exist; portfolio managers and game theorists display higher strategic sophistication (Bosch-Domenech, et al., 2002; Camerer et al., 2004). Kovalchik et al. (2005) find that older adults play similarly to young adults and Bühren and Bjorn (2010) find that chess grandmasters do not play differently than lay people.

A number of studies have found that strategic behaviour responds strongly to the perceived sophistication of opponents. Palacios-Huerta and Volij (2009) find that when students play the centipede game against professional chess players they engage in more rounds of backward induction.[4] Agranov et al. (2012) find that undergraduate students seem to perform further rounds of reasoning when playing the guessing game against graduate students than against computers. Georganas et al. (2015) find a similar result in the undercutting game. Our results confirm these findings.

We are aware of only one experimental study relating strategic sophistication to financial incentives. Alaoui and Penta (2016) find that subjects engage in more rounds of reasoning when the prize from outguessing the opponent increases.[5] However, these authors do not explore gender differences in the response to higher stakes. Fryer et al. (2008) find that

[3] Personal communication with the authors.

[4] This is not contemplated by models of level-k thinking since agents in these models do not factor the presence of individuals more sophisticated than them.

[5] Arad and Rubinstein (2012) run a treatment where they manipulate payoffs so that further levels of reasoning have no monetary cost. They find that nevertheless subjects very rarely perform more than three rounds of reasoning.

the performance of males in a GRE-style mathematical test increases relative to the performance of females when a payment per correct answer is introduced. This is in contrast with our findings, but might be due to the strategic nature of our experiment. Azmat et al. (2016) find that the gender performance-gap in lower-stake tests in favor of female students diminishes as test-stakes increase. In line with our findings, Frick (2011) employs data from professional distance running competitions and finds that differences in the competitiveness of female and male races are significantly smaller in races where higher prizes or more prestige is at stake. Similarly, Petrie and Segal (2015) observe that the gender gap in tournament entry vanishes when prizes become sufficiently large.

By using a competitive game, in which the player who best guesses the average response wins, our paper also relates to the literature on gender differences in competitive performance. Gneezy et al. (2003) and Gneezy and Rustichini (2004) have shown that females underperform in competitive environments. Guenther et al. (2010) find that competitive performance depends on the perceived bias of the task; females perform better than males when the task is perceived as female-biased. Along similar lines, Shurchkov (2012) find that females overtake men in competitions involving a verbal task and low-time pressure. Regarding the effect of gender priming, Iriberri and Rey-Biel (2016) show that omitting information about the gender of the opponent helps to mitigate the underperformance of women in competition. In contrast, we find that gender priming induces females to engage in further rounds of reasoning and to outperform males.

3 Beauty in the lab

3.1 Design

Our experimental study was conducted with four different cohorts of undergraduate students at the University of Barcelona between 2012 and 2015. We made no mention to gender during the recruitment process.[6] A total of 240 subjects participated in the study. This sample was quite homogeneous. Virtually all subjects were Spanish and all of them majored in the School of Economics and Business. They were first year students in order to ensure they had no previous knowledge of game theory. Because of this limitation, we could only run at most two sessions per academic year.

[6] For showing up, subjects received three euros

.

All participants played in gender-balanced groups of 24 subjects. Subjects could see each other but were seated at a considerable distance so they could not communicate. This was intended because we wanted subjects to see the gender composition of the group. The experiment was implemented with pen and paper. No feedback was provided during the session, only at the end. Experimenters answered privately any questions subjects had. The sessions lasted between 40 and 50 minutes. There were always two instructors in each room. Their gender matched the gender composition of participants in the room in order to minimize experimenter demand effects. At the end of each session, participants filled up a short questionnaire aimed to elicit their views about the behaviour and the relative strategic sophistication of males and females in the game and their. We analyze the responses to this questionnaire in Section 4.1.

Each session was divided in two phases. The first phase was common to all sessions and treatments. In this phase, there were no monetary incentives and gender was never referred to or made salient. Subjects were asked to guess a fraction p of the average response in their room. They played nine rounds of this guessing game with different values of p in each round. The values were $p = \left(1, \frac{2}{3}, \frac{11}{10}, \frac{1}{3}, \frac{3}{2}, \frac{1}{5}, \frac{6}{5}, \frac{1}{2}, \frac{4}{3}\right)$. Instructions were provided through white paper booklets where participants also had to record their answers. The experimenters read the instructions aloud to facilitate comprehension. Subjects did not write neither their name nor their gender in these booklets. Each participant was assigned a number that served as their unique identifier. The purpose of this first phase was to help subjects to familiarize with the beauty contest.

In the second phase, we introduced financial incentives and administered two treatments, the *Priming* treatment (n=144) and the *No priming* treatment (n=96). In the *No priming* treatment, participants had to guess two-thirds of the average response in their room. The winner got a prize of 40 euros (around £32); the prize was divided if there was more than one winner. Participants had to provide their answer in a white paper sheet.

In the *Priming* treatment, subjects played two independent rounds where they had to guess two-thirds of the average response in their room. The difference between the two rounds was the gender composition of the group, single gender (SG) or mixed (balanced) gender (MG). These sessions were run in two rooms located in two different corridors. Hence, at the beginning of the second phase, there were two gender-balanced groups of 24 participants in each room. We then simultaneously moved either all the male or female students in each room from one room to the other using different corridors so the two groups could not see each other. We combined these

movements of participants in such a way that different sessions alternated the order of the SG and MG rounds. When moving from one room to the other, participants were guided by an instructor of their same gender who made all efforts to prevent any communication among them. All participants changed room at some point of the session. To help the reader, we provide a graphical representation of these movements in Figure A1 of Appendix A.

In addition to this manipulation, we primed gender by distributing pink booklets to female subjects and blue booklets to male subjects. In the MG round, there was one female and one male instructor in each room. The gender of the two instructors in the room coincided with the gender of the group in the SG round. Payoffs in the *Priming* treatment were determined by selecting randomly one of the two rounds of the second phase. Consequently, there were two prizes of 40 euros each, one per room.

Throughout the paper, we will associate lower entries when $p < 1$ to further rounds of reasoning, but we will be careful not to associate lower responses with more sophisticated strategic reasoning. The latter association is supported by some studies: Subjects who choose lower entries have better scores in the Cognitive Reflection Test (Burnham et al., 2009; Brañas-Garza et al., 2012) and display more activation in areas of the brain associated with theory of mind (Coricelli and Nagel, 2009). But it would be ultimately wrong to infer higher strategic sophistication from lower responses directly. This is because strategic sophistication is a function of both depth of reasoning and beliefs about the sophistication of the opponents. It would be incorrect to label as sophisticated a player who chooses the Nash equilibrium strategy when the rest of players are relatively unsophisticated compared to another player who departs from the game theoretical prediction but takes correctly into account the strategic sophistication of the rest of players. To better explore this association, we use the quadratic distance to the winning response as a measure of strategic sophistication in Appendix B, and show that our main results go through.

3.2 Hypotheses

As outlined in the introduction, our experimental design aims to study the heterogeneity and endogeneity of strategic reasoning by changing financial incentives and priming stereotypes about the strategic sophistication of men and women. Since we are the first ones to address these issues directly, we cannot draw strong hypotheses from previous evidence.

The results in Burnham et al. (2009) suggest that we should expect no gender differences in behaviour in the incentivised gender-neutral round.

Regarding gender priming, the recent theoretical model by Alaoui and Penta (2016) and the experimental evidence in Agranov et al. (2012) suggest that players should play according to more (less) rounds of reasoning when they perceive they are playing against more (less) sophisticated opponents. Hence, we expect the effect of making gender salient and of changing the gender composition of the group to depend on player's beliefs about the relative ability of men and women in the game. On the one hand, the beauty contest involves a relatively complex calculation task: subjects must think what might be the average response, and then multiply the result by the announced factor one or more times. This calculation may trigger gender stereotypes related to the mathematical abilities of females.[7,8] On the other hand, the game could be perceived as female-biased because it is a strategic interaction and there is a well-known stereotype of women being better at knowing what others feel and think. Admittedly, because these two effects may be present, our design does not allow us to isolate one from the other.

For the analysis by round we will employ the Mann-Whitney and the Median tests because we will be comparing responses by gender. When comparing treatments, we will employ the Wilcoxon signed-ranked and the Sign tests because observations are not independent. We must use non parametric tests because responses are not normally distributed in any of the treatments or rounds.

3.3 Results within rounds

3.3.1 First phase: no monetary incentives, no gender priming

Table 1 depicts the aggregate results for the round of the first phase with $p = \frac{2}{3}$. Recall that there were no monetary incentives and no gender priming in this phase. The distribution of male responses has a lower mean and median than the distribution of female entries.

	Mean	Median	Std dev
Males	28.9	25	17.7
Females	32.1	30	20.5

Table 1: Aggregate results by gender in the No monetary incentives round.

[7] Krendl et al. (2008) show that brain areas involved in calculation are less active in females when this stereotype is activated. Some of these areas are also relevant for subjects playing the beauty contest (Coricelli and Nagel, 2009).

[8] The existence of gender differences in math performance is still a much debated issue. To have an effect on behavior, subjects only need to believe such stereotype to be true.

In addition, the distribution of female responses first order stochastically dominates the distribution of male responses. To test this, we employ the test of stochastic dominance introduced by Davidson and Duclos (2000).[9] This test allows us to associate stochastic dominance to a particular range of entries and, hence, to a certain depth of reasoning. The Davidson-Duclos test reports that gender difference emerge in the interval between 42 and 50 (see first column of Table A1 in Appendix A), which corresponds to less rounds of reasoning. Figure 1 illustrates this result. We use blue and pink for males and females respectively.

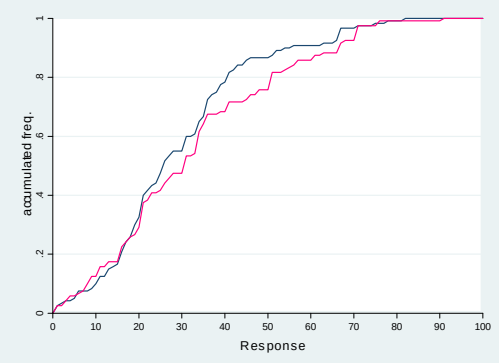


Figure 1: Cumulative distributions by gender in the first phase, p=2/3.

This dominance result holds for all rounds of the first phase. The distribution of females' responses first order stochastically dominates the one of males in all rounds with $p < 1$ and the reverse holds in all rounds with $p > 1$ (see Figure A1 in Appendix A). Gender differences remain very strong even in the last round of this first phase, the one with $p = \frac{4}{3}$, and in the round with $p = \frac{1}{2}$, the last round subjects played with $p < 1$. In that round, again, the distribution of female responses first order stochastically dominates the one for males (Kolmogorov-Smirnov, $p = 0.049$).[10] Hence any feedback-free

[9] This test compares distributions at pre-determined 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 compared distributions at all points between 0 and 100. In Appendix A, we report comparisons at a number of point responses.

[10] This test is based on the significance of the largest positive difference between two CDFs. The outcome of the test is that distribution g first order stochastically dominates f if the largest positive difference of f over g is significant but not the one of g over f. We report the p-value of the largest positive difference in CDFs in favor of the dominated distribution, which in this case is the distribution of responses of male participants.

learning (Weber, 2003) that might take place across rounds does not seem to reduce gender differences in behaviour.

To highlight further these strong gender differences, we classify subjects according to their observed depth of reasoning across the first phase. For each $p \neq 1$, we assign the level $k = \{1, 2, 3, \infty\}$ to individual response x_i when k minimizes $d = \left(x_i - 50p^k\right)^2$. We follow Coricelli and Nagel (2009) and classify a response as a low level response if $k = 1$ (high level otherwise). A subject's type is considered to be low (high) type if at least five out of her/his eight responses are of low (high) level. The rest of subjects are considered random and discarded from the analysis. This classification reflects how close an individual plays with respect to the equilibrium prediction. It partly incorporates beliefs about the sophistication of the opponent: Coricelli and Nagel (2009) show that high type subjects display a more intense activation in areas of the brain associated with theory of mind than low type subjects.[11]

As Table 2 shows, 80.7 % of the classified individuals in the sample are low type; from these, 58.9% are female. The percentages of high and low type subjects out of the whole subject pool (19.2% and 80.7% respectively) are similar to the ones obtained in previous studies.[12] However, these figures mask important gender differences. The small fraction of high type females (9.2%) stands out. It is significantly different from the proportion of high type males (Proportions test, $p < 0.001$).

	Low	High	Total
Males	69	30	99
Females	99	10	109
Total	168	40	208

Table 2: Type classification by gender (first phase).

These gender differences suggest that non-monetary incentives are at stake in these rounds. One non-monetary prize could be the utility of winning. In contests, Sheremeta (2010) finds that about a third of subjects are willing to spend a positive amount of money in order to win a zero value prize. Males might value winning the guessing game more highly than females, leading them to engage in more rounds of reasoning. We explore this issue further when we present our second experimental study in Section 4.2.

[11] For the rounds with $p > 1$, we take the equilibrium where all responses are 100 as the focal one and classify the sophistication of responses accordingly.

[12] Coricelli and Nagel (2009) obtain 41% and 59% respectively (n=20). Brañas-Garza et al. (2012) obtain 22% and 78% (n=191).

3.3.2 Second phase: monetary incentives and no gender priming

Now we move to the second phase of the experimental session, where we introduced monetary incentives. Recall that in this phase we applied two treatments: a control treatment (*No priming* treatment) and another in which we made gender, and therefore stereotypes, salient (*Priming* treatment).

In the control treatment, the gender composition of the group was always balanced. Table 3 shows that the distribution of responses of females has a larger mean and median than the one of males.

	Mean	Median	Std dev
Males	23.2	17	18.8
Females	29.2	24.5	21.3

Table 3: Aggregate results by gender in the *No priming* treatment.

Males seem to play according to more rounds of reasoning than females in the *No priming* treatment. However, the distributions of male's and female's responses in this treatment are not statistically different (Mann-Whitney, $p = 0.128$; Median test, $p = 0.153$). Furthermore, both the Kolmogorov-Smirnov and the Davidson-Duclos tests cannot rank these distributions in terms of first order stochastic dominance (see second column of Table A1 in Appendix A). This result is in line with Burnham et al. (2009) who find no significant gender differences in entries in the beauty contest when monetary incentives are provided.

Let us summarize our findings so far in the following result:

Result 1 Male subjects choose lower responses than females when there are no monetary incentives. Gender differences are not significant when monetary incentives are present.

3.3.3 Gender priming

As mentioned above, we primed gender by manipulating the gender composition of the group and the colour of the instruction booklets. This manipulation was expected to trigger simultaneously stereotypes about the relative mathematical and mentalising ability of males and females. We next study gender differences in observed depth of reasoning when gender is salient and the gender composition of the group changes.

The MG round Recall, that in the MG round of the *Priming* treatment, half of the participants in each room were male and half were female. The difference between the second phase of the *No priming* treatment and the MG round is that gender was made salient in the latter. Table 4 shows that the distribution of responses of females in the MG round has a lower mean and median than those of males.

	Mean	Median	Std dev
Males	27.4	25	18.3
Females	21.4	19	17.7

Table 4: Aggregate results by gender in the *Priming* MG round.

Females display deeper strategic reasoning than males in this round. The distributions of responses across genders in the MG round are statistically different (Mann-Whitney, $p = 0.021$; Median test, $p = 0.009$). Furthermore, the Kolmogorov-Smirnov ($p = 0.011$) and the Davidson-Duclos (see third column of Table A1 in Appendix A) tests provide a clear ordering between them, as Figure 2 illustrates. More specifically, the Davidson-Duclos dominance test establishes that there is a higher number of females than males who choose entries in the interval between 12 and 24.

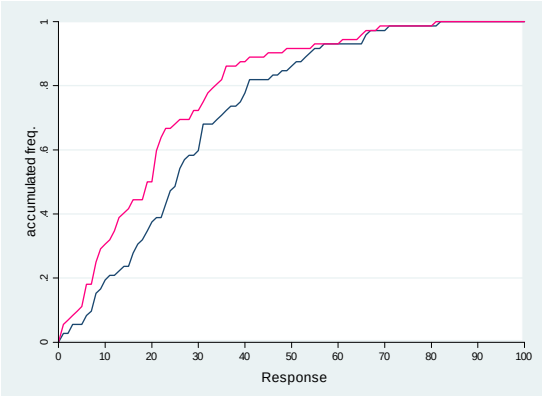


Figure 2: Cumulative distributions of responses by gender in the MG round.

The SG round In this round of the *Priming* treatment, participants played against opponents of their same gender. Table 5 shows the mean and median responses for males and females. Again, the average and median male response are higher than the average and median female responses.

	Mean	Median	Std dev
Males	29.9	26	20.9
Females	20.7	17	14.9

Table 5: Aggregate results by gender in the *Priming* SG round.

The distributions of responses across genders are statistically different (Mann-Whitney, $p = 0.009$; Median test, $p = 0.017$) and the dominance result is even stronger than in the MG round, as Figure 3 illustrates (see fourth column of Table A1 in Appendix A for the result of the Davidson-Duclos test).

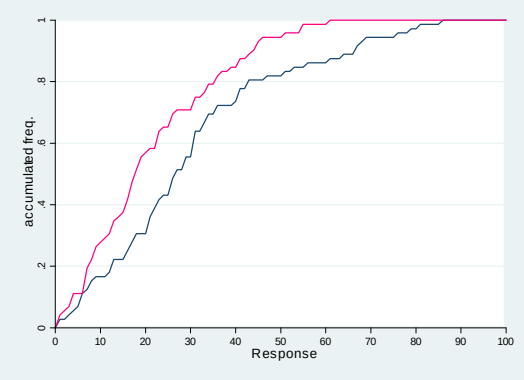


Figure 3: Cumulative distributions of responses by gender in the SG round.

Let us summarize our findings pertaining to the gender-primed rounds in the following result:

Result 2 When gender is primed and there are monetary incentives, females play according to more rounds of reasoning than males.

3.4 Results across rounds

3.4.1 *Priming* versus *No priming*

Let us now compare individual responses, first across the gender-balanced rounds of the *Priming* and *No priming* treatments, and then across the MG and SG rounds of the *Priming* treatment.

We expected gender priming to make beliefs about the relative strategic sophistication of males and females salient. If that were the case, responses

across the two treatments should change. But how? If there is the stereotype that one gender is inferior to the other in the game, members of that gender may feel *stereotype threat* (Steele, 1997), and become anxious about their performance. This might be the case for females if they perceive that the mathematical calculation involved in the guessing game favors males (Quinn and Spencer, 2001), or for males if they believe women are superior in mentalising or in strategic interactions in general. Stereotype threat has been consistently associated with higher emotional loads and cognitive impairment (e.g. Croizet et al., 2004; Krendl et al., 2008; Schmader and Johns, 2003). Hence, we would expect the threatened group to choose higher entries in the *Priming* treatment than in the *No priming* treatment. Individuals can also enjoy *stereotype lift* (Walton and Cohen, 2003) when they belong to the group they believe is superior in the task. If there is the stereotype that one gender is superior to the other, we would expect members of that group to choose lower entries in the *Priming* treatment compared to the *No priming* one.

Table 6 below compares responses by gender and across the gender-balanced rounds of the *Priming* and the *No priming* treatments. We observe that females change their behaviour considerably when gender is made salient. Their mean and median responses are much lower in the *Priming* treatment. Men change their answers to a lesser extent and in the opposite direction.

	Mean	Median	Std dev
Male, *Priming* MG	27.4	25	18.3
Female, *Priming* MG	21.4	19	17.7
Male, *No priming*	23.2	17	18.8
Female, *No priming*	29.2	24.5	21.3

Table 6: Aggregate results by gender and across gender-balanced rounds.

The distributions of responses in the two treatments differ only for females (Mann-Whitney, $p = 0.034$; Median test, $p = 0.052$). Differences in the distribution of males' responses across treatments are weaker (Mann-Whitney, $p = 0.138$; Median test, $p = 0.062$). But the Davidson-Duclos test can rank these distributions in terms of first stochastic dominance (see table A2 in Appendix A). As Figure 4 corroborates, fewer males choose entries between 10 and 18 and fewer females choose entries between 20 and 24 and between 32 and 42 under *Priming*. Let us summarise these findings in the following result.

Result 3 Females display more depth of reasoning in the *Priming* treatment than in the *No priming* treatment. The opposite holds for males.

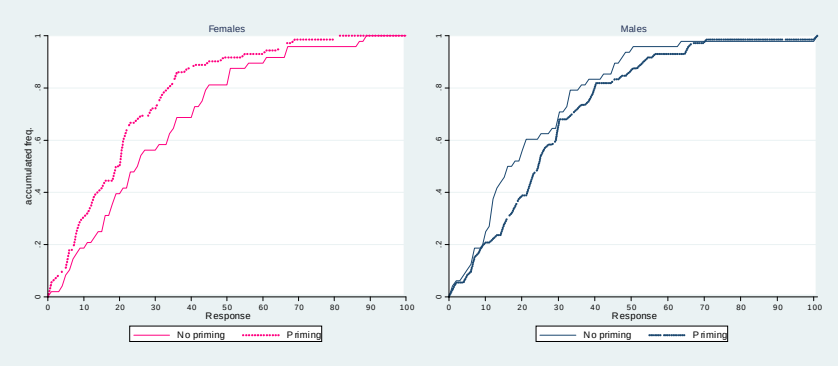


Figure 4: Cumulative distributions of responses by gender and treatment in the gender-balanced rounds.

The beauty contest involves a computation task. Gender priming may trigger negative stereotypes relating to the mathematical ability of females. This could lead female participants to exert lower cognitive effort and thus to respond higher numbers in the *Priming* treatment than in the *No priming* treatment. On the other hand, common wisdom is that women are better at imagining what others think and feel. This is supported by studies reporting female superiority in empathy and mentalising ability (Baron-Cohen, 2002; Krach et al., 2009). Gender priming may trigger this stereotype and thus encourage females' cognitive effort. Result 3 suggests that the second force is stronger, resulting in lower (higher) entries by female (male) participants in the *Priming* treatment. We investigate this idea further in Section 4.1 when analyzing subjects' beliefs about the relative strategic sophistication of men and women.

3.4.2 MG versus SG

Let us now compare the SG and the MG rounds. Recall that in the second phase of the *Priming* treatment we manipulated the gender composition of the groups of participants. The purpose of this manipulation was to explore the role of beliefs about the strategic sophistication of the opponents. In line with the findings in Agranov et al. (2012) and the model by Alaoui and Penta (2016), we conjectured that if an individual believes that a change in the gender composition shifts up (down) the distribution of levels of so-

phistication in the group, he/she will exert more (less) cognitive effort and his/her entry will decrease (increase).

Table 7 shows that the average and median responses of both sexes do not significantly differ across the SG and the MG rounds.

	Mean	Median	Std dev
Male, SG	29.9	26	20.9
Female, SG	20.7	17	14.9
Male, MG	27.4	25	18.3
Female, MG	21.4	19	17.7

Table 7: Aggregate results by gender across the SG and MG rounds.

The distributions of males' responses in the SG and MG rounds are not statistically different (Wilcoxon sign-rank, $p = 0.276$; Sign-test, $p = 0.427$). The same result applies to females' responses (Wilcoxon sign-rank, $p = 0.959$; Sign-test $p = 1.000$).

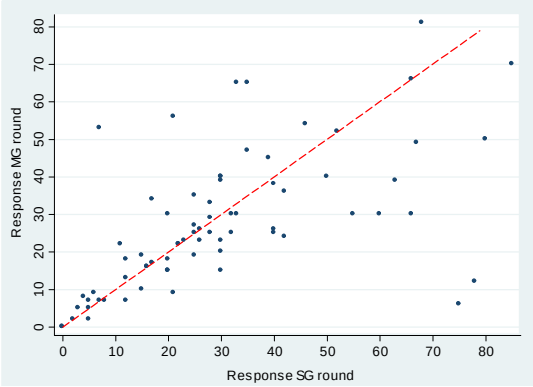


Figure 5: Males' responses in the MG and SG rounds.

A more detailed picture emerges from Figure 5, the scatterplot of males' entries in both rounds. Dispersion from the 45 degree line increases as responses are higher, and is denser below the line. This suggests that a larger number of male subjects decrease their response in the MG round compared to the SG round than in the other way around. Statistical tests confirm this. Male participants who in the SG round respond above the median decrease their entries in the MG round (one tailed Sign-test, $p = 0.040$). There is no significant change for males who choose entries above the median in the MG round. The lower responses in the MG round than in the SG round

are consistent with males increasing their depth of reasoning because they perceive women to be superior in the game. It is also in line with Krach et al. (2009), who in their fMRI study of the Prisoner's dilemma observe men compensating their weaker mentalising abilities. An alternative explanation is that males may expect level-0 female players to randomize over lower numbers than their male counterparts. In the next section, we explore these two hypotheses by analyzing the responses to the questionnaire administered at the end of each session.

Before that, let us summarize the results of the analysis across rounds: 1) Females react strongly to gender priming by engaging in more rounds of reasoning whereas males react to a lesser extent and in the opposite direction. And 2) males with higher responses in the SG round display deeper strategic reasoning in the MG round.

4 Supplementary analysis

In this Section, we analyse further the two dimensions studied above, namely, the effect of gender priming and the effect of financial incentives on strategic behaviour. To do so, we first study the answers to the questionnaire we administered to participants. Second, we present results from a classroom experiment which provides evidence on the existence of differential responses by gender to changes in monetary incentives.

4.1 Questionnaires

Participants in our laboratory study answered a questionnaire at the end of the session but before any feedback was provided. The aim of this questionnaire was twofold: To investigate whether priming was effective in activating gender stereotypes and to explore whether beliefs about the relative strategic sophistication of men and women influenced behaviour.

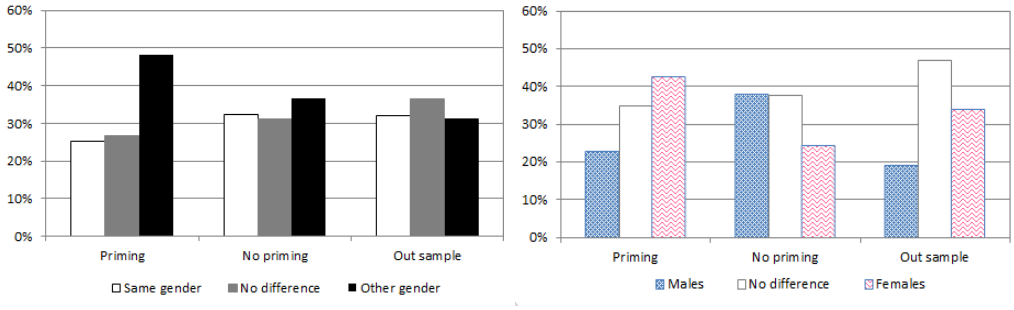
We focus on two questions: "When $p = \frac{2}{3}$, which sex responds higher numbers?" (Q1) and "Which sex is better at this game?" (Q2). We coded the free-text answers into four options, "Males", "Females", "No difference" and "Don't know." Q1 is designed to obtain information on beliefs about the behaviour of others and Q2 is designed to elicit perceptions about the relative sophistication of males and females. We have assumed, as it is customary in the literature, that level-0 behaviour is uniformly distributed over the set of strategies. But participants might have stereotypes on the ran-

dom behaviour of males and females which differ from that assumption.[13] In addition, participants may have different views as to what it means to be better in the game. This is ultimately irrelevant for our purposes because the gender-bias subjects may perceive in the game (if any) can affect their behaviour regardless of whether such perception is correct or not. In any case, responses to Q1 and Q2 display a medium to strong correlation (Contingency coefficient, 0.443; Cramér's V, 0.349), suggesting that participants associated a better performance with lower responses.

Answers to Q1 were not incentivised. Whilst this might reduce their validity, we show below that answers to that question have significant explanatory power. Admittedly, responses to both questions could be affected by the experiment itself. In order to have a cleaner source of information, we also ran this questionnaire on a comparable population of students from the University of Barcelona (n=134) who had not been exposed to the beauty contest game before. This allows us to compare responses across three populations, subjects who participated in the *Priming* treatment, those who participated in the *No priming* treatment, and respondents who did not participate in the experiment and had no knowledge of the game.

4.1.1 Was priming effective?

We saw in previous sections that priming had an effect on entries in the beauty contest, especially for females. Women in the *Priming* treatment chose lower entries than their counterparts in the *No priming* treatment. We observed the opposite effect, albeit weaker, for males. If priming was indeed effective in raising gender salience and stereotypes we should expect it to have an effect on responses to our questionnaire.



[13] A fraction of males responding to Q1 said that females tend to pick lower numbers such as birthdays or lucky numbers. A similar fraction of females responded that males tend to pick higher numbers because they like "speeding" and "big things in general."

Figure 6: Responses by gender to Q1 and Q2 (for females) by sample.

The left panel of Figure 6 shows the histogram of responses to Q1 by subsample. The distribution of responses of participants who played in the *No priming* treatment and those outside our subject pool are not significantly different. Responses are quite evenly distributed across the three possible answers. However, participants in the *Priming* treatment are more inclined to believe that the opposite sex tends to respond higher numbers. The difference with respect to the rest of answers to this question is statistically significant (chi-squared, $p = 0.020$). Our gender priming thus induced participants to believe that there were gender differences in entries in the game. The right panel of Figure 9 shows the effect of priming on responses to Q2 of female participants only. Female subjects in the *No priming* treatment tended to believe that either males are better at the game or that there are no gender differences, whereas female subjects in the *Priming* treatment tend to believe the opposite. The difference is weakly significant (chi-squared, $p = 0.099$). This pattern is in line with our previous result showing that females display higher strategic sophistication when gender is salient. Gender priming had no significant effect on males' responses to Q2.

4.1.2 Gender bias

The next question is whether answers to the questionnaire can help explain observed behaviour. The first issue we tackle relates to the association between behaviour and beliefs about which gender is better in the guessing game. Our aim is to study whether gender stereotypes, expressed in responses to Q2, might be related to observed depth of reasoning.

As the next result shows, this relationship is straightforward for males.

Result 4.1 The distribution of responses in the incentivised gender-balanced rounds of males who believe that males are better is different from the distribution of males who believe that females are better (Mann-Whitney, $p = 0.044$).

The median response in the incentivised gender-balanced rounds for the pooled sample of males who believe that females are better at the game is 25. It is 17 for males who believe that males are better at the game. The perceived gender-bias in the beauty contest is thus associated with depth of strategic reasoning in males. It might be that men who believe that the women have a relative advantage in the game choose lower entries because

they have lower cognitive abilities. It might also be that the causality is reversed, and males who respond higher numbers conclude that their gender is worse in the game (although there was no feedback until the very end of the session). We will come back to this last point below when looking at beliefs and gender composition.

Surprisingly, the data does not provide evidence on the existence of the analogous association in females. Their behaviour in the incentivised gender-balanced rounds is not related to their responses to Q2. However, a more careful exploration shows that priming has a decisive effect.

Result 4.2 Consider the subset of females who believe that females are better in the game. The distribution of responses of participants in the *No priming* treatment first order stochastically dominates the one of participants in the *Priming* treatment (Kolmogorov-Smirnov, $p = 0.048$).

This offers an explanation for Result 2. The belief of women on their own superiority in the game has an effect on their behaviour only when gender is made salient. The difference in median responses is striking: 40 for those women in the *No priming* treatment and 17 in the *Priming* treatment. It is important to note that, in this case, we can pin down the causality from perceptions to behaviour. It cannot be the case that behaviour affected their responses to Q2 because these are all women who believe that females are better in the game. So we can conclude that the combination of gender salience and the belief that women are better in the game boosted the observed depth of reasoning of these participants.

Interestingly, gender priming has no negative effect on females who answer that males are better in the game. This result might be due to the presence of a generalized positive stereotype in favor of women which might countervail their individual belief: In our out-sample survey, we asked an additional question (Q3): "Which gender obtains better results in strategic interactions?" A 42.5% of all respondents (57.5% for females) answered that females obtain better results, and 34.3% answered that no difference exists. This might explain why females who answer that males are better in the game do not display less strategic sophistication when gender is primed.

4.1.3 Gender composition

Let us now explore whether responses to the questionnaire can help us explain the changes we observed between the MG and the SG rounds of the

Priming treatment. Recall that we found that males who responded higher numbers in the SG round decreased their entries in the MG round. First we want to establish that, despite not being incentivised, responses to Q1 can help to explain this change in behaviour across rounds.

Result 4.3 The median responses of subjects who believe their same (the other) gender respond higher numbers is higher (lower) in the SG round than in the MG round (one-tailed Sign-test $p = 0.014$ and $p = 0.020$ respectively).

Now we can return to the question we left open at the end of Section 3. We had observed that males with higher responses in the SG round reduced their entries in the MG round. We mentioned that this was consistent with the perception that females are better in the game. We also mentioned that these males might have picked lower numbers in the MG round because they expected level-0 females to choose lower numbers than level-0 males. The analysis of Q1 and Q2 can shed light on this. Under the first hypothesis, males who changed behaviour should be those who believe that females are better at the game. According to the second hypothesis, males who pick higher entries in the SG round than in the MG one should answer to Q1 that men tend to pick higher numbers.

Males who believe that men respond higher numbers than females change their behaviour between the SG and the MG rounds (Wilcoxon sign-rank, $p = 0.050$). This would lend support to the hypothesis that males expected level-0 females to choose lower numbers than level-0 males. However, we find no significant opposite effect for male subjects who believe that females respond higher numbers. This begs the question of why changes in the gender composition affect only males who believe that their own gender responds higher numbers.

On the other hand, the distribution of entries in the SG round picked by males who believe that females are better at the game first order stochastically dominates the distribution of responses of males who believe the opposite (Kolmogrow-Smirnov, $p = 0.040$). Recall that we observed that males with higher responses in the SG round decrease their entries when playing the MG round. The combination of these two observations implies that males who decrease their answers when playing the MG round are more likely to believe that females are better in the game. The change in the median response across the two rounds for males who respond that females are better at the game is in line with this explanation: The median is 28 in the MG round and 33 in the SG round. All this lends support to the

hypothesis that men increase their cognitive effort in mixed gender groups because they believe that females are relatively more sophisticated. Unfortunately, our sample size does not allow for a more detailed analysis which can discriminate further between these two hypotheses.

4.2 Beauty in the classroom

4.2.1 Design and hypotheses

Let us present the results from a set of classroom experiments that we ran at the University of Edinburgh between 2005 and 2010. This study allows us to establish the robustness of Result 1 and to explore the existence of differential responses to financial incentives by gender. Participants were six cohorts of undergraduate students taking an Intermediate Microeconomics course. As part of the course, students had to fill an online problem set containing several game-theoretic questions implemented via the website *Games and behaviour*[14]. Cohorts ranged between 116 and 170 students. Completing the problem set was compulsory. Students had no previous instruction in game theory. They had a diverse background both by nationality and major of study. In total, 792 students took part, 39.4% of them were female.[15] Although we do not have information about the exact gender composition of each cohort, they mirrored this proportion fairly consistently.

In one of the questions in the problem set, students had to guess two-thirds of the average of all responses of students in the class.[16] In the 2007, 2008 and 2010 cohorts (n=401), a prize of £10 (about 12 euros) was given to the student who made the best guess. If there were more than one winner, the prize was divided. We call this the *Prize* treatment. The *No prize* treatment corresponds to the other three cohorts (n=391) in which no money was awarded to the winner. The instructor did not mention in class that the name of the winner(s) was to be announced publicly. So for the *No*

[14] Developed by Ariel Rubinstein and Eli Zvuluny. Available at http://gametheory.tau.ac.il/.

[15] When retrieving the data from the website, we were provided first with the list of participants' names but without their responses in order to ensure anonimity. We then assigned gender to these names and returned the list. We then received the data associating responses to the gender of the responder.

[16] The exact phrasing was: "Each of you (the students in this course) have to choose an integer between 0 and 100 in order to guess 2/3 of the average of the responses given by all students in the course. Each student who guesses 2/3 of the average of all responses rounded up to the nearest integer, will receive a prize to be announced by your teacher (or alternatively will have the satisfaction of being right!)." This phrasing was identical in the two treatments.

prize treatment, such non-monetary reward was not made explicit.[17]

There are clear differences between this study and the gender neutral part of the lab experiment presented in previous sections. The lab study was within subjects; monetary incentives were bigger and groups were smaller. The first (non-incentivised) phase of the lab study was meant to familiarize subjects with the design, whereas in the classroom study the absence of financial incentives was a full treatment. And perhaps more importantly, students in the classroom study could potentially communicate with each other or look up for solutions elsewhere. We thus lose control over some important elements of the experiment. The advantage is that we can go beyond the lab and enhance the external validity of Result 1.

In addition, this experiment allows us to study gender differences in the response to enhanced monetary incentives because, unlike in the lab study, we are comparing only first responses. The model by Alaoui and Penta (2016) suggests that participants in the *Prize* cohorts should display engage in more rounds of reasoning, but it remains silent on the existence of differential responses by gender. Results in Fryer at al. (2008) suggest that we should expect men to react more strongly than women to the presence of the monetary reward. But their study is not directly comparable to ours since they use a general knowledge test instead of a strategic interaction. Finally, stronger monetary incentives might induce differential responses by gender depending on its perceived gender-bias (Guenther et al. 2010; Shurchkov, 2012). If the game is perceived to be female-biased, as responses to the questionnaire in the lab study seem to suggest, higher financial incentives should inducing females to exert more competitive effort.

4.2.2 Results

Table 8 reports the aggregate results by gender in the two treatments.

	Mean	Median	Std dev	Obs.
Male, *No prize*	37.6	35	23.9	243
Female, *No prize*	41.9	42	23.3	148
Male, *Prize*	35.7	33	22.7	237
Female, *Prize*	36.4	34	23.5	164

Table 8: Aggregate results by gender and treatment.

[17] This does not rule out that students could seek prestige or status among their peers.

The table suggests that, as in the laboratory study, males choose lower entries than females when monetary incentives are absent, and that gender differences vanish when a prize is awarded. Indeed, the distributions of responses across genders differ in the *No prize* treatment (Mann-Whitney, $p = 0.049$; Median test, $p = 0.036$), and do not differ in the *Prize* treatment (Mann-Whitney, $p = 0.688$; Median test, $p = 0.316$).

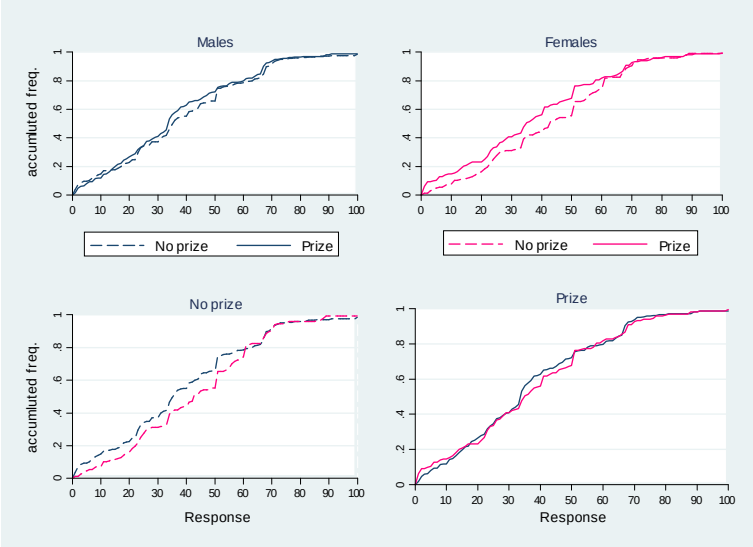


Figure 7: Cumulative distributions of responses by gender and treatment.

Figure 7 breaks down the cumulative distributions of responses by gender across treatments (upper panels) and by treatments across genders (lower panels). The cumulative distributions of responses in the *No prize* treatment are depicted with dashed lines. The lower left panel shows the cumulative distributions of male and female responses under the *No prize* treatment. The Davidson-Duclos test concludes that the distribution of female responses first order stochastically dominates the distribution of male responses (see third column of Table A3 in Appendix A). On the other hand, the lower right panel shows virtually no gender differences in the cumulative distributions of responses in the *Prize* cohorts. We thus corroborate Result 1: When financial incentives are absent, females play according to less rounds of reasoning than males, but when financial incentives are present, gender differences are no longer significant.

The results also show that female participants react more strongly to monetary incentives than males, to the extent that gender differences vanish when a monetary prize is at stake. Table 8 and the upper panels of Figure

7 show that females change their behaviour when there is a monetary prize whereas males' responses do not differ much across the two treatments. The distributions of females' responses differ across treatments (Mann-Whitney, $p = 0.026$; Median test, $p = 0.054$), but the distributions of males' responses do not. (Mann-Whitney, $p = 0.334$; Median test, $p = 0.154$). In addition, the upper right panel shows that the distribution of female responses under *No prize* first order stochastically dominates the distribution of female responses in the *Prize* treatment (Kolmogorov-Smirnov, $p = 0.031$), whereas dominance for males is unclear. The Davidson-Duclos test corroborates this (see Table A3 in Appendix A).

Result 5 Females display deeper strategic reasoning in the *Prize* treatment than in the *No prize* treatment. Males do not respond to the presence of a monetary incentive.

That females respond significantly to monetary incentives runs against the results in Fryer et al. (2008). It also suggests that the conclusion of Camerer and Hogarth (1999) whereby monetary incentives have a small effect in experimental games might not necessarily apply to female populations. If financial incentives constitute a cue indicating that the beauty contest is a competitive interaction, it is to be expected that females react to this contextual information more strongly than men (Croson and Gneezy, 2009).

On the other hand, the fact that males do not react to the presence of a monetary prize also suggests that males may consider that a non-monetary prize is at stake in the *No prize* treatment. This is consistent with males in the *No prize* treatment displaying similar depth of strategic reasoning to females in the *Prize* treatment. But since economic incentives did not affect their strategic effort, it might be the case that males either regarded the prize as of relatively low value or that the monetary incentive crowded out any psychological reward.

5 Discussion and conclusions

In this paper, we explored the existence and endogeneity of gender differences in observed strategic reasoning. We used the beauty contest game as experimental design. We made no attempt to speculate on where the differences we observed come from.

We reported results from two studies. The first study was a laboratory experiment where we manipulated monetary incentives, gender salience

and gender composition. We found that the gender differences in observed depth of reasoning that emerge in zero-stakes rounds disappear when financial incentives are introduced. Gender differences reappear when gender and stereotypes are made salient. Females react very strongly to priming by lowering their responses. Males react less strongly and in the opposite direction. The effect of changes in the gender composition of the group is weak and only applies to a subset of males, who seem to engage in more rounds of reasoning in mixed gender groups compared to single gender groups.

In the guessing game, gender salience may trigger opposed stereotypes relating to potential performance of females. On the one hand, it may trigger negative stereotypes relating to the mathematical skills of females. On the other hand, gender salience may trigger positive stereotypes about female's advantage in understanding opponents. A caveat of our design is that it does not allow us to isolate these two forces, although our results suggest the latter dominates. This might be due to the complexity of the calculation involved in the beauty contest being relatively small compared to the difficulty of mentalising it requires. Responses to a questionnaire we administered to our participants lend support to this interpretation. Females who react to gender priming are those who believe that females are better in the game. In addition, males who answer that females are better at the game choose higher entries when gender is salient, suggesting that they might be experiencing stereotype threat.

The second study was a large classroom experiment. Its main result is that females engage in more rounds of reasoning when a financial reward is introduced whereas men do not. Note that we only study the effect of introducing economic incentives. Given the size of the cohorts and the value of the prize, this probably had a purely cuing effect. Further research might explore the effects of changes from low, but positive, to higher monetary incentives on strategic sophistication.

Our first remark is that we observe gender differences only when we manipulate monetary incentives and priming. This might explain why there are so few studies reporting gender differences (or the lack of) in strategic interactions. In incentivised experiments, gender differences might arise only if gender is made salient. We are aware that gender differences can emerge spuriously when subjects' characteristics correlate with gender, e.g. major of study.[18] Our subject pool in the lab study was relatively homogeneous so we are relatively free from this problem.

In addition, the present paper is one of the few where women are ob-

[18] We thank Colin Camerer for pointing this out.

served to outperform men and where gender salience is beneficial to female performance. One exception is Shurchkov (2012), who obtains that women surpass men in a low-pressure verbal task. Our result on priming is also in sharp contrast with the literature on the effect of gender information on performance in mathematical tests. Inzlicht and Ben-Zeev (2000) find that simply placing a woman in a room with men decreases her test performance. Danaher and Crandall (2008) find that just marking one's gender after an advanced placement calculus test rather than before the test, led to a 33% reduction in the performance gender-gap. Our results suggest that gender salience in strategic interactions may lead to increases in depth of reasoning in females. Since gender priming seems to be detrimental for males, selective gender salience might be even a more effective intervention.

Our final remark refers to the portability of our results. The beauty contest is a relatively complex game with a big strategy space. Hence, it is to be expected that players use simple rules of play, even non-strategic ones (Fragiadakis et al., 2013). In fact, level-k theories can be interpreted as rules of thumb grounded on "an instinctive reaction to the game" (Crawford et al., 2013). These rules might change with how instructions are laid out (Georganas et al., 2015) and with the strategy space (Benhabib et al., 2014). It is natural to expect simple rules of play to be sensitive to individual characteristics and gender salience. Further research should address whether the gender differences in strategic behaviour that we uncover in the guessing game remain in other games where standard equilibrium predictions are more transparent and where subjects may be less prone to resort to simple heuristics.

References

[1] Agranov, M, Potamites, E, Schotter, A, and Tergiman, C. 2012. Beliefs and Endogenous Cognitive Levels: An Experimental Study, *Games and Economic behaviour*, 75(2): 449463.

[2] Alaoui, L, and Penta, A. 2016. Endogenous Depth of Reasoning, *Review of Economic Studies,* 83(4): 1297-1333.

[3] Arad, A, and Rubinstein, A. 2012. The 11-20 Money Request Game: A Level-k Reasoning Study, *American Economic Review*, 102(7): 3561-3573.

[4] Azmat, G, Casalmiglia, C, and Iriberri, N. 2016. Gender Differences in Response to Big Stakes, *Journal of the European Economic Association*, 14(6): 1372–1400.

[5] 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.

[6] Baron-Cohen, S. 2002. The Extreme Male Brain Theory of Autism, *Trends in Cognitive Sciences*, 6(6): 248–54.

[7] Benhabib, J, Duffy, J, and Nagel, R. 2014. De-framing Rules to (De)-Anchor Beliefs through Sentiments in Beauty Contest Experiments, Barcelona GSE, unpublished manuscript.

[8] Bosch-Domenech, A, Garcia-Montalvo, J, Nagel, R, and Satorra, A. 2002. One, Two, (Three), Infinity: Newspaper and Lab Beauty-Contest Experiments, *American Economic Review*, 92(5): 1687-1701.

[9] Brañas-Garza, P, Garcıa-Muñoz, T, and Hernan, R. 2012. Cognitive Effort in the Beauty Contest Game, *Journal of Economic behaviour and Organization*, 83(2): 254–260.

[10] Bühren, C, and Björn, F. 2010. Chess Players' Performance Beyond 64 Squares: A Case Study on the Limitations of Cognitive Abilities Transfer, MAKGS, unpublished manuscript.

[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 behaviour 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] Camerer, C F, and Hogarth, R. 1999. The Effects of Financial Incentives in Experiments: A Review and Capital-Labor Production Theory, *Journal of Risk and Uncertainty,* 19(1–3): 7–42.

[14] Choi, S. 2012. A Cognitive Hierarchy Model of Learning in Networks, *Review of Economic Design*, 16(2): 215-250.

[15] Coricelli, G, and Nagel, R. 2009. Neural Correlates of Depth of Strategic Reasoning in Medial Prefrontal Cortex, *Proceedings of the National Academy of Sciences*, 106(23): 9163-9168.

[16] Costa-Gomes, M A, Crawford, V P, and Broseta, B. 2001. Cognition and behaviour in Normal-Form Games: An Experimental Study, *Econometrica,* 69: 1193-1235.

[17] Crawford, V P, Costa-Gomes, M A, and Iriberri, N. 2013. Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications, *Journal of Economic Literature*, 51: 5-62.

[18] Croizet, J, Després, G, Gauzins, M, Huguet, P, Leyens J, and Méot, A. 2004. Stereotype Threat Undermines Intellectual Performance by Triggering a Disruptive Mental Load, P*ersonality and Social Psychology Bulletin*, 30(6): 721-731.

[19] Croson, R, and Gneezy, U. 2009. Gender Differences in Preferences, *Journal of Economic Literature*, 47(2): 1-27.

[20] Danaher K, and Crandall, C S. 2008. Stereotype Threat in Applied Settings Re-examined, *Journal of Applied Social Psychology*, 38(6): 1639-1655.

[21] 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.

[22] Fragiadakis, D E, Knoepfle, D T, and Niederle, M. 2013. Identifying Predictable Players: Relating behavioural Types and Subjects with Deterministic Rules, Stanford University, unpublished manuscript.

[23] Frick, B. 2011. Gender Differences in Competitiveness: Empirical Evidence from Professional Distance Running, *Labour Economics*, 18(3): 389-398.

[24] Fryer, R, Levitt, S, and List, J A. 2008. Exploring the Impact of Financial Incentives on Stereotype Threat: Evidence from a Pilot Study, *American Economic Review*, 98(2): 370-375.

[25] Georganas, S, Healy, P, and Weber, R. 2015. On the Persistence of Strategic Sophistication, *Journal of Economic Theory*, 159: 369-400.

[26] Gill, D, and Prowse, V L. 2015. Cognitive Ability and Learning to Play Equilibrium: A Level-k Analysis, forthcoming *Journal of Political Economy*.

[27] Gneezy, U, Niederle M, and Rustichini, A. 2003. Performance in Competitive Environments: Gender Differences, *Quarterly Journal of Economics*, 118: 1049–74.

[28] Gneezy, U, and Rustichini, A. 2004. Gender and Competition at a Young Age, *American Economic Review*, 94: 377–81.

[29] 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 behaviour and Organization*, 75: 395-401.

[30] Ho, T-H, Camerer, C F, and Weigelt, K. 1998. Iterated Dominance and Iterated Best Response in Experimental 'p-Beauty Contests', *American Economic Review*, 88(4): 947-969.

[31] Iriberri, N, and Rey-Biel, P. 2016. Stereotypes are Only a Threat when Beliefs are Reinforced: On the Sensitivity of Gender Differences in Performance under Competition to Information Provision, Barcelona GSE, unpublished manuscript.

[32] Inzlicht, M, and Ben-Zeev, T. 2000. A Threatening Intellectual Environment: Why Females are Susceptible to Experiencing Problem-solving Deficits in the Presence of Males, *Psychological Science*, 11: 365-371.

[33] Kocher, M, and Sutter, M. 2005. The Decision Maker Matters. Individual versus Team behaviour in Experimental Beauty-contest Games, *Economic Journal*, 115: 200-223.

[34] Kovalchik, S, Camerer, C F, Grether, D M, Plott, C R, and Allman, J M. 2005. Aging and Decision Making: A Comparison Between Neurologically Healthy Elderly and Young Individuals, *Journal of Economic behaviour and Organization*, 58: 79–94.

[35] Krach, S, Blumel, I, Marjoram, D, et al. 2009. Are Women Better Mindreaders? Sex Differences in Neural Correlates of Mentalizing Detected with Functional MRI, *BMC Neuroscience*, 10, 9.

[36] Krendl, A C, Richeson, J A, Kelley, W M, and Heatherton, T F. 2008. The Negative Consequences of Threat: An fMRI Investigation of the

Neural Mechanisms Underlying Women's Underperformance in Math, *Psychological Science*, 19(2): 168-175.

[37] Nagel, R. 1995. Unraveling in Guessing Games: An Experimental Study, *American Economic Review*, 85(5): 1313-1326.

[38] Nguyen, H-H D, and Ryan, A M. 2008. Does Stereotype Threat Affect Test Performance of Minorities and Women? A Meta-Analysis of Experimental Evidence, *Journal of Applied Psychology*, 93(6): 1314-1334.

[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] Palacios-Huerta, I, and Volij, O. 2009. Field Centipedes, *American Economic Review*, 99(4): 1619-1635.

[41] Petrie, R, and Segal, C. 2015. Gender Differences in Competitiveness: The Role of Prizes, GMU Working Paper, unpublished manuscript.

[42] Quinn, D, and Spencer, S. 2001. The Interference of Stereotype Threat with Women s Generation of Mathematical Problem-solving Strategies, *Journal of Social Issues*, 57: 55–71.

[43] Schmader, T, and Johns, M. 2003. Converging Evidence that Stereotype Threat Reduces Working Memory Capacity, *Journal of Personality and Social Psychology*, 85: 440–452.

[44] Sheremeta, R. 2010. Experimental Comparison of Multi-Stage and One-Stage Contests, *Games and Economic behaviour*, 68: 731-747.

[45] 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.

[46] Stahl, D O, and Wilson, P W. 1994. Experimental Evidence on Players' Models of Other Players, *Journal of Economic behaviour and Organization*, 25: 309–327.

[47] Stahl, D O, and Wilson, P R. 1995. On Players' Models of Other Players: Theory and Experimental Evidence, *Games and Economic behaviour*, 10(1): 218-254.

[48] Steele, C M. 1997. A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance, *American Psychologist*, 52: 613–29.

[49] Strzalecki, T. 2014. Depth of Reasoning and Higher Order Beliefs, *Journal of Economic behaviour and Organization*, 108: 108-122.

[50] Walton, G M, and Cohen, G L. 2003. Stereotype Lift, *Journal of Experimental Social Psychology,* 39: 456–467.

[51] Weber, R A. 2003. 'Learning' With No Feedback in a Competitive Guessing Game, *Games and Economic behaviour,* 44(1): 134–144.

Appendix A: Additional Tables and Figures

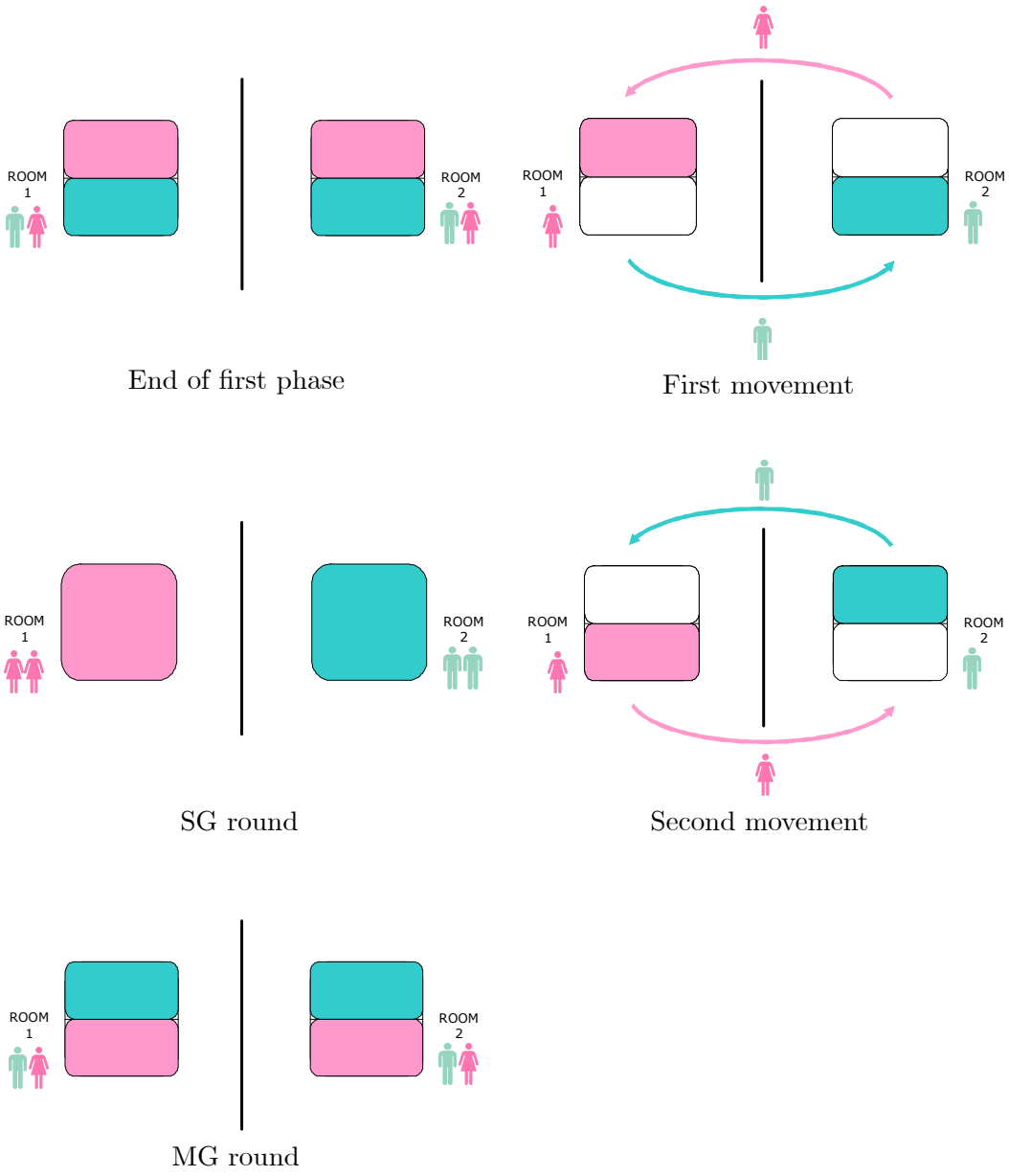


Figure A1: Moves of participants in the SG-MG order of the *Priming* treatment.

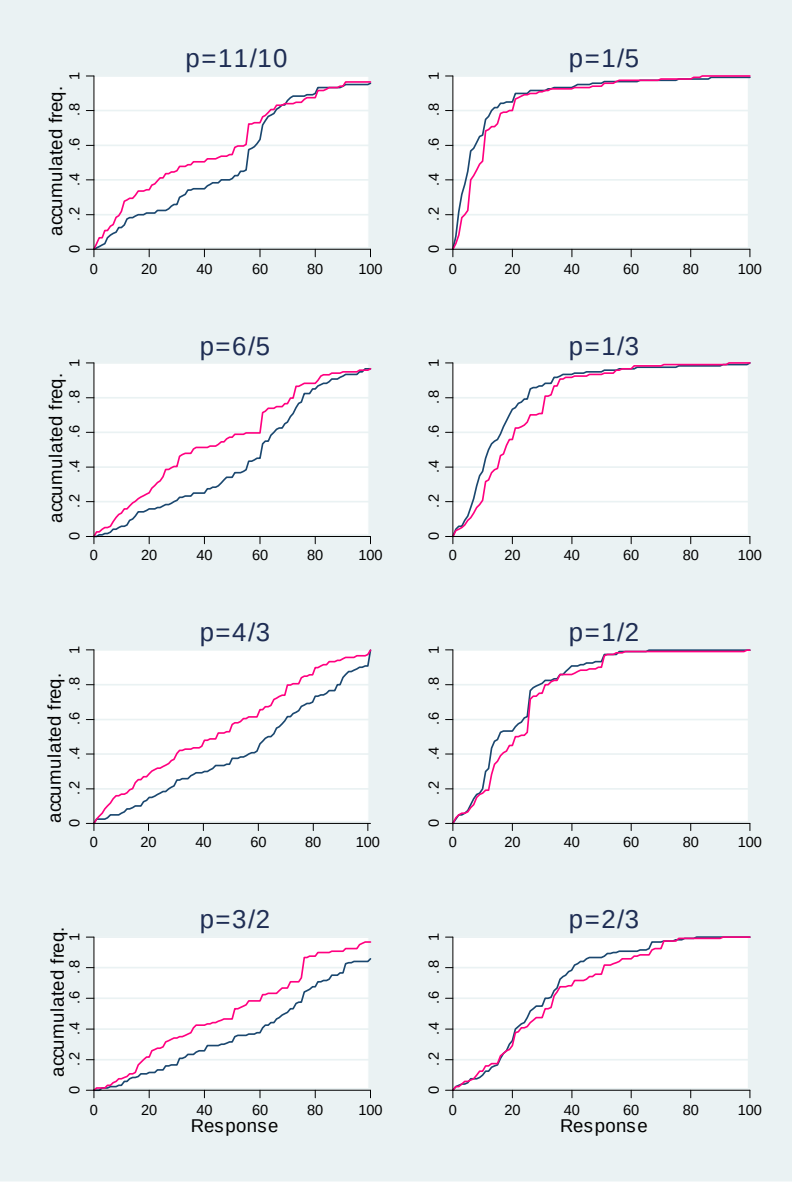


Figure A2: Cumulative distributions of responses by gender to phase 1 rounds.

	No incentives	*No priming* MG	*Priming* MG	*Priming* SG
1	-0.3590	-0.5876	1.1679	0.8360
6	0.0000	1.1424	1.4564	1.1424
10	0.7027	-0.4862	1.5247	1.8045
15	0.2959	-1.4846	2.1139**	2.1552**
22	-0.3697	-1.2388	2.9304***	2.7395***
33	-0.5036	1.1395	1.5524	1.1395
44	-2.4640***	-0.5485	1.4564	2.2548**
50	-1.1939	-1.4941	0.8203	2.5082***
67	-1.3783	-0.5876	0.0000	2.3180**
	Positive (negative) t-statistics indicate that the accumulated frequency of female (male) responses is higher than for the other sex.			

Table A1: Female-male DD test t-statistics per round of the lab study.

	No priming vs *Priming* Males	*No priming* vs *Priming* Females	MG vs SG Males	MG vs SG Females
1	0.3997	-1.3367	0.0000	0.3444
6	0.4696	-0.5091	-0.5308	-0.2135
10	2.0239**	-1.3829	0.6414	0.3619
15	1.8956*	-1.4839	0.3783	0.3367
22	1.2168	-2.4439***	0.1686	0.3502
33	1.2168	-2.1491**	0.0000	0.4106
44	0.5091	-1.3620	0.2135	-0.6037
50	1.7186*	-0.7210	0.7095	-1.0366
67	0.2455	-0.3997	1.1679	-1.4342
	Positive (negative) t-statistics indicate that the accumulated frequency of the first (second) element in the comparison is higher than the other.			

Table A2: DD test t-statistics for round comparisons in the lab study.

	No prize vs *Prize* Males	*No prize* vs *Prize* Females	Males vs Females *No prize*	Males vs Females *Prize*
1	-1.2772	3.1908***	-3.1385***	1.7107*
6	-0.9446	2.3097**	-2.3406***	1.0942
10	-0.7632	1.8925*	-1.9509*	0.2484
15	0.8216	2.3290***	-1.7346*	0.1031
22	0.4790	1.7681*	-1.6279	-0.1163
33	1.4628	1.4961	-1.4248	-1.1050
44	0.7612	1.8036*	-2.0286**	-0.7584
50	0.1599	2.0818**	-1.9510*	0.1594
67	1.0358	0.6773	-0.3669	-0.5475
	Positive (negative) t-statistics indicate that the accumulated frequency of the first (second) element in the comparison is higher than the other.			

Table A3: Davidson-Duclos (DD) test t-statistics for the classroom study.

Appendix B: Accuracy

A key assumption in the analysis above has been the association between lower entries and higher strategic sophistication. However, this assumption does not take into account that positive entries are a better response than the Nash equilibrium strategy when opponents exhibit imperfect strategic sophistication. In this appendix we check the robustness of our results to the use of an alternative measure of sophistication: The quadratic distance to the winning response. This measure of (lack of) sophistication is similar to strategic IQ in Coricelli and Nagel (2009). It accounts both for depth of reasoning and for the correctness of beliefs about others' responses. We compute the average quadratic distance to the winning response (the inverse of accuracy) for the eight rounds of the first phase with $\rho \neq 1$, and for each of the rounds of the second phase in both treatments, *Priming* and *No priming*.

There are substantial gender differences in the distributions of the average distance to the winning responses in the first phase (Mann-Whitney, $p < 0.001$; Median test, $p < 0.001$). Figure B1 depicts the corresponding kernel densities. Female players (flatter curve) tend to be less accurate than male players. This confirms our results in Section 3.3.1.

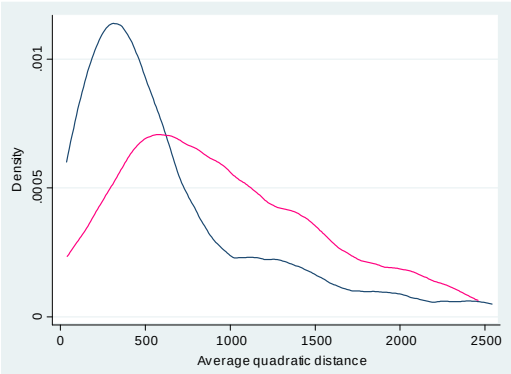


Figure B1: Average quadratic distance to winning response in the first phase by gender.

In Figure B2 we extend the analysis to the classification by levels of sophistication introduced in Section 3.3.1. The distributions of average quadratic distances for low and high sophisticated individuals are statistically different (Mann-Whitney, $p < 0.001$; Median test, $p < 0.001$). The average quadratic distance to the winning response is significantly higher for individuals we classified as low sophisticated. Hence, there is a

close relationship between that classification and accuracy in the first phase.

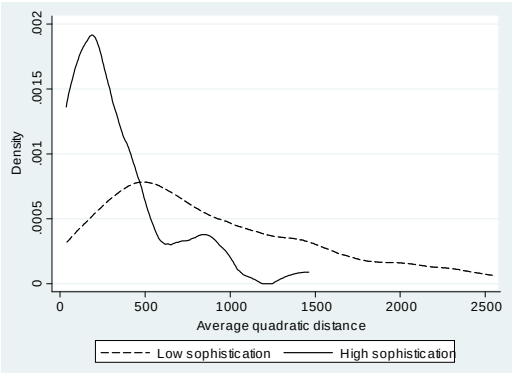


Figure B2: Average quadratic distance to winning response in the first phase by level of sophistication.

Next we compare the accuracy of responses in the gender-balanced rounds of the *Priming* and *No priming* treatments. Recall that Result 2 stated that gender priming had the effect of significantly lowering the entries of female participants and of increasing males' responses. These results remain, albeit less sharply, when looking at accuracy. The upper panels of Figure B3 below present the comparison of accuracy across treatments, *Priming* (solid line) versus *No priming* (dotted line), for males and females. The upper right panel shows that females' entries are indeed closer to the winning response in the *Priming* treatment than under *No priming* (Mann-Whitney, $p = 0.066$; Median test, $p = 0.044$). *Priming* does not change males' accuracy though.
The lower panels of Figure B3 display the comparison across genders by treatment. There are no gender differences in accuracy in the *No priming* treatment. The lower right panel shows that accuracy is higher for females than for males in the *Priming* treatment (Mann-Whitney, $p = 0.066$; Median test, $p = 0.046$).[19] This corroborates Result 2.

[19] A similar effect is also observed in the same gender round of the *Priming* treatment (Mann-Whitney, $p = 0.034$; Median test, $p = 0.067$).

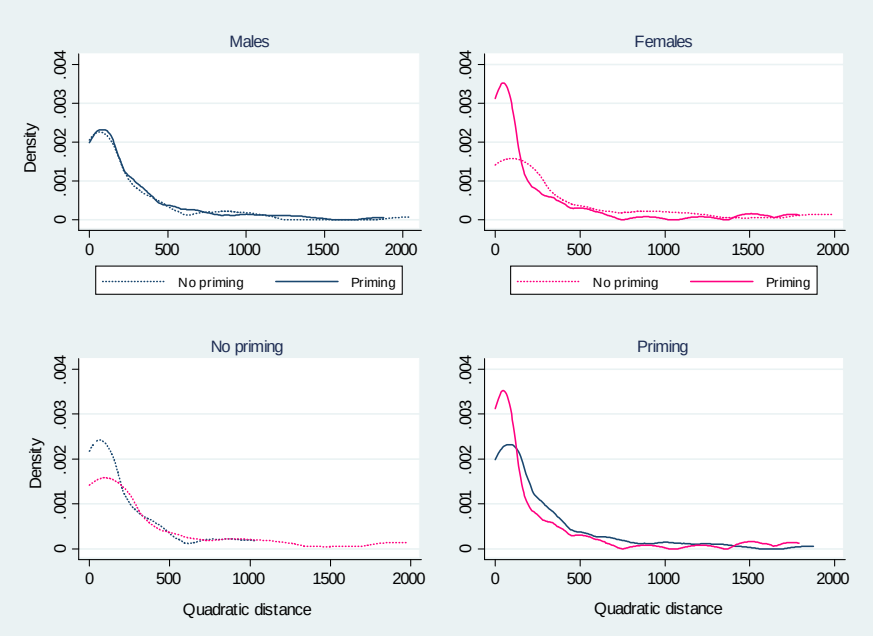


Figure B3: Quadratic distance to winning response by gender and treatment.

Finally, we do not find any substantial effect on accuracy due to changes in the gender composition of the group. The distributions of quadratic distances for both males and females do not differ across the SG and the MG rounds of the *Priming* treatment, as Figure B4 below shows.

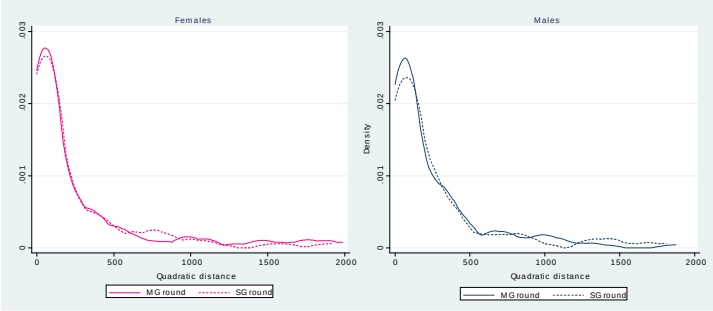


Figure B4: Quadratic distance to the winning response by gender across rounds of the *Priming* treatment.

Appendix C: Instructions of Study 2 (translated from Spanish)

GENERAL INSTRUCTIONS

Hello. Many thanks for taking part in this session.

The purpose of this session is to study how people make decisions in strategic settings.

The session is organized in two parts:

In the first part, you should answer a series of independent questions with the objective of becoming familiar with the rules of the experiment.

In the second part, you should answer another series of independent questions. You will compete with the rest of participants in your room for a monetary prize. The participant with the most correct answer will be the winner.

After reading these instructions you will find the first set of questions. We will read each question aloud. You will have time to answer each question before moving to the next one.

Read carefully each question and take the time you need to answer it.

It is very important that you remain silent during the whole session. Otherwise, the data collected will be useless.

Please do not go to the next question until we tell you to.

Before starting the experiment please write in the box below your participant number.

GENERIC ROUND QUESTION (PHASE 1)

Each one of you should choose a number between 0 and 100 with the objective of guessing (p fraction of) the average of the numbers chosen by all the participants in this room.

The winner will be the participant(s) whose answer is the closest to the (p fraction of the) average of all numbers chosen.

Which number do you choose?

Do not go to the next question until being instructed to do so.

INSTRUCTIONS PHASE 2

Now the second phase of the experiment begins.

In this phase, you will participate in tow independent rounds. The structure and rules are similar to those of phase 1 but there are two main differences:

1. The identity of the participants you will compete with will change in each round.

2. There will be two monetary prizes of 40 euros each.

At the end of the second phase, one of the two rounds will be chosen randomly. The winner of this round will obtain the prize. If there is more than one winner in the chosen round, the prize will be split among the winners.

Again questions will be read aloud.

Read carefully each question and take the time you need to answer it.

Recall that it is very important that you remain silent during the whole session. Otherwise, the data collected will be useless.

Please do not go to the next question until we tell you to**.**

Before continuing please write in the box below your participant number.

GENERIC ROUND QUESTION (PHASE 2)

Each one of you should choose a number between 0 and 100 with the objective of guessing the "2/3 of the average" of the numbers chosen in this question by all the participants in this room.

The winner will be the participant(s) whose answer is the closest to the 2/3 of the average of all numbers chosen in this question by all the participants in this room.

Which number do you choose?

Now close the booklet and remain silent.