CESifo Working Papers

www.cesifo.org/wp

Gender Differences in Strategic Reasoning

Marcus Dittrich
Kristina Leipold

CESifo Working Paper No. 4763
Category 13: Behavioural Economics
April 2014

CESifo *Working Paper No. 4763*

Gender Differences in Strategic Reasoning

Abstract

This paper studies gender differences in strategic situations. In two experimental guessing games - the beauty contest and the 11-20 money request game - we analyze the depth of strategic reasoning of women and men. We use unique data from an internet experiment with more than 1,000 participants. We find that men, on average, perform more steps of reasoning than women. Our results also suggest that women behave more consistently across both games.

JEL-Code: C720, C900, J160.

Keywords: gender differences, strategic reasoning, level-*k*, beauty contest, 11-20 money request game.

Marcus Dittrich
Chemnitz University of Technology
Department of Economics and Business Administration
Germany – 09107 Chemnitz
marcus.dittrich@wirtschaft.tu-chemnitz.de

Kristina Leipold
Dresden University of Technology
Faculty of Business and Economics
Germany – 01062 Dresden
kristina.leipold@mailbox.tu-dresden.de

March 31, 2014

1 Introduction

Gender differences can be observed in a number of different domains, e.g., education, consumption, charitable giving and, perhaps most important, in the labor market (Croson and Gneezy, 2009). Traditional approaches in labor economics focus on differences in the accumulation of human capital and discrimination as the main factors responsible for the persistent gender differences (Bertrand, 2011). However, even after controlling for various measures of human capital, the wages of men and women differ significantly. Recently, alternative explanations have been analyzed in laboratory experiments. These experimental studies report systematic gender differences in social preferences, risk preferences, and behavior in competitive situations (see Croson and Gneezy (2009) for an overview).

This paper provides an alternative channel to explain gender differences. In many real-life situations, decision making depends not only on an individual's own actions but also on the actions of other persons involved. Thus, making the best possible decision requires thinking strategically and taking the decision making of other persons into account. In this paper, we examine whether women and men behave differently in such strategic situations. While gender differences in social preferences are extensively reported in the literature, differences between women and men in strategic reasoning have rarely been studied. Our findings help fill this research gap.

Our analysis is based on two experimental guessing games, the beauty contest (Nagel, 1995) and the money request game (Arad and Rubinstein, 2012). In these games, we are able to observe the participants' level-k reasoning, i.e., the depth of their strategic reasoning. We use data from an internet experiment with a large number of participants.[1] We find that men have a higher depth of reasoning than women. Our results suggest that behavioral differences between women and men might be due not only to social preferences—as discussed in the literature—but also to behavior in strategic situations.

The remainder of the paper is organized as follows. Section 2 provides a brief review of the related literature. Section 3 describes the experimental design and the procedure. Section 4 analyzes the effect of gender on strategic reasoning. Section 5 presents conclusions.

[1] The internet experiment was originally conducted for a research project on the relationship between social preferences and strategic reasoning. In the recent paper, we only analyze the gender-related data regarding strategic reasoning, which has not been analyzed before. Our interest in gender differences was not accidental but rather complements our previous research, in which we examined gender differences in bargaining behavior (Dittrich et al., 2014) and time preferences (Dittrich and Leipold, 2014).

2 Related literature

A considerable number of experimental studies report systematic gender differences in social preferences (Croson and Gneezy, 2009; Eckel and Grossman, 1998, 2001), time preferences (Dittrich and Leipold, 2014), risk preferences (Charness and Gneezy, 2012; Croson and Gneezy, 2009; Dwyer et al., 2002; Jianakoplos and Bernasek, 1998), and bargaining behavior (Dittrich et al., 2014). Some of these studies using dictator games report that women are more inequality averse and less selfish than men (Eckel and Grossman, 1996, 1998; Fehr et al., 2006; Selten and Ockenfels, 1998). Andreoni and Vesterlund (2001) find that women prefer to equalize payoffs and are more generous when giving is expensive while men are more generous when giving is rather cheap. In an ultimatum game, Eckel and Grossman (2001) find no significant gender differences in mean offers. Güth et al. (2007) find in a three-person ultimatum newspaper experiment that women prefer the equal split solution more often than men. García-Gallego and Jaramillo-Gutiérrez (2012) apply an ultimatum game and focus on risk attitudes. Their results suggest that both gender and risk-related effects co-exist, but differences in risk attitudes cannot explain gender effects. Castillo and Cross (2008) find in various ultimatum and dictator games that gender differences are due not only to social preferences, such as altruism, but also to beliefs about others players' strategic behavior. Dittrich et al. (2014) find in an alternating-offer bargaining game that male players are able to obtain better bargaining outcomes than female players.

In summary, the evidence on gender differences in laboratory experiments is mixed (Camerer, 2003). Croson and Gneezy (2009) suppose that these mixed results might be obtained because women are more risk averse than men. A second explanation is the greater context sensitivity of women, i.e., their social preferences vary more strongly over different experimental settings than those of men (Croson and Gneezy, 2009). Moreover, women seem to have a smaller propensity to enter competitive situations (Datta Gupta et al., 2013; Gneezy et al., 2009, 2003; Niederle and Vesterlund, 2007).

In this paper, we argue that differences in strategic behavior between women and men can provide an alternative explanation for the observed gender differences. In many situations, individuals have to think and act strategically, and decision making in those situations often depends on the actions of other individuals. Making optimal decisions thus requires predicting others' behavior and subsequently adapting one's own behavior. Analyzing these situations, standard game theory often assumes that players play equilibrium strategies. However, recent experiments suggest that individuals' initial choices often deviate systematically from equilibrium. An alternative approach to explain observed behavior employs structural non-equilibrium level-k models (Crawford et al., 2013). The level-k model, first proposed by Stahl and Wilson (1995) and

Nagel (1995), assumes that individuals are heterogeneous in their levels of strategic reasoning. A level-0 type behaves non-strategically, while a level-1 type behaves as if he responds best to the belief that the other player is a level-0 type (Arad and Rubinstein, 2012). A level-2 type assumes that all other players are level-1 types or he responds best to a mixture of lower types, and so on. Variations of the level-k model can be found in Camerer et al. (2004) and Costa-Gomes et al. (2001). In most models, level-0 behavior is assumed to be a uniform distribution over the action space. Burchardi and Penczynski (2014) provide a generalized version of the level-k model that allows for non-uniform level-0 behavior and heterogeneous level-0 beliefs. Experimental evidence for the level-k model was proposed by guessing games, such as the beauty contest (Nagel, 1995), normal-form games (Costa-Gomes et al., 2001; Croson and Gneezy, 2009; Stahl and Wilson, 1995), and, more recently, by the 11–20 money request game (Arad and Rubinstein, 2012).[2] For further experimental evidence, see the overview by Crawford et al. (2013).

These studies, however, do not examine whether women and men differ in their strategic reasoning. Our paper aims to fill this gap.

3 Experimental design and procedure

To test gender differences in strategic reasoning, we use two experimental guessing games: the beauty contest (Nagel, 1995) and the money request game (Arad and Rubinstein, 2012). In both games, each player's payoff depends on the number she chooses and the number(s) chosen by the other player(s) and, thus, by the players' depth of strategic reasoning.

3.1 Experimental design

The beauty contest is described in general form by Nagel (1995) as follows:

> *"A large number of players have to state simultaneously a number in the closed interval [0, 100]. The winner is the person whose chosen number is closest to the mean of all chosen numbers multiplied by a parameter p."*

We apply this game structure to our experiment and set $p = 2/3$. Moreover, we restrict the choice to integers instead of decimal numbers. Thus, each participant chooses an

[2] In contrast to the level-k model, Breitmoser (2012) provides experimental evidence that choices are described more adequately as mixtures of quantal response equilibrium and noisy introspection than as level-k mixtures. Weizsäcker (2003) finds that subjects act as if underestimating the response precision of their opponents. This result could be interpreted as a critique of the assumption made in game-theoretic models that subjects are aware of the level of randomness in their opponents' motivations.

integer between 0 and 100. Then, the average of all chosen numbers is calculated. The winner is the person whose number is closest to 2/3 times this average. The payoff to the winner is €50.[3]

Our second game is a slight variation of the 11–20 money request game introduced by Arad and Rubinstein (2012). Arad and Rubinstein (2012) describe their basic version of the game as follows:

> *"You and another player are playing a game in which each player requests an amount of money. The amount must be (an integer) between 11 and 20 shekels. Each player will receive the amount he requests. A player will receive an additional amount of 20 shekels if he asks for exactly one shekel less than the other player."*

We apply this game structure to our experiment, but convert the amount of money into €-cent. A person participating in our experiment requests an amount of money between 10 and 100 €-cent that is divisible by 10. Each person receives the amount requested and receives an additional amount of 100 €-cent if asking for exactly 10 €-cent less than his opponent.

The experimental analysis is based on unique data from an anonymous online experiment. One advantage of conducting this type of experiment is the large number of participants. Moreover, in contrast to our internet experiment, the participants in lab experiments are often not representative of the population at large. The shortcoming of online experiments, however, is that we have less control over the sample than in a controlled lab experiment.

The experiment was conducted in November 2012. Participants were recruited from an online panel of about 90,000 panelists living in Germany by a certified professional research company. 4,291 potential participants were invited via e-mail. Finally, a total of 1,004 subjects (501 females and 503 males) between 18 and 66 years of age took part. The sample was chosen according to certain criteria so that our data is representative of the German population in terms of key characteristics such as gender, age, religious affiliation, and residential region. Tables 1 presents some descriptive statistics of the socio-demographic characteristics of the participants.

The basic procedure of the experiments was as follows. Participants were invited via e-mail by the organizing research company. The e-mail included a personalized hyperlink directing subjects to the web-based experiments. After following the hyperlink, the participants received some general information about the experiment. To avoid possible biases and framing effects, however, we did not provide any information about

[3]Appendix A contains translated versions of the experimental instructions.

		Women	Men	Total
Participants		501	503	1,004
Residential region	West Germany	360 (71.9%)	379 (75.3%)	739 (73.5%)
	East Germany	103 (20.6%)	83 (16.5%)	186 (18.5%)
	Berlin	38 (7.6%)	41 (8.2%)	79 (7.9%)
Religious affiliation	Catholicism	101 (20.2%)	136 (27.0%)	237 (23.6%)
	Protestantism	146 (29.1%)	143 (28.4%)	289 (28.8%)
	Other religion	54 (10.8%)	41 (8.2%)	95 (9.5%)
	No religion	200 (39.9%)	183 (36.4%)	383 (38.2%)
Age	18-24 years	77 (15.4%)	51 (10.1%)	128 (12.7%)
	25-49 years	285 (56.9%)	294 (58.5%)	579 (57.7%)
	50-66 years	139 (27.7%)	158 (31.4%)	297 (29.6%)
	Mean	40.57	42.32	41.45
	Std. dev.	12.71	12.50	12.63
Monthly income (€, after tax)	Mean	2,391.61	2,854.91	2,633.05
	Std. dev.	3,673.36	3,334.28	3,506.34

Table 1: Socio-demographic characteristics of the participants.
93 women and 59 men did not indicate their income, i.e., we have 852 observations for income.

the aim of the study. In particular, the participants did not know that the study is focused on gender differences.

All participants played the beauty contest first and then the money request game. Both games were played once. After having read the instructions, the participants should explain in detail why they choose the respective numbers in both games. The participants were informed that the reasoning is incentivized, i.e., they can receive an additional payment, which depends on how conclusive they explain their reasons. There was no time pressure to give reasons for their decision. After the participants explained their decisions, they stated their numbers chosen. This order allows participants to write down their reasoning while they are thinking about their reasons to choose a number. Moreover, this procedure meets the requirements of protocol analysis known from social psychology and supports the validity of verbal reports (Ericsson, 2002, 2003). Finally, the participants completed a short questionnaire about their gender and other socio-demographic characteristics, such as age, income, residential region, and religious affiliation.

3.2 Classification procedure

There are different approaches to assign a level-k type to a person. Nagel (1995) classifies the level-k types on the basis of the numbers chosen in the guessing game. An alternative method uses multiple choices of one person from one class of game to obtain an overall decision rule that represents the observed behavior (Costa-Gomes and Crawford, 2006; Costa-Gomes et al., 2001; Stahl and Wilson, 1995). Bosch-Domènech et al. (2002) analyze choices and ex post comments of participants. These comments could be made optionally with a time delay after the decision and were not incentivized. Burchardi and Penczynski (2014) use incentivized written accounts of individual reasoning for assigning a level-k type. They apply a team communication protocol which ensures that written accounts are stated at the time of the decision making.

We use a similar approach as Burchardi and Penczynski (2014) and classify participants by their written statements, which are incentivized to elicit the individual reasoning behind the choices. In addition to the beauty contest, we apply the 11–20 money request game (Arad and Rubinstein, 2012). According to Arad and Rubinstein (2012) this simple game triggers level-k reasoning very well, but no possible alternative decision rules. Moreover, the classification of the level-0 type and the higher-order types is clear-cut. Based on the written statements, each participant is assigned a level-k type according to a classification scheme. The classification scheme consists of decision rules for each level in both games. It is based on experimental evidence (Arad and Rubinstein, 2012; Burchardi and Penczynski, 2014; Nagel, 1995) and written statements of a pretest. For the beauty contest, the decision rules reflect the iterated best reply model (Nagel, 1995) and iterated dominance. A person acting non-strategically and following a simple decision rule is a level-0 type (Arad and Rubinstein, 2012). In the money request game, a level-0 type chooses the natural anchor 100 or any other number without realizing the iterative reasoning process. The same holds true for the beauty contest, in which a level-0 type plays a focal strategy or chooses non-strategically any number in the given interval. A typical example is as follows: "I choose 8 because it is my birthday." A level-1 type thinks more strategically and responds best to any level-0 decision. For example, in the money request game, the best response to 100 is 90. In the beauty contest, assuming that the numbers chosen by the other players are uniformly distributed, a level-1 type's best response is 33. Similarly, we classify subjects as level-2 types according to their description. We restrict the classification to a maximum of level-3. In this category, we summarize level-3 and players who comprehend the whole iteration process.[4]

The most important difference between our approach and traditional assignments

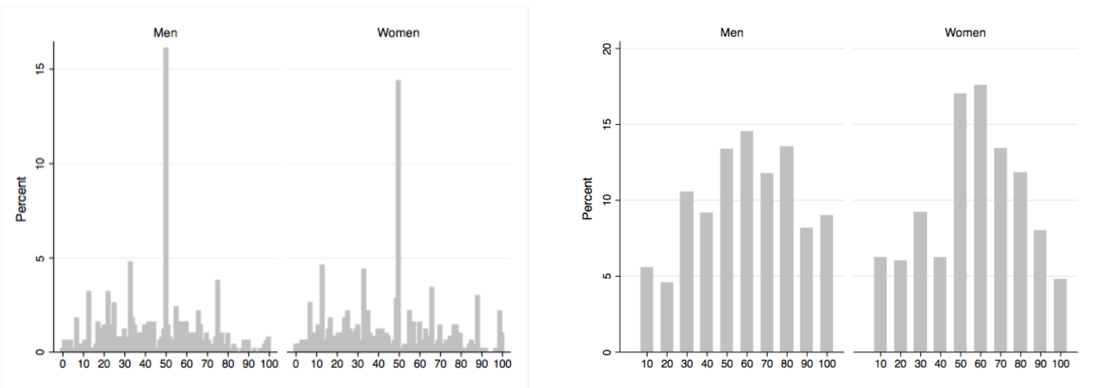
[4]Further explanations concerning the classification scheme can be found in Appendix B.

is that we do not elicit the subjects' cognitive processes by their actions, but by the verbal description of their action choice. The implementation was as follows: Two independent research assistants assigned a level-k type to each subject according to the mentioned classification scheme. The assistants did not have any information on the subjects' gender or other socio-demographic data. First, both research assistants independently classified the written statements of each subject without information about the chosen number and indicated the level-k. Afterwards, they were informed about the chosen number of the participant and could reconcile their classification, if necessary. The research assistants agreed on the classification of 91% of all participants, i.e., the intraclass correlation coefficient is 0.91. In the remaining 9% in which they did not agree, we were consulted to arbitrate.[5]

4 Results

4.1 Gender differences in level-k reasoning

We now turn to the results of our experimental games. First, we compare the action choices of women and men in the beauty contest and in the 11–20 money request game. Figure 1 shows the relative frequencies of the numbers chosen in both games.



(a) Beauty contest.

(b) Money request game.

Figure 1: Numbers chosen in the beauty contest and the money request game by gender.

A two-sample t-test as well as a Mann–Whitney test indicate that there are no gender differences between the numbers in neither of the two games. As described above, however, we measure the subjects' cognitive processes not by the action choices but by the verbal description of these action choices. We asked the participants in our experiments to give reasons for their decision, i.e., to describe exactly why they chose

[5]This approach is standard in social psychology (see, e.g., Channon and Crawford (2000)).

their numbers. Using this information as measure of the depth of strategic reasoning, we assign each subject a level-k type. Table 2 presents the relative frequencies of level-k types in both the beauty contest and the money request game. L0, L1, and L2 denote the percentages of level-0, level-1, and level-2 types, respectively. L3 denotes all types with level-3 or higher.

Level-k	Beauty contest		Money request	
	Women	Men	Women	Men
L0	91.62	86.28	85.43	77.93
L1	6.39	9.15	7.19	12.13
L2	1.40	3.98	4.79	6.96
L3	0.60	0.60	2.60	2.99

Table 2: Relative frequencies (in %) of level-k types by gender. Observations: 501 women, 503 men.

A Mann–Whitney test reveals significant gender differences in the level-k types in the beauty contest ($p < 0.01$) and in the money request game ($p < 0.01$). These results indicate that, on average, men exhibit a higher depth of strategic reasoning.

Finally, we use a logit analysis. Our dependent variable, the level-k type in both games, has four categories that can be sequentially ordered, because L3 is "higher" than L2, and so on. We therefore use an ordered logit model and regress *BCLEVEL* (= 0 if L0, = 1 if L1,... in the beauty contest) and *MRLEVEL* (in the money request game), respectively, on *GENDER* (= 1 if male, = 0 if female). Table 3 presents the estimation results of the ordered logit models.

	BCLEVEL	*MRLEVEL*	*DIFFLEVEL*
GENDER	0.561***	0.487***	0.459***
	(0.206)	(0.166)	(0.161)
Log likelihood	−428.483	−655.719	−660.039

Table 3: Ordered logit estimation results. Robust standard errors in parentheses. *** indicates significance at 1%. Observations: 501 women, 503 men.

The coefficients in the models are ordered log-odds coefficients. A positive coefficient for *GENDER* implies that the likelihood of having a higher level-k is higher for men in both games. For men (i.e., going from 0 to 1 in *GENDER*), we expect increases of 0.561 and 0.487 in the beauty contest and the money request game, respectively, in the log-odds of being a higher-level type. Thus, the results confirm the results of the

Mann–Whitney test indicating that men, on average, exhibit a higher depth of strategic reasoning. Because participants were asked about their socio-demographic characteristics, we are able to control for age, income, residential region, and religious affiliation. The significance of the coefficients, however, does not change when controlling for these variables.

4.2 Stability of level-k reasoning across games

Our within-subject design enables us to compare the level-k reasoning in the beauty contest and the money request game. This is important for two reasons: First, from a more general perspective, this comparison provides information on the portability of the level-k approach. Since the level-k classification is our key measure used for strategic reasoning, it is important to discuss how portable it is from one game to the other.[6] Second, this comparison sheds light on the stability of strategic behavior as a possible source of gender differences. Previous research on gender differences in social preferences has shown that women have a greater context sensitivity than men (Croson and Gneezy, 2009). With our within-subject we are able to examine whether there is a similar gender pattern in different strategic reasoning games.

A Spearman test shows a positive and significant correlation between the respective level-k in the beauty contest and the money request game for all subjects ($\rho = 0.41$, $p < 0.01$). Analyzing behavior on an individual level, we next calculate for each subject the difference between level-k in both games. This variable, *DIFFLEVEL*, has the value 0, if a subject was assigned the same level-k in the beauty contest and the money request game; it has the value 1, if both assignments differ by one level, and so on. *DIFFLEVEL* can have the (absolute) values 0, 1, 2 or 3 and can be interpreted as a measure of stability. Table 4 shows that 806 subjects (80.28%) behave consistently, i.e., their level-k is the same in both games. Only 67 subjects (6.67%) differ by more than one level-k in the two games. Although these results are clearly driven by the large number of level-0 types, they are nevertheless indicative of a relatively high stability of individual behavior across games.

Table 4 also shows that the stability of individual behavior differ between the genders. 421 women (84.03%), but only 385 men (76.85%) were assigned the same level-k in both games.[7] A Mann–Whitney test indicates that the gender difference in stability is significant ($p < 0.01$). We also use an ordered logit model and regress *DIFFLEVEL* on *GENDER*. The estimation results confirm the finding that women behave more con-

[6] Hargreaves Heap et al. (2014) examine the portability of level-0 assumptions in level-k theory in experimental coordination, discoordination and hide and seek games.

[7] Table 5 in Appendix C provides more information on the number of subjects in each level-k combination.

DIFFLEVEL	Women	Men	Total
0	421	385	806
1	50	81	131
2	22	27	49
3	8	10	18
Total	501	503	1,004

Table 4: Differences in level-k types across both games by gender.

sistently than men (see Table 3). Controlling for age, income, residential region, and religious affiliation do not change the significance of the coefficient.

4.3 Discussion of the results

In addition to gender differences, our results differ from other guessing game experiments particularly in the proportion of level-0 types. This may be because in our experiments, a sample of the German population, which is representative of that population in terms of some key characteristics, took part. In laboratory experiments, the participants are very often students who are, on average, better able to think strategically than ordinary people (Bosch-Domènech et al., 2002). However, there also seem to be significant differences between our results and the famous results of beauty contest newspaper experiments conducted with a large number of participants in the UK, Spain, and Germany. Those experimental results are reported in Bosch-Domènech et al. (2002). A possible explanation for the differences is that the newspapers—the British *Financial Times*, the Spanish daily business newspaper *Expansión*, and the monthly German science magazine *Spektrum der Wissenschaft*—are relatively prestigious publications whose readers are not representative segments of the populations of those three countries. The results thus might be biased such that the guessing games in those newspaper experiments demonstrate the (higher) strategic reasoning of a special group (Güth et al., 2007) rather than the decision making of an average citizen, as in our experiments. This might explain the higher proportion of level-0 types in our experiments.

Another explanation for the different results in our experiment may be due to the level-k classification approach. Nagel (1995) classifies the level-k types on the basis of the numbers chosen in the beauty contest . For example, this approach implies that someone choosing 33 is classified as level-1. Figure 1 shows peaks around the numbers 14, 22, 33, and 50 in the beauty contest which would correspond to the levels 3, 2, 1, and 0, respectively. Other approaches use subjects' optional ex post comments in addition to the chosen numbers (Bosch-Domènech et al., 2002) and subjects' chosen numbers in multiple games of the same class to obtain a decision rule that represents the observed overall behavior (Costa-Gomes and Crawford, 2006; Costa-Gomes et al., 2001; Stahl and Wilson, 1995). Instead of such approaches, we classify the participants by their incentivized comments on why they have chosen the number. As our data shows, these classification approaches can lead to very different results. For example, one participant in our beauty contest chose 33 but motivated the decision by: "33 is my lucky number". A classification by the chosen number would have resulted in a level-1 type, while in our approach the participant is a level-0 type since he plays some focal strategy. Another example from our data is a participant who chose the number 13. Since this number may have resulted from $33 \cdot 2/3 \cdot 2/3$, one could have classified this person as level-3 type. However, the motivation for choosing this number was "this is my birthday". Obviously, this participant played a focal strategy and was thus classified as level-0 type. Our data shows that not only the number 50, but also the majority of numbers around the peaks 14, 22, and 33 (see Figure 1) are motivated by such a focal strategy leading to a level-0 classification in our approach.

5 Conclusion

This paper examines gender differences in strategic reasoning in two experimental guessing games conducted with a large number of participants. In the beauty contest and the 11–20 money request game, we find that men, on average, seem to employ higher levels of reasoning than women. Our results are furthermore indicative of gender differences in stability across the games. We find that more women than men were assigned the same level-k in both games and interpret this as meaning that women behave more consistently.

These results provide an additional channel through which differences in the economic behavior of women and men can be explained. The findings may have important implications for economic policy, e.g., labor market policies aimed at reducing the gender wage gap. Traditionally, discrimination and differences in human capital have been seen as the main factors responsible for this wage gap. Recent studies point out that differences in social preferences and behavior in competitive situations are further influencing factors. Our results suggest that at least a part of the gender differences occurring in real-life situations, e.g., in wage bargaining, team production or promotion decisions, may be explained by behavioral differences in strategic situations.

Appendix

The first two parts of the appendix (A and B) provide the translated experimental instructions for the beauty contest and the money request game as well as the classification scheme for both games. All instructions and the questionnaire were originally written in German. Part C contains more information on the stability of level-k reasoning across games.

A Experimental instructions

A.1 Beauty contest

Your task is to choose an integer between 0 and 100. Then the average of all chosen numbers will be calculated. The winner is the person whose number is closest to 2/3 times this average. The winner receives a prize of €50.
Please give reasons for your decision. Dependent on how conclusive you explain your reasons you will receive an additional payment.

Fill in the number you have chosen:

A.2 Money request game

You and another person are playing a game in which each player requests an amount of money. The amount must be a number between 10 and 100 €-cent, which must be divisible by 10. Each person will receive the amount he requests. A person will receive an additional amount of 100 €-cent if she asks for exactly 10 €-cent less than the other person.
Please give reasons for your decision. Dependent on how conclusive you explain your reasons you will receive an additional payment.

Fill in the amount of money you request:

B Classification scheme

(The two independent research assistants received the following instructions to assign the levels to the participants according to the written statements.)

Please read the written statements of each participant and assign a level according to the following classification scheme. Please use only the verbal report of the participant for your classification. Do not try to interpret what the participants might have meant. You should only classify what you clearly extract from the written statement. Afterwards, you will be informed about the chosen numbers and get the opportunity to reconcile your classification.

B.1 Beauty contest

Level 0: There is no hint that the player thinks strategically in the way that she includes other players in her decision making. A level-0 type plays a focal strategy or chooses any number in the given interval for a non-strategic reason or she might not understand the game. Examples: "I choose 8 because it is my birthday.", "I choose 20 because it is my lucky number.", "I choose 50 because it is the mean".

Level 1: The player thinks strategically and responds best to any level-0 decision. The player calculates two thirds of her assumed level-0 distribution. Example: "I suppose other players choose 40, so I choose $40 \cdot 2/3 = 27$".

Level 2: The level-2 player responds best to any level-1 decision. Example: "The others may think everybody plays 40 and, thus, they subsequently choose 27. Taking this into account, I have to choose 2/3 of 27".

Level 3: The player assumes that others are able to do iteration step 2. Therefore the level-3 player best responds to any level-2 decision. The player argues that there are other players who realise that players best respond to level-1. So the level-3 type has to best respond to these level-2 players.

Level 4 and higher: The process of best responding to lower levels continues.

B.2 Money request game

Level 0: The player chooses the natural anchor 100 or any other number without realizing the iterative reasoning process.

Level 1: The player thinks strategically and responds best to the assumed level-0 decision. Example: "I suppose the other player chooses 100, so I choose 90 to earn the additional 100 €-cent".

Level 2: The level-2 player responds best to any level-1 decision. Example: "I suppose the other player thinks I choose 100, so he chooses 90. Therefore I choose 80 to earn the additional 100 €-cent".

Level 3: The player thinks strategically and responds best to the assumed level-2 decision. Example: "The other player thinks that I will choose 90 to win the additional 100 €-cent, so she should choose 80. Therefore I choose 70 to earn the additional 100 €-cent".

Level 4 and higher: The process of best responding to lower levels continues.

C Stability of level-k reasoning across games

Table 5 shows the frequencies of subjects for each level-k combination by gender. For example, 5.19% of women are assigned L0 in the beauty contest and L1 in the money request game. The sum of the frequencies on both diagonals (bold) are measures for stability across both games and correspond to the numbers for $DIFFLEVEL = 0$ presented in Table 4 (421 women (84.03%) and 385 men (76.85%)).

		Women					Men				
		Money request					Money request				
		L0	L1	L2	L3	All	L0	L1	L2	L3	All
Beauty contest	L0	**81.84**	5.19	3.19	1.40	91.62	**73.16**	8.15	3.18	1.80	86.28
	L1	2.99	**1.60**	1.00	0.80	6.39	3.38	**2.19**	2.78	0.80	9.15
	L2	0.40	0.40	**0.40**	0.20	1.40	1.19	1.59	**0.99**	0.20	3.98
	L3	0.20	0.00	0.20	**0.20**	0.60	0.20	0.20	0.00	**0.20**	0.60
	All	85.43	7.19	4.79	2.60	100.00	77.93	12.13	6.96	2.99	100.00

Table 5: Relative frequencies (in %) of level-k types across both games by gender.

References

Andreoni, J. and L. Vesterlund (2001). Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics 116*(1), 293–312.

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

Bertrand, M. (2011). New perspectives on gender. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4B, pp. 1543–1590. Amsterdam: Elsevier.

Bosch-Domènech, A., J. G. Montalvo, R. Nagel, and A. Satorra (2002). One, two, (three), infinity...: Newspaper and lab beauty-contest experiments. *American Economic Review 92*(5), 1687–1701.

Breitmoser, Y. (2012). Strategic reasoning in p-beauty contests. *Games and Economic Behavior 75*(2), 555–569.

Burchardi, K. B. and S. P. Penczynski (2014). Out of your mind: Eliciting individual reasoning in one shot games. *Games and Economic Behavior 84*, 39–57.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton, New Jersey: Princeton University Press.

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

Castillo, M. E. and P. J. Cross (2008). Of mice and men: Within gender variation in strategic behavior. *Games and Economic Behavior 64*(2), 421–432.

Channon, S. and S. Crawford (2000). The effects of anterior lesions on performance on a story comprehension test: Left anterior impairment on a theory of mind-type task. *Neuropsychologia 38*(7), 1006–1017.

Charness, G. and U. Gneezy (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization 83*(1), 50–58.

Costa-Gomes, M. A. and V. P. Crawford (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review 96*(5), 1737–1768.

Costa-Gomes, M. A., V. P. Crawford, and B. Broseta (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica 69*(5), 1193–1235.

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri (2013). Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature 51*(1), 5–62.

Croson, R. T. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic Literature 47*(2), 448–474.

Datta Gupta, N., A. Poulsen, and M. C. Villeval (2013). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry 51*(1), 816–835.

Dittrich, M., A. Knabe, and K. Leipold (2014). Gender differences in experimental wage negotiations. *Economic Inquiry 52*(2), 862–873.

Dittrich, M. and K. Leipold (2014). Gender differences in time preferences. *Economics Letters 122*(3), 413–415.

Dwyer, P. D., J. H. Gilkeson, and J. A. List (2002). Gender differences in revealed risk taking: Evidence from mutual fund investors. *Economics Letters 76*(2), 151–158.

Eckel, C. C. and P. J. Grossman (1996). The relative price of fairness: Gender differences in a punishment game. *Journal of Economic Behavior & Organization 30*(2), 143–158.

Eckel, C. C. and P. J. Grossman (1998). Are women less selfish than men? Evidence from dictator experiments. *Economic Journal 108*(448), 726–735.

Eckel, C. C. and P. J. Grossman (2001). Chivalry and solidarity in ultimatum games. *Economic Inquiry 39*(2), 171–188.

Ericsson, A. K. (2002). Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology 16*(8), 981–987.

Ericsson, A. K. (2003). Valid and non-reactive verbalization of thoughts during performance of tasks. *Journal of Consciousness Studies 10*(9-10), 1–18.

Fehr, E., M. Naef, and K. Schmidt (2006). The role of equality, efficiency, and Rawlsian motives in social preferences: A reply to Engelmann and Strobel. *American Economic Review 96*(5), 1912–1917.

García-Gallego, Aurora, N. G. and A. Jaramillo-Gutiérrez (2012). Gender differences in ultimatum games: *Despite* rather than *due to* risk attitudes. *Journal of Economic Behavior & Organization 83*(1), 42–49.

Gneezy, U., K. Leonard, and J. List (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica 77*(5), 1637–1664.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics 118*(3), 1049–1074.

Güth, W., C. Schmidt, and M. Sutter (2007). Bargaining outside the lab – a newspaper experiment of a three-person ultimatum game. *Economic Journal 117*(518), 449–469.

Hargreaves Heap, S., D. Rojo-Arjona, and R. Sugden (2014). How portable is level-0 behavior? A test of level-k theory in games with non-neutral frames. *Econometrica*, forthcoming.

Jianakoplos, N. A. and A. Bernasek (1998). Are women more risk averse? *Economic Inquiry 36*(4), 620–630.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review 85*(5), 1313–1326.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics 122*(3), 1067–1101.

Selten, R. and A. Ockenfels (1998). An experimental solidarity game. *Journal of Economic Behavior & Organization 34*(4), 517–539.

Stahl, D. O. and P. W. Wilson (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior 10*(1), 218–254.

Weizsäcker, G. (2003). Ignoring the rationality of others: evidence from experimental normal-form games. *Games and Economic Behavior 44*(1), 145–171.