WORKING PAPERS

Col.lecció d'Economia E16/338

Gender differences and stereotypes in strategic thinking

Maria Cubel
Santiago Sanchez-Pages

UNIVERSITAT DE BARCELONA

UB Economics Working Papers 2016/338

Gender differences and stereotypes in strategic thinking

Abstract: Recent literature has emphasized that individuals display varying levels of strategic reasoning. This paper presents ten years worth of experimental data from two countries exploring the existence and endogeneity of gender differences in strategic sophistication. We report results from two experimental studies employing the beauty contest game, one from the classroom and one from the laboratory. We observe robust and signi…cant gender differences in strategic sophistication 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 signi…cantly higher levels of strategic sophistication than males when gender is made salient. This effect of gender priming is driven by females who believe women are superior in the game.

JEL Codes: C72, C91, D81, J16.

Keywords: guessing game, strategic sophistication, gender, stereotype threat, beliefs.

Maria Cubel
Universitat de Barcelona

Santiago Sanchez-Pages
Universitat de Barcelona

Acknowledgements: We thank E Adamopoulou, Larbi Alaoui, Ayala Arad, Peter Backus, Colin Camerer, Dirk Foremny, Tatiana Kornienko, John Morgan, Muriel Niederle, Ana Nuevo-Chiquero, Amedeo Piolatto, Lise Vesterlund and audiences at Alicante, Barcelona, Edinburgh, Pittsburgh, the University of New South Wales, the COSME-FEDEA workshop and the RES 2016 conference for their help and useful comments. We are particularly indebted to Ariel Rubinstein for retrieving the data featured in Study 1 and for his constructive criticisms. All remaining errors are completely ours. Both authors acknowledge …nancial support from the Spanish Ministry for Science and Innovation research grant ECO2012-33243 and from the Generalitat de Catalunya grant 2009SGR1051.

ISSN 1136-8365

1 Introduction

The experimental literature provides substantial evidence of the existence of strong individual heterogeneity in strategic sophistication. Observed individual behavior 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 seems to reflect differences in the extent to which individuals engage in mentalizing processes or "theory of mind", that is, the activity of thinking about others' thoughts, emotions and intentions (Baron-Cohen, 1991). Models of k-level thinking have been proposed in order to account for these experimental results. These models acknowledge that individuals have different cognitive levels and non-equilibrium beliefs about the sophistication of others.[1]

But strategic sophistication is a complex concept to pin down. Depth of strategic reasoning might depend on innate mentalizing or cognitive abilities, beliefs about the sophistication of others, and on the incentives provided. An individual may be sophisticated enough to choose the strategy corresponding to the standard game theoretical prediction but that choice may fail to acknowledge that the rest of the population might be incapable of that. It would be questionable to label such choice as more sophisticated than one that departs from the game theoretical prediction but takes correctly into account the heterogeneity in strategic sophistication in the population. On the other hand, a person may be reluctant to engage in further levels of reasoning, which require extra mental effort, when stakes are low or when opponents are perceived as strategically unsophisticated. Responses in those cases might not reflect the mentalizing ability of individuals, but rather their lack of motivation to engage in the process. In short, it should be natural to expect observed strategic sophistication to depend on both beliefs about the sophistication of others and on incentives.[2]

In this paper, we explore the heterogeneity and endogeneity of strategic sophistication in the context of gender. Gender constitutes an obvious source of observable heterogeneity across individuals. Hence, gender can bring up

[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. Level-k models have been applied to a number of strategic interactions such as communication and auctions. For a survey, see Crawford et al. (2013).

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

relevant questions in the analysis of strategic sophistication. In particular, we investigate three questions that, to the best of our knowledge, have not been addressed in the literature before.

The first question is whether there exist gender differences in depth of strategic reasoning.[3] In the psychology literature, superior mentalizing ability is typically ascribed to women (Baron-Cohen, 2002). But no study has attempted to study whether this translates into higher levels of strategic sophistication. The second question is whether such gender differences (if any) are mediated by beliefs about the relative strategic sophistication of men and women. Gender stereotypes affect daily behavior 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). Stereotypes might also influence depth of strategic reasoning. We investigate whether gender salience and changes in the gender composition of the group of players alter observed strategic sophistication. Our third question relates to the endogeneity of depth of reasoning to incentives. We study whether males and females respond differently to the presence of monetary incentives. Alaoui and Penta (2016) show that higher incentives induce deeper strategic reasoning. But higher 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 behavior.

We explore these questions in the p-beauty contest/guessing game (Nagel, 1995). This game is well suited for our purposes for a number of reasons. First, it is competitive. Players must predict the average response of others in order to win a prize. Incentives are easy to adjust by changing its monetary value. In addition, beliefs about the sophistication of others are extremely important, as highlighted by models of k-level thinking. Finally, the game 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.[4,5]

[3] This is definitely a thorny issue, more so since the "Summers affair" in 2005. We believe that our understanding of strategic sophistication is better served by tackling such question rather than by ignoring it altogether.

[4] 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).

[5] 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.

We present ten years worth of experimental data from two studies run in two countries and encompassing over a thousand individuals. Study 1 is a large classroom experiment. Students in six different cohorts played the guessing game. In half of these cohorts, no monetary prize was given to the winner(s), whereas a monetary prize was awarded in the other half. We find substantial gender differences in the zero-stakes treatment. Female subjects display lower strategic sophistication than males. However, no gender differences exist when a monetary prize is awarded because there is a significant shift down in females responses across treatments. Males do not react significantly to the presence of monetary incentives.

Study 2 is a laboratory experiment where we manipulate gender priming and gender composition. This design allows us to compare the effect on strategic sophistication of gender priming and of changes in gender composition. Our findings corroborate the result in Study 1, namely that females display lower strategic sophistication than males when incentives are absent and that no gender differences exist when a monetary prize is at stake. Gender differences re-emerge in the opposite direction when we prime gender: Females display higher strategic sophistication than males. Changes in gender composition only affect a subset of males; they decrease their responses when playing in mixed gender groups compared to single gender groups.

We explore the reasons behind these results by analyzing the responses to a questionnaire administered to participants in Study 2 at the end of each session and to non-participants of similar characteristics. We find that beliefs about the relative advantage of males and females have a significant effect on behavior. Males who believe females are better in the game display lower strategic sophistication 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, gender priming has the effect of increasing strategic sophistication among women who believe females are better in the game. Our gender manipulation seems to activate this perception and boost these women's depth of strategic reasoning. We conclude that the overall positive effect of gender priming on the sophistication of women and of mixed gender groups on the sophistication of men is due to females perceiving themselves and being perceived as superior in the game.

The rests of the paper proceeds as follows: Section 2 reviews the related literature. Sections 3 and 4 present the results of Study 1 and 2 respectively. We analyze the responses to questionnaires in Section 5. Section 6 performs a robustness check with an alternative measure of strategic sophistication. In Section 7 we conclude and discuss further the relevance of our results.

2 Related literature

To the best of our knowledge, ours is the first experimental study analyzing the existence of gender differences in depth of reasoning, beliefs and sensitivity to incentives in games. The reason behind this might be a genuine lack of differences, but also a conscious choice of researchers due either to ideological reasons or to the potential controversy of the topic. Very few experimental studies report evidence of gender differences in observed strategic sophistication as a by-product of their design either. Camerer et al. (2004) report in their Table 2 results for a beauty contest in same gender groups, but they only show summary statistics. Burnham et al. (2009) find no gender differences in choices in the beauty contest. This is consistent with the results we obtain when gender is not primed and monetary incentives are given. Östling et al. (2011) and Arad and Rubinstein (2012)[6] report that females display slightly lower strategic sophistication in the Lowest Unique Positive Integer (LUPI) game and in the Colonel Blotto games respectively. Let us reiterate that 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. Behavior 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 depth of strategic reasoning 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.[7] Agranov et al. (2012) find that undergraduate students display higher strategic sophistication when playing the guessing game against graduate students than against computers. Georganas et al. (2015) find a similar result in the undercutting game.

We are aware of only one experimental study relating strategic sophistication to incentives. Alaoui and Penta (2015) find that subjects engage

[6] Personal communication with the authors.

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

in more rounds of reasoning when the prize from outguessing the opponent increases.[8] However, these authors do not explore gender differences in the response to higher stakes. Fryer et al. (2008) find that 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. Azmat et al. (2015) find that the gender performance-gap in exams in favor of female high school students vanishes as stakes increase. This is in contrast with what we find in our two studies, but this might be due to the strategic nature of our experiment. In line with our findings, Frick (2011) employs data from professional distance running competitions and finds that differences in the competitiveness between female and male races are significantly smaller in races where higher prizes or more prestige are 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 in Economics which studies 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 display higher strategic sophistication and to outperform males.

3 Study 1: Beauty in the classroom

Participants in this study were six cohorts of undergraduate students taking an Intermediate Microeconomics course at the University of Edinburgh between 2005 and 2010. As part of the course, students had to fill an online problem set containing several game-theoretic questions implemented via the website *Games and Behavior* developed by Ariel Rubinstein and

[8] 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.

Eli Zvuluny[9]. Cohorts were large, ranging between 116 and 170 students. Completing the problem set was compulsory and liable to a small mark penalty. Response rates were 91.83% on average. Students had no previous instructions in game theory before answering the questions and they had a diverse background both by nationality and by major of study.[10] In total, 792 students took part; 480 of them were male and 312 were female.[11]

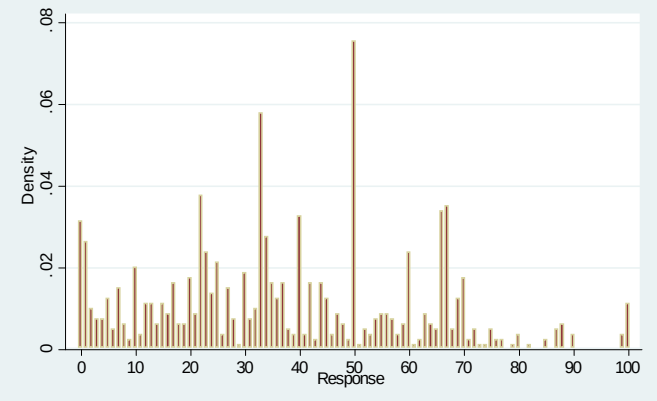

Figure 1:Histogram of responses in Study 1.

One of the questions in the problem set was a beauty contest. Students had to guess two-thirds of the average of all responses of students in the class.[12] Figure 1 contains the histogram of responses for the entire sample. The graph shows the typical pattern of responses in experimental beauty contests (Bosch-Domenech et al., 2002): A 3% or participants responded zero, the Nash equilibrium prediction. The spikes of frequency at 50, 33 and 22, according to the theory of k-level thinking, correspond to individuals with sophistication of level 0, 1 and 2 respectively.

[9] Available at http://gametheory.tau.ac.il/.

[10] Under the Scottish university system, students have the option of taking courses outside their major during their first two years.

[11] 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.

[12] 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!)."

However, this histogram masks important heterogeneity. We ran two different treatments with three cohorts each: 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 in the class, the prize was divided among them. 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 prize* treatment, such non-monetary reward was not made explicit.[13]

Table 1 shows the aggregate results for the two treatments in Study 1, and compares them with the aggregate results of other experimental beauty contests. The studies in italics correspond to subject pools composed by non-students.[14] The first clear thing to observe is that the mean and median responses for the *Prize* treatment are in line with those in previous experiments. We can then safely conclude that despite being implemented online, this treatment is comparable to other experiments.

	Mean	Median	Std dev	Group size
Study 1 - Prize	**36.1**	**33**	**23**	**110-170**
Study 1 - No prize	**39.2**	**37**	**23.3**	**103-156**
Nagel (1995)	37.2	33	20	14-16
Ho et al. (1998)	38.9	NA	24.7	7
Camerer (2003)	32.5	NA	18.6	20-32
Kovalchik et al. (2005)	37	NA	17.5	33
Kocher and Sutter (2005)	34.9	32	NA	17
Buhren and Bjorn (2010)	32.1	29.6	22.2	6,112
Agranov et al. (2012)	36.4	33	21	8

Table 1: Aggregate results in Study 1 and in other experimental beauty contests.

The second observation is that the *No prize* treatment shows the highest average and median responses of all studies reported in Table 1. It is natural to expect that the lack of monetary prizes should induce students to think less about the game. From now on, and as it is customary in the literature, we will associate lower responses to deeper strategic reasoning. Under this assumption, students in the *No prize* treatment should respond higher

[13] This does not rule out that students could seek prestige or status among their closer peers by winning.

[14] Camerer (2003) uses CEOs; Kovalchik et al. (2005) 80 year olds; Buhren and Bjorn (2010) employs chess players, from amateurs to Grand masters.

numbers than in the *Prize* treatment. The data supports this prediction. The distribution of responses under *Prize* differs from the distribution under *No prize* (Mann-Whitney, $p = 0.041$; Median test, $p = 0.047$). If lower responses are associated to higher strategic effort, we should also observe the accumulated distribution of responses for the *No prize* treatment to first order stochastically dominate the accumulated distribution of responses for the *Prize* treatment. Figure 2 suggests this is the case and the Kolmogorov-Smirnov test of first stochastic dominance corroborates this ($p = 0.032$).[15] This evidence supports that higher incentives induced students to engage in deeper strategic reasoning.

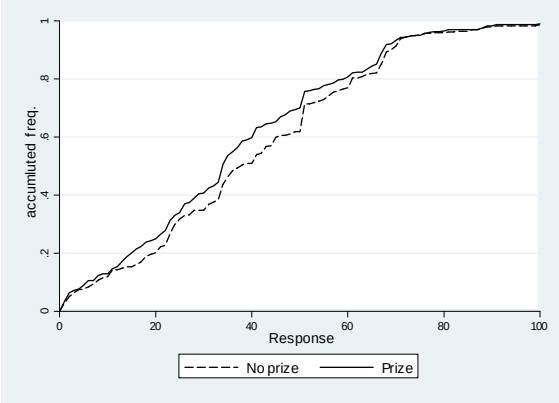

Figure 2: Cumulative distributions of responses by treatment.

A closer look at the data shows however that male and female students responded very differently to the presence of a monetary prize. Table 2 shows the aggregate results by gender in the two treatments. The striking result there is that average and median responses for females in the *No prize* treatment are much higher than any other. Another interesting observation is that males' responses do not differ much across the two treatments.

[15] This test is based on the significance of the largest positive difference between two CDFs. The outcome of the test is that distribution f first order stochastically dominates g 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 difference in the favor of the dominating distribution.

	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 2: Aggregate results by gender and treatment.

Result 1 The distribution of responses of females differs between *Prize* and *No prize* treatments (Mann-Whitney, $p = 0.026$; Median test, $p = 0.054$), and differs between males and females in the *No prize* treatment (Mann-Whitney, $p = 0.049$; Median test, $p = 0.029$).

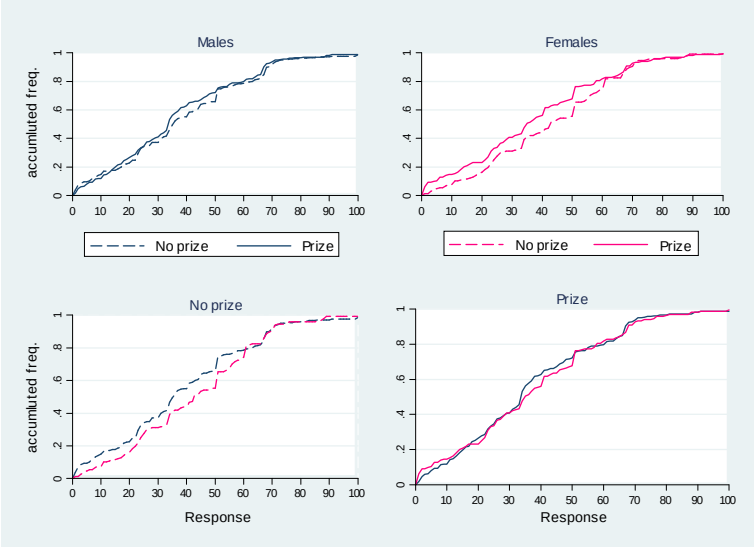

Figure 3: Cumulative distributions of responses by gender and treatment.

Figure 3 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 is depicted with dashed lines. We use the stereotypical colours blue and pink for males and females respectively. The graphs show clearly 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. In addition to the Kolmogorov-Smirnov test, we will also employ throughout the paper the test of stochastic dominance introduced by Davidson and

Duclos (2000).[16] This tests allows us to associate stochastic dominance to a particular range of responses and, hence, to a certain degree of strategic sophistication. The Davidson-Duclos test yields that the distribution of female responses under the *No prize* treatment first order stochastically dominates the distribution under the *Prize* treatment. In addition, the distribution of male responses is not statistically different across the two treatments. Although a higher proportion of males responded numbers between 30 and 50 compared to the *No prize* treatment, the difference is too small to be significant (see Table A1 in Appendix A).

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 A1 in Appendix A). Dominance comes from responses between 0 and 18 and between 35 and 54. According to the k-level theory, these are the responses roughly corresponding to very sophisticated (level-3 and higher) and relatively unsophisticated subjects (level-0 and 1) respectively. More females seem to populate the medium sophistication range of responses (between 19 and 34) and the quite unsophisticated range (55 and above), although no gender differences exist in the proportion of completely irrational responses (68 and above).

The lower right panel shows a very different picture: when a monetary prize is given, gender differences become insignificant (see fourth column of Table A1 in Appendix A). Females show deeper levels of reasoning, as suggested by their lower responses, when a monetary prize is at stake. The extent of this reaction is such that gender differences vanished when a monetary prize was at stake. This might explain why Burnham et al. (2009) find no gender differences in entries in the beauty contest. On the other hand, males do not respond significantly to monetary incentives. This would suggest 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).

[16] 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.

But the fact that males do not react to the presence of monetary prize also suggests that males may consider that a non-monetary prize is at stake in the *No prize* treatment. This 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. 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.

4 Study 2: Beauty in the lab

The aim of our second experimental study was to check whether strategic sophistication might be affected by gender priming and changes in the gender composition of the set of opponents. This study was conducted with four different cohorts of undergraduate students at the University of Barcelona between 2012 and 2015. We made sure that subjects were recruited without they noticing that the experiment related to gender.[17] A total of 240 subjects participated in the study. This sample was more homogeneous than the sample in Study 1. Virtually all subjects were Spanish and all of them majored in the School of Economics and Business.

All participants played in gender-balanced groups of 24 subjects except in one of the treatments, as specified below. 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.

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. During this phase, 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

[17] For showing up, subjects received three euros

.

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 twofold: first to help subjects to familiarize with the beauty contest and second, to replicate the results from Study 1 on the existence of gender differences in behavior in the absence of incentives.

In the second phase we introduced incentives and administered two treatments, the *Priming* treatment (n=144) and the *No priming* treatment (n=96). In the second phase of the sessions pertaining to 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 second phase of the sessions pertaining to 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.

The purpose of this manipulation was to prime gender. Gender was also made salient by distributing pink booklets to female subjects and blue booklets to male subjects. In addition, the gender of the pair of instructors present in each room matched the gender composition of the subjects in it. This means that in the MG round there was one female and one male instructor in each room, and in the SG round, the gender of the two instructors in the room coincided with the gender of the group.

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. At the end of the session, participants filled up a short questionnaire aimed to elicit their views about the behavior of males and females in the game and their relative strategic sophistication. We analyze the responses to this questionnaire in Section 5.

4.1 Results within rounds

4.1.1 First phase: no incentives, no gender priming

Table 3 depicts the aggregate results for the round of the first phase with $p = \frac{2}{3}$. Recall that in that phase, there were no incentives and no gender priming. 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 3: Aggregate results by gender in the No incentives round.

In addition, the distribution of female responses first order stochastically dominates the distribution of male responses. The Davidson-Duclos test reports that this gender difference emerges in the interval of responses from 42 to 50 (see first column of Table A2 in Appendix A), which corresponds to a low level of strategic sophistication. Figure 4 illustrates this result.

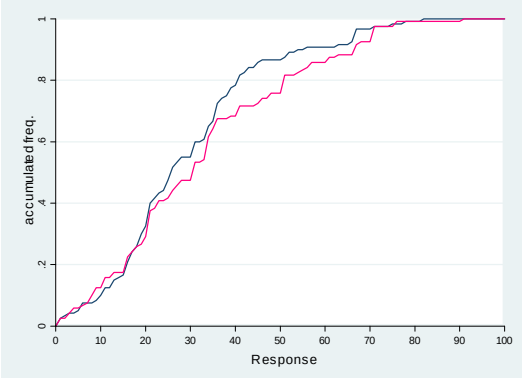

Figure 4: Cumulative distributions of responses by gender in the first phase with p=2/3.

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 A2 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, $p = 0.049$). Hence any feedback-free learning (Weber, 2003) that might take place across the phase does not seem to help to reduce gender differences in strategic sophistication.

To highlight further these strong gender differences, we classify subjects according to their sophistication across the 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 is considered to be of low (high) sophistication 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 also indirectly incorporates beliefs about the sophistication of the opponent: Coricelli and Nagel (2009) show through fMRI that subjects classified as highly sophisticated display a more intense activation in areas of the brain associated with theory of mind than low sophisticated subjects.

As Table 4 shows, 80.7 % of the classified individuals in the sample are low sophisticated; from these, 58.9% are female. The percentages of High and Low sophistication subjects out of the whole pool (19.2% and 80.7% respectively) are very similar to the ones obtained in previous studies.[18] However, these figures mask important gender differences. The small fraction of females who exhibit high strategic sophistication (9.2%) stands out. It is significantly different from the proportion of highly sophisticated males (proportions test, $p < 0.001$).

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

Table 4: Sophistication by gender (first phase).

[18] Coricelli and Nagel (2009) obtain 35% and 50% respectively (n=20). Brañas-Garza, Garcıa-Muñoz and Hernan (2012) obtain 13% and 78% (n=191).

All this evidence reinforces the result we obtained in Study 1: When financial incentives are absent, females display lower levels of strategic sophistication than males.

4.1.2 Second phase: incentives and no gender priming

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

In the control treatment, the gender composition of the group was always balanced. Table 5 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 5: Aggregate results by gender in the *No priming* treatment.

Males display higher strategic sophistication 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 A2 in Appendix A). We again corroborate the results obtained in Study 1: Gender differences in strategic sophistication disappear when monetary incentives are present.

4.1.3 Gender priming

As mentioned above, we primed gender by manipulating the gender composition of the group and the colour of the instruction booklets. Thus, we are able to analyse gender differences in 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 only difference between the second phase of the *No priming* treatment and the MG round is that gender was made salient in the latter. In contrast

with previous results, Table 6 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 6: Aggregate results by gender in the *Priming* MG round.

Females display higher strategic sophistication than males in this round. The comparison of the distributions of responses across genders in the MG round shows indeed that they 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 A2 in Appendix A) dominance tests provide a clear ordering between them.

Result 2.1 The distribution of male responses first order stochastically dominates the distribution of female responses in the MG round of the *Priming* treatment.

Figure 5 illustrates this result. The Davidson-Duclos dominance test establishes that there is a higher number of females than males who choose responses in the interval between 12 and 24. This suggests that more females display relatively higher levels of sophistication than males.

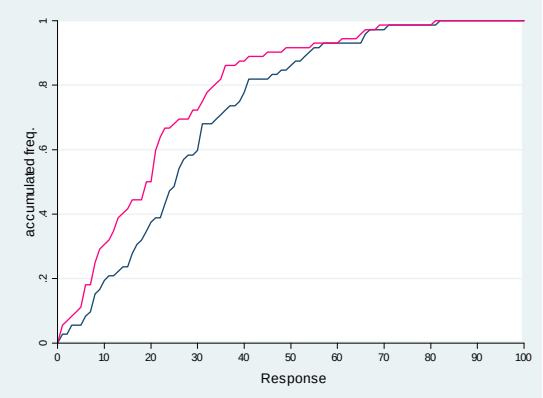

Figure 5: Cumulative distributions of responses by gender in the MG round.

The SG round In this round of the *Priming* treatment, participants played against opponents of their same gender. Table 7 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 response.

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

Table 7: 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 6 illustrates. The interval of significant dominance according to the Davidson-Duclos tests ranges from 14 to 27 and from 44 to 74 (see fourth column of Table A2 in Appendix A).

Result 2.2 The distribution of male responses first order stochastically dominates the distribution of female responses in the SG round.

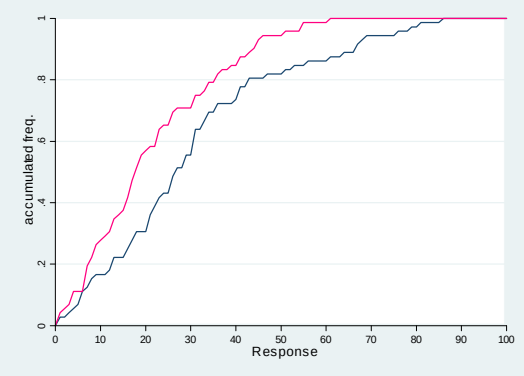

Figure 6: Cumulative distributions of responses by gender in the SG round.

To summarize, the main findings of the analysis by round are: 1) We confirm that males show higher levels of sophistication than females in the absence of incentives. 2) We confirm that gender differences disappear when there are monetary incentives. And 3), when gender is primed, females display higher levels of strategic sophistication than males.

4.2 Results across rounds

4.2.1 *Priming* versus *No priming*

Let us now exploit our design in order to 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 expect gender priming to make beliefs about the relative strategic sophistication of males and females salient. If this is the case, responses across the two treatments should change. But how? If there exists the stereotype that a 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 mentalizing 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 display lower levels of sophistication, and 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 a gender is superior to the other, we would expect members of that group to display higher levels of sophistication, and thus lower entries, in the *Priming* treatment compared to the *No priming* one.

Table 8 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 behavior 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 8: 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.

Result 3 The distributions of female responses under *No priming* first order stochastically dominates the one under *Priming*. The opposite holds for males' responses.

This dominance test (see table A3 in Appendix A) establishes that under *Priming* fewer males display high levels of sophistication (entries between 10 and 18) whereas fewer females display low and medium levels of sophistication (entries between 20 and 24 and between 32 and 42). Figure 7 corroborates this. In summary, females respond strongly to gender priming by increasing their level of sophistication whereas males react to a lesser extent and display slightly lower strategic sophistication when gender is salient.

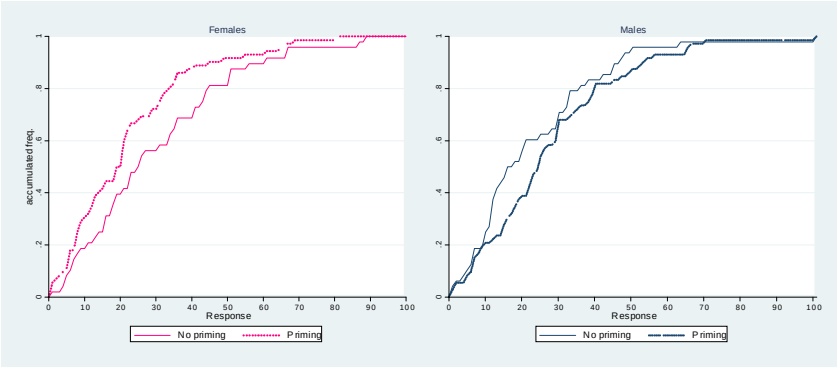

Figure 7: Cumulative distributions of responses by gender and treatment in the gender-balanced rounds.

Result 3 suggests that strategic interactions might be a context where females might perceive themselves and be perceived to be superior. 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 mentalizing ability (Baron-Cohen, 2002; Krach et al., 2009). Our gender priming seems to activate this stereotype, boosting women's sophistication and reducing males's depth of strategic reasoning. We investigate this idea in Section 5 when analyzing subjects' beliefs about the relative strategic sophistication of men and women.

4.2.2 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), we conjecture that if an individual believes that a change in the gender composition shifts up (down)

the distribution of levels of sophistication in the group, he/she will exert more (less) effort and his/her entry will decrease (increase).

Table 9 shows that the average and median responses of both sexes do not significantly differ across the SG and the MG rounds and that the gender differences observed in the MG round persist in the SG round.

	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 9: 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 $p = 1.000$).

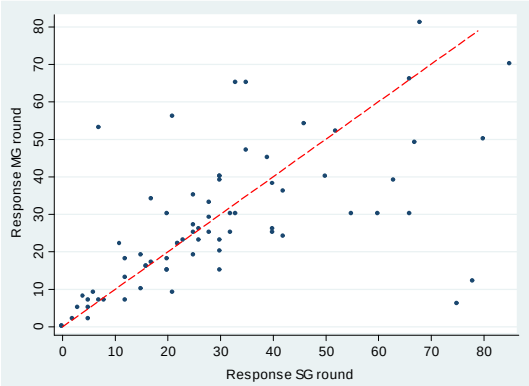

Figure 8: Males' responses in the MG and SG rounds.

A more detailed picture emerges from Figure 8, the scatterplot of males' responses 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, especially those with higher entries in the SG round, decrease their response in the MG 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 below the median. This reduction of responses in the MG round is consistent with males increasing their depth of reasoning compared to the SG round because they perceive women to be superior in the game. Alternatively, it might be that these males expect level-0 female players to randomize on 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 the session.

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

5 Beliefs and stereotypes

In this section, we explore responses to the questionnaire we administered to participants in our Study 2. Participants answered these questions at the end of the session, before any feedback was provided. The aim of this questionnaire was twofold. First, to investigate whether priming was effective in activating gender stereotypes. Second, given that depth of strategic reasoning depends on the perceived sophistication of others (Georganas et al., 2015; Agranov et al., 2012), to explore whether beliefs about the relative strategic sophistication of men and women influence behavior.

We focus on the responses to two questions. The first question is "When $p = \frac{2}{3}$, which sex responds higher numbers?" (Q1) and the second is "Which sex is better at this game?" (Q2). These two questions capture different factors which might be important to understand subjects' behavior. Q1 is designed to obtain information on beliefs about the behavior of others and Q2 is designed to elicit perceptions about the relative sophistication of males and females. Although we have assumed in our analysis, that lower entries are associated with higher strategic sophistication, this might not be true in the mind of subjects. In addition, note that this association is based on the assumption, customary in the literature, that level-0 behavior is a uniform distribution on the set of strategies. But participants engaging in deeper strategic reasoning might have stereotypes on the random behavior of males and females which depart from that assumption.[19]

[19] 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."

Answers were free-text, so we coded them in four options, "Males", "Females", "No difference" and "Don't know". Responses display a medium to strong correlation across questions (Contingency coefficient, 0.443; Cramér's V, 0.349). This implies that participants seemed to understand the basics of the game and 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) and 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.

5.1 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 displayed higher strategic sophistication 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.

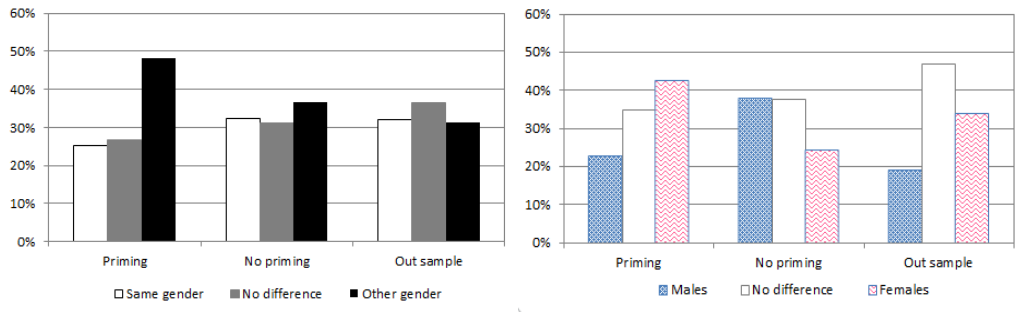

Figure 9: Responses by gender to Q1 and Q2 (for females) by sample.

The left panel of Figure 9 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, $p = 0.0202$). 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.

5.2 Gender bias

The next question is whether answers to the questionnaire can help explain observed behavior. The first issue we tackle relates to the association between behavior and beliefs about which gender has a relative advantage in the guessing game. Our aim is to study whether gender stereotypes, expressed in responses to Q2, might be related to strategic sophistication.

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 male subjects. Males subjects who believe that their own gender has a relative advantage in the game, choose lower entries. Of course, we cannot establish a casual relationship between perceptions and behavior. It might be that males who respond lower numbers conclude that their gender is better in the game (although there was no feedback until the very end of the session). We come back to this 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 behavior 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 Take 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 3. The belief of women on their own superiority in the game has an effect on their behavior only when gender is made salient. The difference in median responses is striking: 40 for these 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 behavior. It cannot be the case that behavior 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 depth of reasoning of these participants.

Interestingly, gender priming has no significant effect on females who answer that males are better in the game or that no gender differences exist. This might be due to the absence of a negative stereotype against women in the game. 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 gender priming has a neutral to positive effect on our female subjects.

5.3 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. First we want to establish that, despite not being incentivised, responses to Q1 can help to explain differences in behavior across these 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 (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 4. 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 behavior should be those who believe that females are better at the game. According to the second hypothesis, these males should answer to Q1 that men tend to pick higher numbers.

Males who believe that men respond higher numbers than females change their behavior between the SG and the MG rounds (Wilcoxon sign-rank, $p = 0.050$). This would lend support to the second hypothesis. 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 their own gender responds higher numbers. On the other hand, notice that Result 4.1 extends to the SG round: Males who answer that females are better at the game respond higher numbers in the SG round than those who believe the opposite (Mann-Whitney, $p = 0.024$). Since males with higher responses in the SG round are the ones who decrease their entries when playing the MG round, this suggests that these males might display higher strategic sophistication in mixed gender groups because they believe that females play better than males. In one of the few studies looking at sex differences in mentalizing, Krach et al. (2009) use fMRI on subjects playing a Prisoner's dilemma'. The brain activation patterns they observe are consistent with men compensating their weaker mentalizing abilities by increased effort. The change in the median response across the two rounds for males who respond to Q2 that females are better at the game is also consistent with this explanation. The median is 28 in the MG round and 33 in the SG round. Unfortunately, our sample size does not allow for a more detailed analysis which can discriminate further between these two hypotheses.

6 Robustness check: Accuracy

A key assumption in the analysis so far 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. So as a robustness check, we use 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*. The analysis below shows that our main results hold when we use the quadratic distance to the winning response as a measure of strategic sophistication.

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 10 depicts the corresponding kernel densities. Female players (flatter curve) are clearly less accurate than male players. This confirms our results in Section 4.1.1.

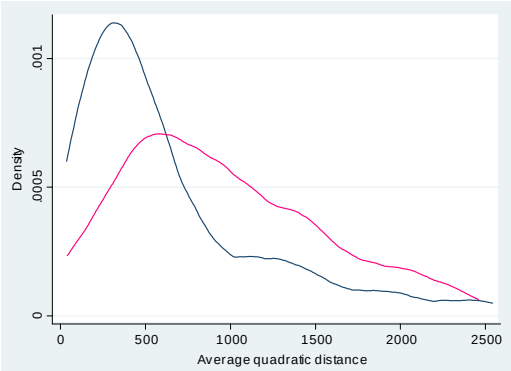

Figure 10: Average quadratic distance to winning response in the first phase by gender.

In Figure 11 we extend the analysis to the classification by levels of sophistication introduced in Section 4.1.1. 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.

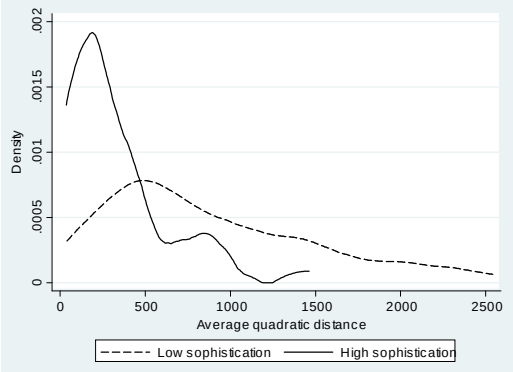

Figure 11: Average quadratic distance to winning response in the first phase by level of sophistication.

Next we compare the accuracy of responses in the gender-balanced rounds of the *Priming* and *No priming* treatments. Recall that Result 3 stated that gender priming had the effect of significantly lowering the entries of female participants and of increasing males' responses. The strength of this effect was such that females displayed higher strategic sophistication than males in the *Priming* treatment. These results remain, albeit less sharply, when looking at accuracy. The upper panels of Figure 12 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. Hence, we cannot conclude that gender priming makes males less accurate despite the fact that their answers are higher in average when gender is made salient.

The lower panels of Figure 12 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, $p = 0.046$).[20] This corroborates our results in Section 4.1.3, namely that gender priming makes females more sophisticated than males.

[20] A similar effect is also observed in the same gender round of the *Priming* treatment (Mann-Whitney, $p = 0.034$; Median test, $p = 0.067$). Kernel densities for the quadratic distance in this round can be found in Figure A3 in Appendix A.

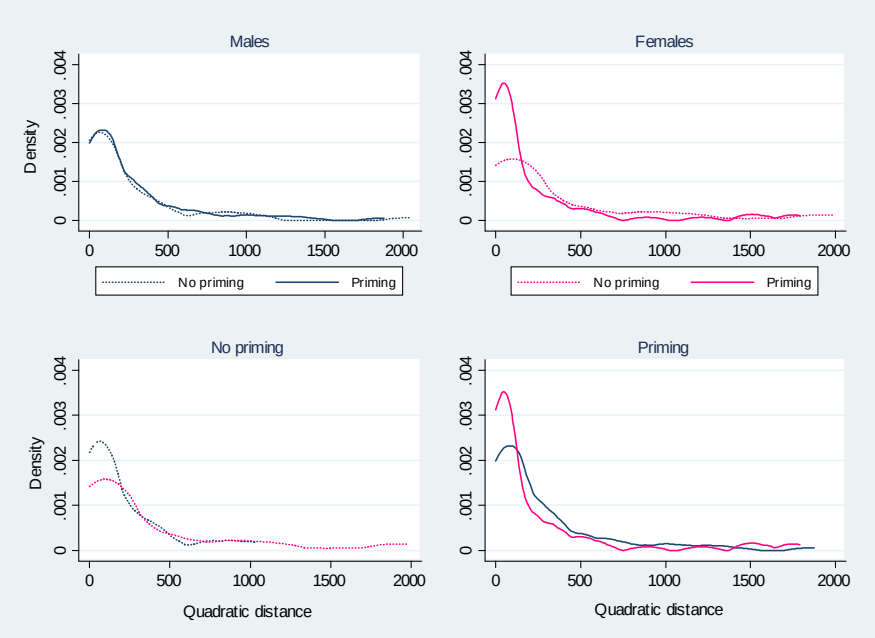

Figure 12: Quadratic distance to winning response by gender and treatment.

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 (see Figure A4 in Appendix A).

7 Discussion and conclusions

In this paper, we explored the existence and endogeneity of gender differences in observed strategic sophistication. Depth of strategic reasoning might depend on individual cognitive abilities, beliefs about the sophistication of others and the size of incentives. Gender might be relevant to all these three factors. We used the beauty contest game as experimental design. We chose this game because it is competitive, because incentives can be easily manipulated and because beliefs about the sophistication of others are important in it.

We reported results from two studies encompassing ten years of sessions and over one thousand individuals. Study 1 was a large classroom experiment. The main result of this study is that gender differences in behavior exist only when no monetary prize is awarded. We interpret this result in line

with Croson and Gneezy (2009): Monetary incentives frame the interaction as competitive. Females, being more sensitive to contextual information, think more deeply about the game and gender differences disappear. An alternative, but not incompatible, explanation is that males derive utility from winning regardless of whether a monetary prize is at stake.

Study 2 was a laboratory experiment where we manipulated gender priming and gender composition. Results of this study corroborate those in Study 1: Gender differences in strategic sophistication disappear when incentives are introduced. In addition, females react very strongly to gender priming by increasing their level of sophistication. Males react to a lesser extent and in the opposite direction. Gender differences reappear when gender is made salient but they are favourable to women. The effect of changes in the gender composition of the group was smaller and only applied to a subset of males who display higher sophistication in mixed gender groups compared to single gender groups.

To understand the forces driving these results, we explored the responses to a questionnaire we administered to our participants. Responses to these questionnaires show that females who react to gender priming are those who believe that females are better in the game. We conclude that the combination of gender salience and the belief that women are better in the game boosted the depth of reasoning of these participants. Males who answer that females are better at the game display lower strategic sophistication suggesting that these males might be experiencing stereotype threat.

Indirect evidence (e.g. Burnham et al., 2009) seemed to suggest that no gender differences existed in the beauty contest. We observe differences only when we manipulate incentives and gender 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. Nevertheless, we are aware that subjects' characteristics could correlate with gender, e.g. major of study in undergraduate populations, and thus create spurious gender differences.[21] Our subject pool in Study 2 was relatively homogeneous. Our participants were students of Economics or Business, of very similar age and ethnic and cultural background, so we are relatively free from this problem.

In sum, our results show that strategic sophistication, especially for females, is endogenous to incentives and gender priming. These results confirm previous findings in the experimental literature highlighting the role of beliefs about the strategic sophistication of other players (Agranov et al.,

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

2012; Georganas et al., 2015) and incentives (Alaoui and Penta, 2015) on depth of strategic reasoning. Iriberri and Rey-Biel (2016) find that just mentioning gender is enough to reduce women's performance in real-effort tasks perceived as male-biased. In contrast, when gender is made salient in the guessing game, females increase their depth of reasoning and display higher strategic sophistication than males. Our conjecture is that this positive effect of gender priming is due to women perceiving themselves and being perceived as superior in strategic settings. This is in line with studies showing that women's performance is higher in tasks perceived to be female-biased (e.g. Guenther et al., 2010). Results from a questionnaire administered to a set of comparable non-participants with no previous knowledge of the guessing game provide additional evidence of this stereotype.

The present paper is one of the few where women are observed 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 gender priming is in even sharper 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, especially if confidence on their own superiority is widespread among women. 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 sophistication that we uncover in the guessing game remain in other games where standard equilibrium predictions are more transparent and where subjects may resort to simple rules of play to a lesser extent.

References

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

[2] Alaoui, L, and Penta, A. 2015. Endogenous Depth of Reasoning, forthcoming in *Review of Economic Studies*.

[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. 2015. Gender Differences in Response to Big Stakes, forthcoming *Journal of the European Economic Association*.

[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, 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 Behavior 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, 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 Behavior and Organization*, 72(1): 171–175.

[12] Camerer, C F. 2003. *Behavioral Game Theory: Experiments on Strategic Interaction*, Princeton. NJ: Princeton University Press.

[13] Camerer, C F, Ho, T-H, and Chong, J K. 2004. A Cognitive Hierarchy Model of Games, *Quarterly Journal of Economics*, 119(3): 861-898.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

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

[21] Danaher K, and Crandall, C S. 2008. Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology*, 38(6): 1639-1655.

[22] 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.

[23] Fragiadakis, D E, Knoepfle, D T, and Niederle, M. 2013. Identifying Predictable Players: Relating Behavioral Types and Subjects with Deterministic Rules," unpublished manuscript.

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

[25] 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.

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

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

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

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

[30] 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.

[31] 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.

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

[33] 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.

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

[35] 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 Behavior and Organization*, 58: 79–94.

[36] 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.

[37] 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.

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

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

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

Appendix A: Additional Tables and Figures

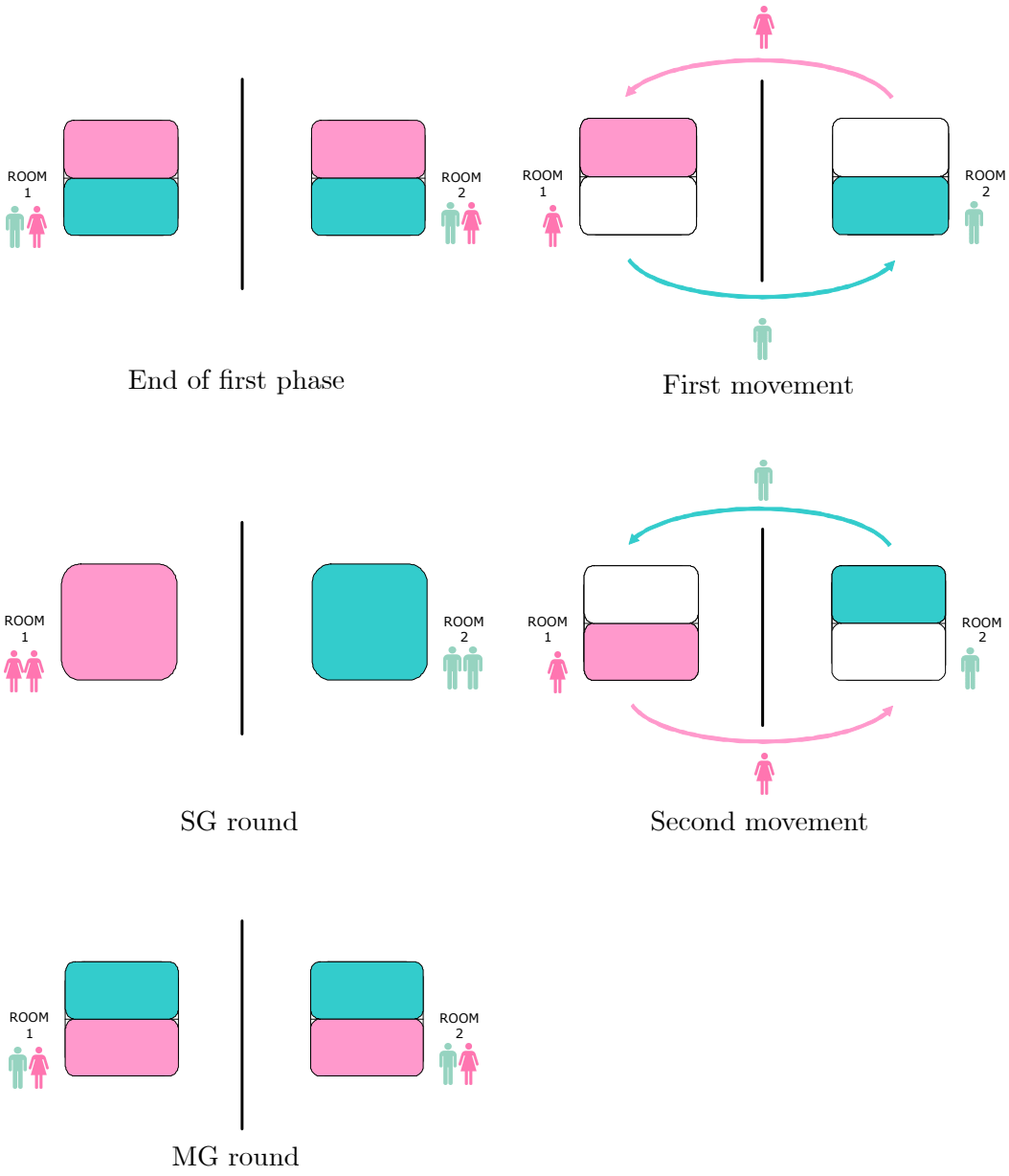

Figure A1: Moves of participants in the SG-MG order of the *Priming* treatment.

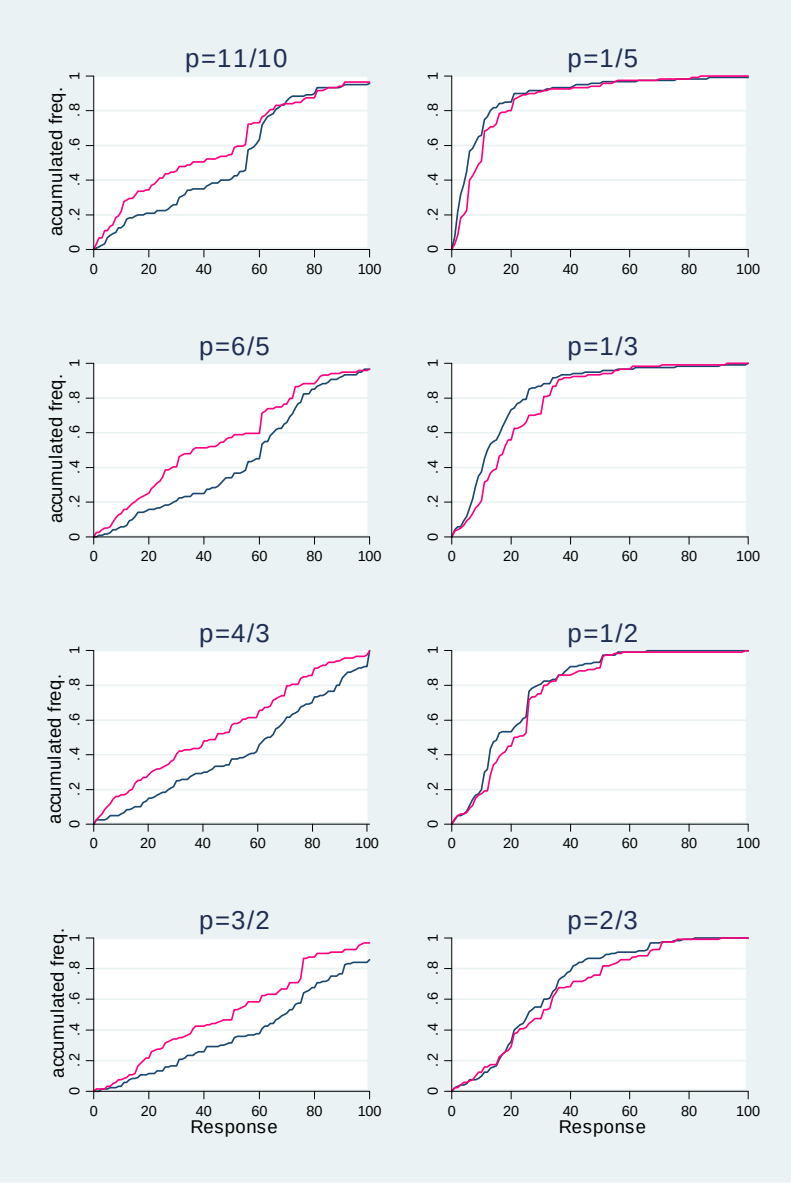

Figure A2: Cumulative distributions of responses by gender to phase 1 rounds.

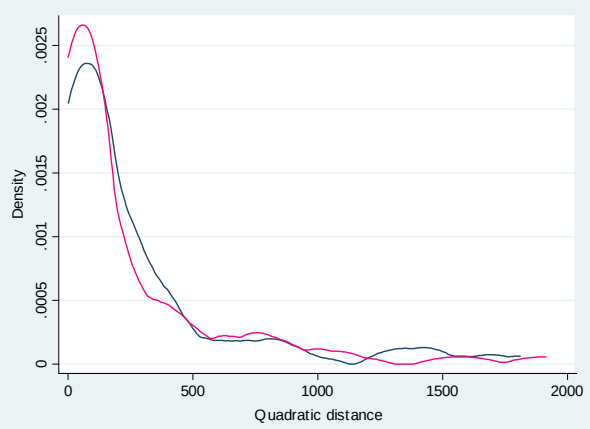

Figure A3: Quadratic distance to the winning response by gender in the SG round.

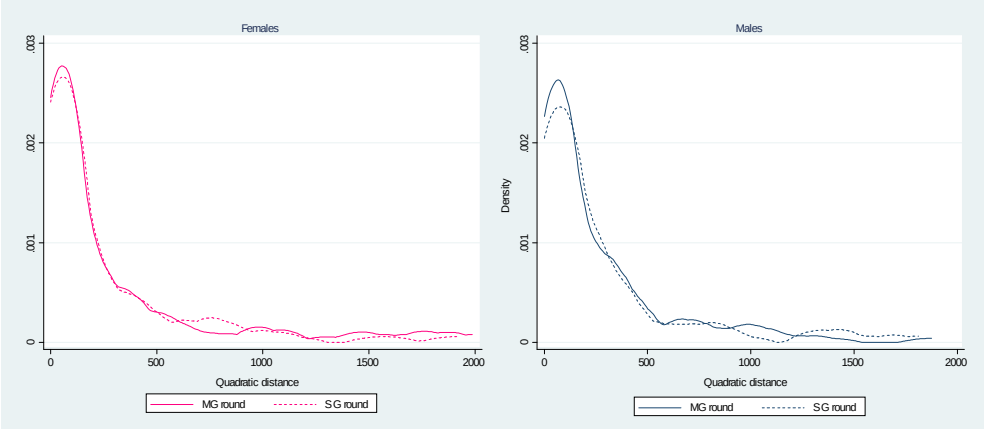

Figure A4: Quadratic distance to the winning response by gender across rounds of the *Priming* treatment.

	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 A1: Davidson-Duclos (DD) test t-statistics for Study 1.

	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 A2: Female-male DD test t-statistics per round of Study 2.

	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 A3: DD test t-statistics for round comparisons in Study 2.

Appendix B: 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.