University of Heidelberg

Department of Economics

Discussion Paper Series | No. 516

More than Meets the Eye: an Eye-tracking Experiment on the Beauty Contest Game

Julia Müller and
Christiane Schwieren

September 2011

More than Meets the Eye: an Eye-tracking Experiment on the Beauty Contest Game*

Julia Müller† Christiane Schwieren ‡

September 22, 2011

Abstract

The beauty contest game has been used to analyze how many steps of reasoning subjects are able to perform. A common finding is that a majority seem to have low levels of reasoning. We use eye-tracking to investigate not only the number chosen in the game, but also the strategies in use and the numbers contemplated. We can show that not all cases that are seemingly level-1 or level-2 thinking indeed are – they might be highly sophisticated adaptations to beliefs about other people's limited reasoning abilities.

Keywords: beauty contest game, levels of reasoning, level-k model, strategic reasoning, cognitive hierarchy, iterated elimination of dominated strategies, experiment.

JEL-Classifications: C72, C91, D03

*We would like to thank Sven Brüssow, Christoph Brunner, Giorgio Coricelli, Daniel Holt, Dan Levine, Rosemarie Nagel and Jörg Oechssler for valuable comments and sugesstions. Financial support from the START-Professorship of Heidelberg University, from the DFG Initiative of Excellence is gratefully acknowledged.

†Alfred-Weber-Institut for Economics, University of Heidelberg, Bergheimer Strasse 58, 69115 Heidelberg, Germany; email: julia.mueller@awi.uni-heidelberg.de

‡Corresponding author, Alfred-Weber-Institut for Economics, University of Heidelberg, Bergheimer Strasse 58, 69115 Heidelberg, Germany; email: christiane.schwieren@awi.uni-heidelberg.de

1 Introduction

The beauty contest game is frequently used to analyze the depth of strategic thinking of ordinary people. In this game all players have to state a number between 0 and 100 simultaneously. The payoff is a fixed amount for the winner; all other players get nothing. The winner of the game is the person whose chosen number is closest to the mean of all chosen numbers multiplied by a predetermined positive parameter. (If more than one person chooses the same number the prize is divided equally among the winners.) The game has only one unique Nash equilibrium: all players pick zero. This equilibrium can be reached via several steps of either iterated elimination of dominated strategies or iterated best response.

Empirically, however, players usually do not state zero, but rather choose a number that indicates only one or two iterations; in other words, people seem to apply only low levels of reasoning. This is, however typically only inferred from the numbers stated. Another possibility is that people in fact have higher levels of reasoning, but after geting through many steps of iterated reasoning, decide that others might not be as smart and therefore choose a number being interpreted as showing only low levels of reasoning.

We used eye-tracking to get a deeper understanding of how people choose the number they state in the guessing game. Eye-tracking has recently been used in economic experiments to distinguish between different possible decision processes leading to similar results (see e.g., Arieli et al. (forthcoming), Knoepfle et al. (2009), Wang et al. (2010), Reutskaja et al. (2011)). Eye-tracking technology records what the subject is looking at. The assumption underlying the use of eye-tracking is that people tend to look at data they process. Our design permits us to investigate the procedures that subjects use in choosing a number. While following their eye movements, we first presented the rules of the game for a fixed time span and then a number ray from 0 to 100. Knowing which numbers subjects looked at informs us which numbers subjects contemplated in the beauty contest game. Monitoring the sequence of numbers considered in the thinking process gave us a deeper insight into the strategies used.

We found that the evidence with respect to levels of reasoning is less clear than has so

far been assumed. We found different strategies that look similar when just focusing on the stated number: Choosing a number associated with level-1 or level-2 reasoning can in fact be the outcome of level-1 or level-2 reasoning, but it can also be the outcome of higher level types adjusting their chosen number to their beliefs of what other people might do. In many cases we discover that subjects contemplate choosing low numbers and later go back and choose a number consistent with level-1 or level-2 reasoning.

We cannot reject with our data that some people only engage in level-1 or level-2 reasoning, but we can show that not all cases that are seemingly level-1 or level-2 thinking indeed are – they might be a form of highly sophisticated adaptation to beliefs about other people's limited reasoning abilities.

2 The beauty contest game

2.1 Definition of the game

The beauty contest game was first mentioned by Keynes (1936) and later introduced formally by Moulin (1986).[1]

In this game each of n players, $n \geq 2$, simultaneously choose a number x_i from a given interval, usually $[0, 100]$. The player whose chosen number is closest to p times the mean of all numbers $x_i, i = 1, \ldots n$, wins a fixed and known prize. If there is a tie among m players with $m \leq n$, then the prize is divided equally among them. All other players get nothing. The value of the parameter p is common knowledge before the game starts.

2.2 Nash equilibrium and dominance

For p with $0 \leq p < 1$ there exists only one Nash equlilibrium: all players announce zero.[2]

The beauty contest game is dominance solvable and was originally experimentally tested in the laboratory to see how many steps of reasoning subjects are performing. Iterated

[1] For the history of the beauty contest game see Bühren et al. (2009).

[2] The uniqueness of the equilibrium is only true for beauty contest games where players can choose a real number out of the interval; if only integer numbers are allowed there are muliple equilibria; see López (2001) for a charaterization of these.

elimination of dominated strategies starts in the first step with the elimination of all numbers larger than $100p$ and then those larger than $100p^2$, $100p^3$ and so forth. An infinite number of steps will lead towards zero, the only undominated number and the unique equilibrium point of the game. People have infered the level of reasoning by the number of steps of elimination of dominated strategies, identifying a choice as level k if it is included in the interval $[100p^{(k+1)}, 100p^k]$.[3]

2.3 Level-k models

Deviations from Nash equilibrium play have been widely demonstrated and one of the approaches to model this behavior is often referred to as the "level-k"- or "cognitive hierarchy"-model (e.g. Nagel (1995), Stahl and Wilson (1995), Costa-Gomes et al. (2001), Camerer et al. (2004), Ho et al. (1998), Crawford and Iriberri (2007a), Crawford and Iriberri (2007b)). The key assumption here is that, departing from the idea of complete rationality and consistency in strategies and beliefs, one allows for a hierachy of beliefs or differing depths of reasoning. All variants model step by step reasoning with heterogeneous types, where the number of steps define the level of reasoning, thinking or sophistication. All models have in common that the depth of strategic thinking is incorporated into the number of applications of a best response procedure that subjects do. So in general, level-k is defined by best responding to level-$k-1$. The level-0 type is defined differently: playing uniformly distributed over the given intervall or the choosing of a focal strategy.

2.4 Experimental results

The beauty contest game has been used in various experiments in the laboratory and in the field. The first experiments on the beauty contest game showed quite unambiguously that most players do not play the unique Nash equilibrium especially in the first round. From the perspective of iterated elimination of dominated strategies, this means that players do not perform the infinite number of iterated eliminations leading to zero. Instead Nagel

[3] One could also use other classifications since choosing numbers lower than $100p^{(k+1)}$ may also be a best response given type-k's beliefs.

(1995) proposed in her first experimental study on the beauty contest game a model of boundedly rational behavior to explain the behavior observed in the first period.[4] This model of iterated best reply with limited elimination captures the following types of players: level-0-players choose randomly between 0 and 100, level-1-players best reply[5] to this with $50p$, level-2-players choose $50p^2$ (a best reply to level-1) and level-3-players choose $50p^3$. The experimental results of Nagel (1995) fit well to this model: for $p = 1/2$ and $p = 2/3$ no one picked zero and the average chosen numbers are 27 and 36, respectively. Many subjects perform one step of reasoning, choosing 33 with $p = 2/3$, or 2 steps, choosing 22. Ho et al. (1998) were the first to replicate these main findings, and analyze different learning models for games where feedback was given.

Building on these first experiments, others (for a survey, see Nagel (1999), Camerer (2003), Crawford et al. (2010)) varied different aspects of the game, like the number of players, repetitions, or the payoff. The results with respect to first round behavior are always quite clear. The Nash equilibrium (zero) is reached in less than 2.5% of the cases, and a proportion of 5 to 25% of the players play a dominated strategy, i.e. choose a number between $100p$ and 100. Inferring the level of iterated dominance from the data, all studies find relatively low levels: the modal level is two, and there are only few subjects with a level higher than 3 (less than 5%). Going from the laboratory to the field, Bosch-Domenech et al. (2002) corroborate the results in large newspaper-based experiments. In addition to playing the game, participants of their experiments were asked how they chose their number. By classifying these answers, the authors were able to find different types of reasoning processes. They find five different types: two of them use a game theoretic argument (fixed point argument and iterated elimination of dominated strategies), two types use arguments mentioned in the beauty contest game literature (where the first type starts his analysis with the mean of 50 and then uses the reasoning described above and the other type is best replying to a probability distribution of types) and the last type, called "experimenter",

[4]There are also models for the other periods, after subjects got feedback. We do not concentrate on these learning models, because in our experiment we gave no feedback and are not interested in learning.

[5]Breitmoser (2010) shows that $50p$ is not in general a best reply to uniformly randomizing players, but that the best reply significantly differs when the number of players is low compared to an approximately infinite number of players.

conducts his or her own experiments with friends to find out what they are doing. Additonally Bosch-Domenech et al. (2002) introduce a classification of those subjects who reason until equilibrium, but then choose non-equilibrium strategies. In a neuroeconomic study Coricelli and Nagel (2009) identify different neural substrates of subjects with low and high levels of reasoning.

Psychologists use the term *theory of mind* to describe the ability to understand other minds. Ohtsubo and Rapoport (2006) review the beauty contest game (and the investment game) and assume that one underestimates the depth of reasoning, because subjects may perform many steps of the iteration towards the equilibrium solution and even figure it out, but then state a higher number because they think that the $n-1$ other players are not as smart as they are. Therefore, in the beauty contest game, high levels of reasoning and a theory of mind of the other players can cause a number associated with low levels of reasoning. The same number could be reached by truly low levels of reasoning.

There are attempts to use other data than choices to learn more about the decision process underlying choices. Costa-Gomes et al. (2001) and Costa-Gomes and Crawford (2006) used MouseLab to ascertain information search behavior. Verbal data also is used to get more information on the reasoning of subjects, for example Nagel (1993), Bosch-Domenech et al. (2002), Sbriglia (2008), Burchardi and Penczynski (2011). So far, there is one other experiment that has used eye-tracking together with the beauty contest game. Chen et al. (2009) introduce a two-person beauty contest game played spatially on a two-dimensional plane. The authors classify subjects into various types, based on choices and on eye-tracking data. They find that more than half of the subjects are classified in the same class by both procedures, and that some subjects are classified into a higher level-k-type using the eye-tracking data than using the choice data. But as they use the two-person game, we aim to show that a similar result can also be verified with the standard beauty contest game.

3 Experiment

3.1 Method

We recorded subjects' eye movements using the EyeLink II Eyetracking System made by SR Research Ltd./Canada. The EyeLink II is a head mounted video-based eye tracker. It consists of three miniature cameras mounted on a padded headband. Two eye cameras allow binocular eye-tracking or easy selection of the subject's dominant eye without any mechanical reconfiguration. An optical head-tracking camera integrated into the headband allows accurate tracking of the subject's point of gaze. We used a chin rest to inhibit movement of the subjects.

3.2 Design and Procedures

Subjects played six rounds of a repeated one-shot beauty contest game with no information or feedback in between rounds.

round	1	2	3	4	5	6
p	0.125	0.2	0.33	0.5	0.66	0.75

Table 1: Values of p used in the experiment

They had to choose a number out of the interval $[0, 100]$. Subjects chose a number by saying it aloud. They were instructed that when viewing at the screen shown in figure 1 they should think about which number to choose and then pronounce the number chosen. Using eye-tracking technology it is important that subjects focus on the monitor; and therefore, typing in the chosen number on the keyboard is impossible.

The different parameters were always presented in the same order as shown in table 1.

The number of players n was ten, but subjects came one by one to the eye-tracking laboratory. Each subject was informed that he was playing with nine other players who either had already played or would play later, up to reaching ten players in total.[6] The design of our experiment differs from most other beauty contest games in laboratories, but

[6]Translated instructions of the experiment can be found in the appendix B

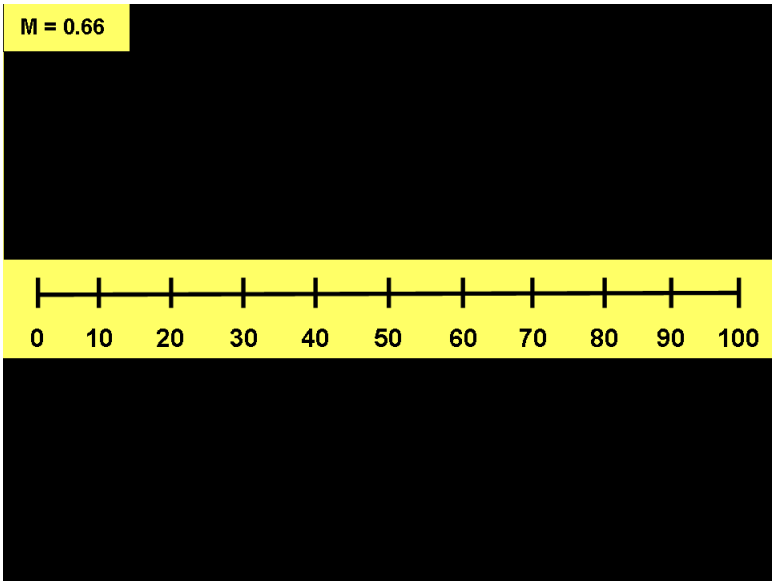


Figure 1: Screen when subjects choose the number (example of round 5, parameter 0.66 is shown in the upper right corner)

it is very similar to the design by Coricelli and Nagel (2009), because the needs of the eye-tracking technology are comparable to the needs of neuroeconomic experiments. We also use the same parameters as Coricelli and Nagel (2009), but the order of parameters differs.

For each subject we first determined his or her dominant eye. We fixed the headmounted system and chose the respective camera. After fixing the system confortably on subjects' head the experiment started with a calibration phase. Only after reaching a good fit we proceded with the experiment. First subjects saw a general instructing screen telling that the experiment will now start. Then subjects saw the specific instructing screen telling how to determine the target number. For each round of the beauty contest game the exact timeline of events on the monitor was as follows: First subjects were informed about the parameter; they saw the number in two formats, i.e. as 0.66 and in percent (66%). This screen was followed by a calibration to secure exact measurement for the following trial. Subjects saw the number ray (together with the parameter, see figure 1) and knew that they now should choose the number for this round. They could think about which number to choose without any time restrictions, and finally had to press any key on the keyboard

to proceed to the next round.

The experiment was conducted in December 2009 and March 2010 in the eye-tracking laboratory of the Psychology Department of Heidelberg University. We had 39 subjects in total[7]. Participants were students of various fields of study. Subjects received the general part of the instructions at the beginning of the experiment and could ask questions.

The fixed prize of each round in the beauty contest game was 10€ and we paid all rounds. The experiment lasted for about 30 minutes. Participants earned €16.01 on average. All subjects were paid in cash and in private. Subjects were paid after all ten players had played.

4 Results

4.1 Behavioral results

Because we ran six rounds of the beauty contest game, we have 231 decisions[8] in total. For the first analysis we do not treat rounds differently and do not take learning into account, as we gave no feedback and no information between rounds. We do analyze the data on subject level later.

In table 2 we list the mean and median of the chosen numbers.

	$p = 0.125$	$p = 0.2$	$p = 0.33$	$p = 0.5$	$p = 0.66$	$p = 0.75$
N	39	38	38	38	39	39
Mean	29.72	34.79	39.84	41.68	46.41	46.72
Median	24	30	35	40.5	45	39

Table 2: Mean and median of the chosen numbers

If we compare the chosen numbers with the typical number choices in other experiments, we find that our subjects chose somewhat higher numbers. For example our mean of 46.41 (for p = 0.5) is definitely larger (t-test two-tailed,t =2.839, p=.007) than the mean of 36

[7]Originally we had eye-tracking data of 40 subjects, but we excluded one subject from the analysis who was familiar with the beauty contest game.

[8]We should have $6 \times 39 = 234$ decisions, but for one subject three decisions are missing, because they got not recorded.

reported in Nagel (1995). Figure 2 gives an overview of the distribution of number choices for the different parameters of p we used. None of our subjects chose zero, which is in line with the usual laboratory finding that only very few people chose the Nash equlibrium - and if the Nash equilibrium is chosen then it usually happens in later rounds, after learning from feedback.

For analysing how many subject choose a weakly dominated strategy we are going to look for chosen numbers larger than $100p$. The frequency of the choices larger than $100p$ for the different parameters p can be found in table 3. For the larger parameters we find the usual percentage, around 20 to 25% choose dominated strategies, while for the smaller parameters we have more subjects choosing a dominated strategy.

	$p = 0.125$	$p = 0.2$	$p = 0.33$	$p = 0.5$	$p = 0.66$	$p = 0.75$
N	39	38	38	38	39	39
Frequency	26	26	22	10	9	6
Percent	66.7%	68.4%	57.9%	26.3%	23.1%	15.4%

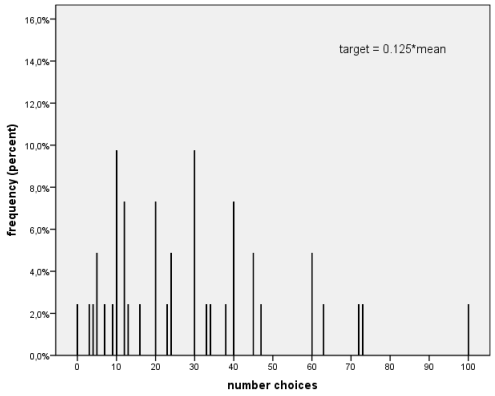
Table 3: Choice of weakly dominated strategies

We use two different methods to calculate levels of reasoning using only the chosen number. We use one level-k model which determines the level of k by $100p^k$ and a second model that determines the level by $50p^k$. To determine the level of thinking by the chosen number x_{ip} (in the round with parameter p) we choose the k resulting in the smallest quadratic distance

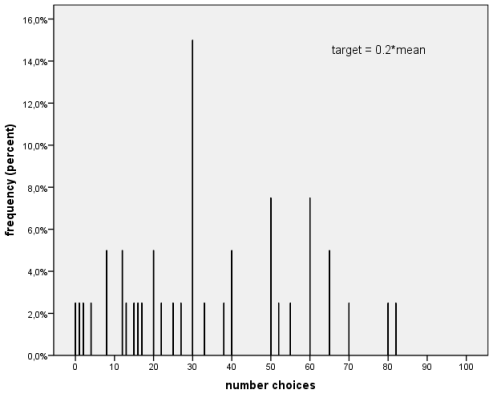
$$QD_i = (x_i - 100p^k)^2 \text{ or } \quad QD_i = (x_i - 50p^k)^2.$$

An array of frequencies and percentages for the calculated levels for each parameter can be found in tables 4 and 5.

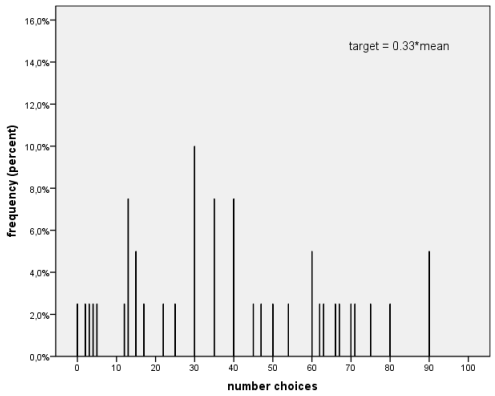
Using both methods of calculation we find that a majority of subjects have low levels of reasoning. With the level-k model using 100 as a starting point and with the level-k model using 50 respectively, we find levels strictly larger than three in about 13% respectively 6% of cases, which is close to the usual fraction (below five percent for the level-k-50) mentioned in



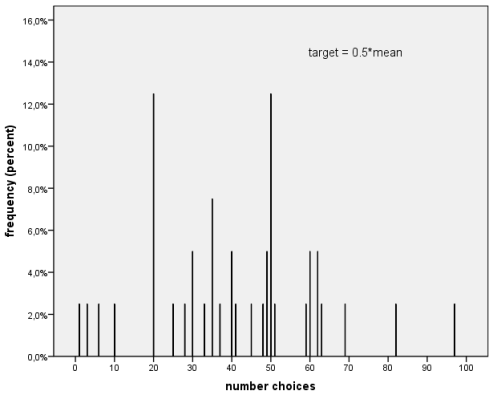
(a) $p = 0.125$



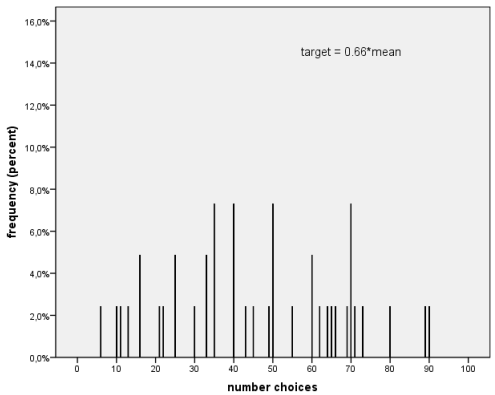
(b) $p = 0.2$



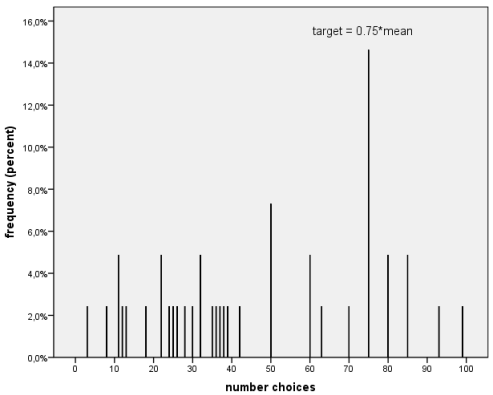
(c) $p = 0.33$



(d) $p = 0.5$



(e) $p = 0.66$



(f) $p = 0.75$

Figure 2: Number choices

	$p = 0.125$		$p = 0.2$		$p = 0.33$		$p = 0.5$		$p = 0.66$		$p = 0.75$		Total	
0	6	15.4%	5	12.8%	7	17.9%	2	5.1%	2	5.1%	2	5.1%	24	10.2%
1	28	71.8%	26	66.7%	20	51.3%	19	48.7%	13	33.3%	10	25.6%	116	49.5%
2	5	12.8%	5	12.8%	7	17.9%	14	35.9%	9	23.1%	6	15.4%	46	19.6%
3			2	5.1%	3	78.7%	1	2.6%	8	20.0%	4	10.3%	18	7.7%
4			1	2.6%	1	2.6%	0	0.0%	3	7.7 %	6	15.4%	11	4.7%
5					1	2.6%	1	2.6%	2	5.1%	4	10.3%	8	3.4%
6							2	5.1%	1	2.6%	1	6.6%	4	1.7%
7									1	2.6%	2	5.1%	3	1.2%
8											2	5.1%	2	0.8%
9											1	2.6%	1	0.4%
10											0	0.0%	0	0.0%
11											0	0.0%	0	0.0%
12											1	2.6%	1	0.4%
N	39		39		39		39		39		39		234	

Table 4: Levels via level-k model using 100

	$p = 0.125$		$p = 0.2$		$p = 0.33$		$p = 0.5$		$p = 0.66$		$p = 0.75$		Total	
0	19	48.7%	17	43.6%	21	53.8%	21	53.8%	20	51.3%	18	46.2%	116	49.5%
1	19	48.7%	18	46.2%	13	33.3%	14	35.9%	9	23.1%	6	15.4%	79	33.7%
2	1	2.6%	2	5.1%	2	5.1%	1	2.6%	4	10.3%	6	15.4%	16	6.8%
3			1	2.6%	2	5.1%	0	0.0%	3	7.7%	2	5.1%	8	3.4%
4			1	2.6%	1	2.6%	1	2.6%	2	5.1 %	1	2.6%	6	2.6%
5							2	5.1%	1	2.6%	4	10.3%	7	2.9%
6											1	2.6%	1	0.4%
7											0	0.0%	0	0.0%
8											0	0.0%	0	0.0%
9											0	0.0%	0	0.0%
10											1	2.6%	1	0.4%
N	39		39		39		39		39		39		234	

Table 5: Levels via level-k-model using 50

the literature (2.3). Clearly, one gets different, but similar levels by using differing methods of estimating levels. For the rest of the paper we assign levels using the level-k model using 50. We chose this method because it is more frequently used in the literature, and we mainly use the idea of the number of steps that indicate the level of reasoning in the following analyses. However, nothing substantial would change with respect to our results using the idea of iterated elimination as the basis for defining levels.

So far we have calculated the levels of reasoning for each round of the beauty contest game. Now we attempt to assign one level of reasoning to each of our subjects. 5 of our 39 subjects show the same level for all different parameters, and if we decide the level by majority rule[9] we can assign a unique level to 26 subjects. Three more subjects have level-0 in half of the cases and level-1 in the other half, so these subjects would have either level-0 or level-1. For three subjects we find clearly increasing levels from the first to the last round. We can not assign one level to these subjects, but can classify them as "learners". (Recall that we give no feedback between the different rounds of the game, but Weber (2003) found in his experiments with the beauty contest game that players seem to learn even in conditions without feedback by mere experience. Subjects played 10 rounds of the beauty contest game and were exposed to the same parameter ($p = \frac{2}{3}$).) There remain only seven subjects that we cannot classify.

The distribution of levels can be found in table 6. Comparing this assignment of levels by subject with the levels assigned by decision we can conclude that the main pattern persists: We find few subjects with high levels (namely only three subjects with levels strictly larger than level-3 or in total only four subjects with level-2 or higher) and the majority of subjects has level-1 (10 subjects) or level-0 (18 subjects).

That a majority of our subjects seem to have low levels of reasoning replicates the findings of many beauty contest games with different parameters and with different subject pools[10] in the literature.

[9]This means that we assign, for example, level-2 to a subject if he showed level-2 in at least four out of six choices. In this assignment of levels we follow Coricelli and Nagel (2009) who used the same rule to determine levels on subject-level based on their choices.

[10]Only game theorists and self-selected newspaper readers show higher levels and pick the equilibrium more often, see Camerer (2003).

by majority		**tie**		**learner**	
level	fequency	level	frequency	level	frequency
0	18				
1	7	*0 or 1*	3		
2	1				
3					
4				*0 to 5*	1
5				*1 to 5*	1
6				*1 to 6*	1
?	7				

Table 6: Levels per subject

Strategic IQ To have a unique measure for each participant we finally calculated a "strategic IQ" for each subject as first introduced by Bhatt and Camerer (2005). We based our calculation on the procedure developed by Coricelli and Nagel (2009). We employ the quadratic distance of choices to the winning numbers. We then calculated the winning numbers for each round using a recombinant estimation method (compare Mullin and Reiley (2006) and Mitzkewitz and Nagel (1993)).

We have a measure of strategic IQ for each round and take the average over the rounds to generate one aggregate measure of strategic IQ.

Additionally we asked our participants, as a proxy for intelligence, to provide their grade in the "Abitur" (the German Highschool diploma) and for their grades in Mathematics and German separately. None of these three measures correlates with our measure of strategic IQ.

The measure of strategic IQ yields low values for high strategic reasoning and high values for low strategic reasoning. We see in figure 3, which shows the distribution of strategic IQ, that most subjects have a rather high value on the strategic IQ measure. This is in line with the rather low levels of reasoning we find.

4.2 Eye-tracking results

So far we have used only the chosen numbers to relate our data to the existing literature on guessing games. In the following, we will use our eye-tracking data to analyze the decision process of our participants.

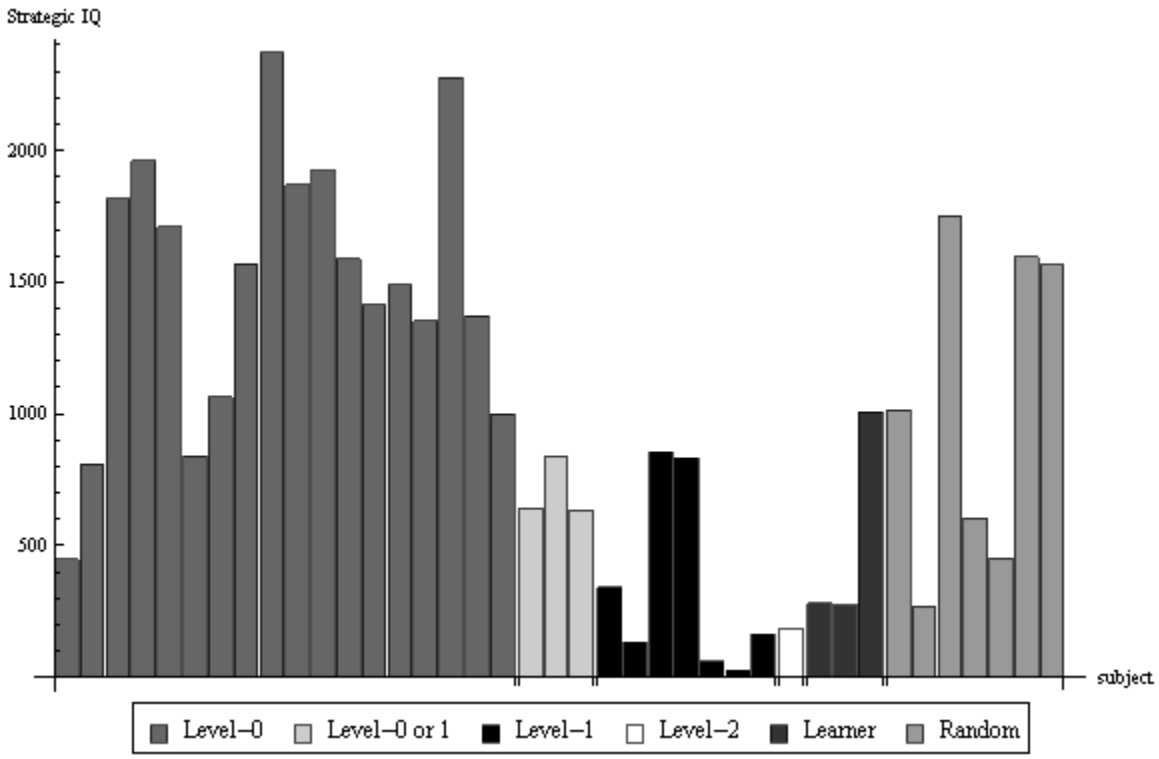


Figure 3: Strategic IQ for each subject, orderd by the level assignment as listed in table 6

Data analysis While our subjects could choose any number between 0 and 100, with eye-tracking we are not able to identify exact numbers (like 33), but rather areas a subject focused on, e.g., around 30.

Therefore we analyze the data using interest areas. We divide the number ray equally into rectangular interest areas: *around zero*, *around ten*, *around twenty*, ..., *around hundred* and one interest area *parameter*, which captures when subjects looked at the parameter of the round, which we presented alternating in one of the upper corners.

For the analysis of our data we mainly use fixations. Fixations are states where the eye is in relative motionlessness. To define a state as a fixation we use the preset definitons of the Eyelink II System.

Reaction Times Remember that subjects pronounced the number they chose in one round of the beauty contest game, and then had to press any key on the keyboard in front of them to proceed to the next round. We use the duration of a given round of the beauty contest game between the onset of the screen and the pressing of the key as a proxy for the reaction time.

The average duration of a round of the game was 14832 milliseconds. Average durations seperately for each round can be found in table 7. Although we do not provide feedback, subjects seem to become familiar with the game in the sense that they are able to decide faster in later rounds.

round	**parameter**	**duration**
1	$p = 0.125$	28538
2	$p = 0.2$	16942
3	$p = 0.33$	15000
4	$p = 0.5$	9301
5	$p = 0.66$	8971
6	$p = 0.75$	9225

Table 7: Duration for the different rounds

Use of number ray and parameter As we gave subjects the necessary information about the game before the number ray appeared, subjects could essentially have closed

their eyes to think about the given problem, without looking at the number ray or even at the screen at all. But we can see that subjects do use the number ray. They do not look randomly (uniformly distributed) across the whole screen, but use the available information in a systematic way. Fixations are either on the number ray or on the parameter p, given in one of the upper corners.

Subjects used the information given by the parameter while they decided. In table 8 one could find the percentage of fixations in the interest area *parameter* while subjects saw the number ray and decided which number to chose. Subjects use this information more in the early rounds, which is in line with the declining reaction times for later rounds.

round	**parameter**	**percent**
1	$p = 0.125$	21.9
2	$p = 0.2$	22.6
3	$p = 0.33$	18.7
4	$p = 0.5$	12.0
5	$p = 0.66$	13.4
6	$p = 0.75$	12.5

Table 8: Fixations in the interest area *parameter*

4.2.1 Comparison of number stated and last contemplated number

As a control for the feasibility of our design we tested whether subjects at the end of each round looked at the number they chose in that round. As we could not identify exact numbers we interpret when a subject looked at the corresponding area as looking at the number, i.e. a subject stating 27 should have a last fixation in the interest area *thirty*. We find that in 67.9% of the cases subjects looked at the number they choose. This might seem like a relatively small percentage, but our design might provide an explanation for this.

As we did not set a time constraint per round it is possible that the remaining fraction of subjects stated the number not at the very end of the round (i.e., right before pressing a key to continue), but rather stated the number and then kept watching the number ray for a while before pressing a key to proceed to the next round. We also calculated, therefore, in how many of the remaining cases subjects had a fixation in the respective interest area

among the last five fixations in that round. In total, 85.5% of the cases fulfill this relaxed criterion.

Hot Spots Hot spots were analysed seperately for each of the six rounds and seperately for both locations of the parameter (in the left or the right upper corner). In table 9 we listed the average total number of fixations seperately, for parameter in the left corner (first row) and right corner (second row).

	0	10	20	30	40	50	60	70	80	90	100	p
1	1.21	7.08	10.37	8.83	12.62	**14.71**	7.46	3.04	1.67	2.08	2.33	**20.04**
1	1.13	6	6.53	6.13	7.73	**9.46**	7.8	6	1.93	2.33	1.73	**22.67**
2	0.83	2.13	4.22	3.09	3.30	**7**	2.78	2.22	1	0.69	1.04	**11.39**
2	0.06	2	4.31	6.69	8.44	**8.87**	7.19	3	1.87	0.87	0.43	**13.12**
3	0.21	1.37	2.25	6.25	6.67	**9.46**	6.58	3.08	2.17	1.42	0.46	**7.71**
3	0.27	1.53	2.67	5.07	3.33	**7.33**	3.47	1.6	1.13	0.6	0.8	**12.07**
4	0.13	1.13	3.13	2.87	3.22	**6.48**	2.96	1.04	0.26	0.35	0.17	**5.34**
4	0.06	0.69	2.94	5.31	**6.75**	**5.69**	3.19	1.87	3	0.81	0.12	2.75
5	0	0.5	2.79	3.21	3.92	**5.29**	4.08	2.08	0.71	0.21	0.29	**4.5**
5	0.14	1.86	4.07	2.28	3.07	**4**	3.36	3	2.07	1.43	0.28	**7.93**
6	0.09	2.09	1.76	3.05	1.95	**4.14**	3	3.86	1.95	0.81	0.14	**4.81**
6	0.23	0.53	2	3.88	**5.41**	**5.53**	2.41	2	1.65	2.06	0.59	4.88

Table 9: Total number of fixations at each interest area, averaged across subjects, 1st row: parameter left, 2nd row: parameter right (in bold: the two largest numbers)

In most rounds the maximum number of fixations is in the interest area of the parameter. The second highest number of fixations is at *around 50*. Most people seem to use 50 as an anchor when choosing their number. Recall however that our fixation point was in the middle of the screen, thus being in the interest area of 50. This might have drawn people to this focal point rather than to 100. Most other hot spots are below 50, which is a hint that people behave in line with a level-k model rather than according to dominance.

4.2.2 Best reply model

Using the eye-tracking data we could trace which numbers subjects contemplated while deciding which number to choose. 55 of the 234 decisions, that is 23.50% of the choices, were made looking only at numbers below 50. This is an indication that the decision was

made in a way described by a level-k model, starting with 50 as the average of random choices (exact statements about the starting point are contaminated by the fixation in the middle before the rounds; compare the analysis of the Hot Spots).

choices out of all 6	**frequency**	**percent**
6	1	2.6%
5	0	0%
4	1	2.6%
3	6	15.4%
2	8	20.5%
1	12	30.8%
0	11	28.2%

Table 10: Fixations only below 50

On the subject-level only one person looked only at numbers below 50 (for all the six choices); for seven subjects this was the case in three or four out of the six choices (see table 10). Only 11 of our 39 subjects never followed this pattern. Most of our subjects seem to decide – at least for some of their decisions – following a level-k model.

Classification Our main interest regarding the eye-tracking data is to test the hypothesis that levels are estimated too low when only taking the chosen number into account. Using the chosen number alone one can not distinguish between a process of a certain, low number of iterations leading to the chosen number from a process of *more* iterations resulting in a lower number (closer to zero), choosing at the same time a higher number (as with less iterations) because of the beliefs about the opponents' choices. We also expect to learn something about the decision process in more general terms.

Therefore we are interested in the following category, which we name *sophisticated*. A classification as *sophisticated* indicates that there is more information than given by the chosen number. This classification requests a level-k reasoning process: starting at some number, going step-wise into the direction of the equilibrium and stopping at some number. But then, the subject now goes up again and chooses a number higher than the lowest contemplated. The information given by the level assigned through the chosen number is

missing some information: more steps in the direction of the equilibrium, that is more levels of reasoning, have been made than indicated by the chosen number.

For a classification one could start with a very loose definiton, we simply assess how many subjects used lower numbers in their decision process than the number they state. A comparison of lowest contemplated number and chosen number leads to 64.53% of the decisions that are such that subjects contemplated lower numbers than the number they choose. But satisfying this criterion must not automatically mean that subjects are sophisticated, it could also be that they choose randomly and just for some reason look at a lower number than the number they chose.

To avoid an overclassification with a simple definiton as above we use the following, more rigorous classification rule. Subjects are classified *sophisticated* depending on a precise pattern of eye-tracking data, which is described in the following. In the reasoning process for choosing the number subjects must follow a level-k-type analysis which means that they perform steps leading towards zero, ending at some lowest contemplated number. Instead of stating this number, the subject goes up again and states a number greater than the lowest contemplated number. This indicates that the subject has completed more steps towards the equilibrium or has a higher level than the level interferred from the chosen number.

To be precise, in the eye-tracking data we require that the subject started at some number and went stepwise in the direction of the equilibrium, that is downwards. That means we must have a sequence of fixations starting at higher and going to lower numbers. We also require that the ending points of this downward reaching analysis are lower than the chosen number.

If we use this classification we can classify 21.37% of our observations as *sophisticated*.

Splitting our analysis by the level we calculated merely from the number choice we could conclude that the result for the *sophisticated*-observations was driven mainly by the lower levels. Of the level-0 observations (by chosen number), 19.8% were indeed *sophisticated*, of the level-1 observations 32.9% were *sophisticated*, but of the higher level observations, none could be classified as *sophisticated*.

So far we classified the decisions. It would be desirable to also have a classification on

subject level. We find that 4 of the 39 subjects, that is 10.3% have no decision classified as *sophisticated*, so these can easily be called non-sophisticated. All other have at least one decision classified as *sophisticated*: 23 (59.0%) have one decision, 9 (23.1%) have two and 3 (7.7%) have three, no subject has more than three. But what exactlly does that tell us about a classification as sophisticated on subject level? This question is not simple to answer. It is not clear that a subject should be classified as sophisticated if and only if all six decisions are classified sophisticated. Maybe it is enough to have one sophisticated decision process, and using the insights gained during that decision process in all further decisions? Most of the decisions classified *sophisticated* were decisions in early rounds, to be precise in the first and second round. There are only 3 decisions classified *sophisticated* in later rounds. So it indeed seems that the subjects engage in this sophisticated reasoning once (ore twice) and then in later rounds just apply it and directly stop at the "chosen" level.

5 Conclusion

We used the beauty contest game with eye-tracking to obtain novel data on the decision process. Until now laboratory studies on the beauty contest game have used the chosen number to assign levels of reasoning to subjects or used comments by the subjects. We attempted to classify a subject not only by using the chosen number, but also using the eye-tracking data we recorded while subjects decided which number to choose.

While on a behavioral level (using only the chosen number) we could replicate the finding that a majority of people seem to apply low levels of reasoning, using the eye-tracking data we find that more than 20% of the observations fall into our *sophisticated*-category. In these cases, subjects seem to do more steps of reasoning than indicated by the chosen number. This happens mainly in the first two rounds. Assigning them a "true level" leads to a higher level than the level assigned by using the chosen number. We find these subjects mainly among our seemingly low-level subjects. We therefore conclude that more people are reasoning in a more sophisticated manner than one might think.

In our experiment, as it is standard when categorizing people, some subjects remain

uncategorized. Also other econometric models, e.g. mixture models, find around 30% of random players, see Bosch-Domènech et al. (2010). The fact that in our study subjects remain uncategorized is partially due to the strictness in our categorization. We request a very specific pattern in the eye-tracking data to file an observation in that category. It would have been nice to be able to better understand what drives level-0 behavior, but even with the eye-tracking data we cannot draw clear conclusions, clear patterns do not arise. The reason for that might be that we cannot detect simple heuristics people use with our method (e.g., choosing their birthday or street number). One additional aspect of our results is that most subjects seem to use 50 as the focal point to start with. This can be partially influenced by our method, as the fixation point was in the middle of the screen and thus in the interest area of 50, but on the other hand subjects usually used the parameter in the beginning and from there they could theoretically have been drawn anywhere - even to zero and 100, which we do not find.

Using eye-tracking, we were able to learn more about people's decision processes and their strategic abilities compared to conducting the experiment without eye-tracking. Our results are good news for economists in that they show that people are more strategically sophisticated than behavioral data on the guessing game has suggested so far. It is also good news for those promoting the use of eye-tracking and similar methodology, as by using this method we were clearly able to gain additional insights relevant to economists.

References

Arieli, Amos, Yaniv Ben-Ami, and Ariel Rubinstein, "Tracking decision makers under uncertainty," forthcoming. AEJ-Micro.

Bhatt, Meghana and Colin F. Camerer, "Self-referential thinking and equilibrium as states of mind in games: fMRI evidence," *Games and Economic Behavior*, 2005, *52*, 424–459.

Bosch-Domenech, Antoni, José G. Montalvo, Rosemarie Nagel, and Albert Satorra, "One, two,(three), infinity,...: Newspaper and lab beauty-contest experiments," *American Economic Review*, 2002, *92* (5), 1687–1701.

Bosch-Domènech, Antoni, José G. Montalvo, Rosemarie Nagel, and Albert Satorra, "A finite mixture analysis of beauty-contest data using generalized beta distributions," *Experimental Economics*, July 2010, *13* (4), 461–475.

Breitmoser, Yves, "Hierarchical Reasoning versus Iterated Reasoning in p-Beauty Contest Guessing Games," 2010. MPRA Paper No. 19893.

Bühren, Christoph, Björn Frank, and Rosemarie Nagel, "A historical note on the Beauty Contest," 2009. `https://cms.uni-kassel.de/unicms/fileadmin/groups/w\_030516/BC/A\_historical\_note\_on\_the\_Beauty\_Contest.pdf`.

Burchardi, Konrad B. and Stefan P. Penczynski, "Out of your mind: eliciting individual reasoning in one shot games," 2011. `http://personal.lse.ac.uk/burchark/research/BurchardiPenczynski2011.pdf`.

Camerer, Colin F., *Behavioral Game Theory*, Princeton University Press, 2003.

_ , Teck-Hua Ho, and Juin-Kuan Chong, "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, August 2004, *119* (3), 861–898.

Chen, Chun-Ting, Chen-Ying Huang, and Joseph Tao yi Wang, "A Window of Cognition: Eyetracking the Reasoning Process in Spatial Beauty Contest Games," 2009. `http://homepage.ntu.edu.tw/~josephw/SpatialBeautyContest\_09July14.pdf`.

Coricelli, Giorgio and Rosemarie Nagel, "Neural correlates of depth of strategic reasoning in medial prefrontal cortex.," *Proceedings of the National Academy of Sciences of the United States of America*, June 2009, *106* (23), 9163–8.

Costa-Gomes, Miguel A. and Vincent P. Crawford, "Cognition and behavior in two-person guessing games: An experimental study," *The American economic review*, 2006, *96* (5), 1737–1768.

_ , _ , and Bruno Broseta, "Cognition and Behavior in Normal Form Games: An Experimental Study," *Econometrica*, 2001, *69* (5).

Crawford, Vincent P. and Nagore Iriberri, "Fatal attraction: Salience, naivete, and sophistication in experimental 'Hide-and-Seek' games," *The American Economic Review*, 2007, (5), 1731–1750.

_ and _ , "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?," *Econometrica*, 2007, *75* (6), 1721–1770.

_ , Miguel A. Costa-Gomes, and Nagore Iriberri, "Strategic Thinking," March 2010. `http://dss.ucsd.edu/~vcrawfor/CGCI27Dec10.pdf`.

Ho, Teck-Hua, Colin F. Camerer, and Keith Weigelt, "Iterated dominance and iterated best response in experimental" p-beauty contests"," *American Economic Review*, 1998, *88* (4), 947–969.

Keynes, John Maynard, *The general theory of employment, interest and money*, Vol. VII, Macmillan, 1936.

Knoepfle, Daniel T., Joseph Tao yi Wang, and Colin F. Camerer, "Studying learning in games using eye-tracking," *Journal of the European Economic Association*, 2009, *7* (2-3), 388–398.

López, Rafael, "On p-Beauty Contest Integer Games," 2001. UPF Economics and Business Working Paper No. 608.

Mitzkewitz, Michael and Rosemarie Nagel, "Experimental results on ultimatum games with incomplete information," *International Journal of Game Theory*, June 1993, *2* (2), 195–198.

Moulin, Herve, *Game Theory in the Social Sciences*, New York University Press, 1986.

Mullin, Charles H. and David H. Reiley, "Recombinant estimation for normal-form games, with applications to auctions and bargaining," *Games and Economic Behavior*, 2006, *54* (1), 159–182.

Nagel, Rosemarie, "Interactive Competetive Guessing," 1993. Bonn Working Paper.

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

_ , "A survey of experimental guessing games: a study of bounded rationality and learning," in David V. Budescu, Ido Erev, and Rami Zwick, eds., *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, 1999, pp. 105–145.

Ohtsubo, Yohsuke and Amnon Rapoport, "Depth of reasoning in strategic form games," *Journal of Socio-Economics*, 2006, *35*, 31–47.

Reutskaja, Elena, Rosemarie Nagel, Colin F. Camerer, and Antonio Rangel, "Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study," *The American Economic Review*, 2011, *101* (2), 900–926.

Sbriglia, Patrizia, "Revealing the depth of reasoning in p-beauty contest games," *Experimental Economics*, March 2008, *11*, 107–121.

Stahl, Dale O. and Paul W. Wilson, "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 1995, *10* (1), 218–254.

Wang, Joseph Tao-Yi, Michael Spezio, and Colin F. Camerer, "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *American Economic Review*, June 2010, *100* (3), 984–1007.

Weber, Roberto A., "'Learning' with no feedback in a competitive guessing game," *Games and Economic Behavior*, July 2003, *44* (1), 134–144.

Appendices

A Conduction of a session

Before handing out the instructions to the subject we gave him or her some information about the eye-tracking method, written down on a sheet of paper with illustrations. We informed the subject about the timeline of the eye-tracking session and how we would set up the head mounted system and the camera. Moreover we told the subjects not to move while being eyetracked.

B Instructions

These instructions have been translated into English from the original German.

Please read these instructions carefully and ask the experimenter if you have any questions.

B.1 Part 1

You are taking part in a game, in which you will play along with nine other participants. This game has six rounds. Parts of the instructions you will get on the screen while the experiment is running.

decision In each round you have to choose a number between 0 and 100 (for example $0, 1, 2, 3, \ldots, 54, 55, \ldots,$ up to $99, 100$).

Payoff Your payoff will be determined as follows: in each round the person who choose the number closest to a target number, receives €10. If two persons choose the same number, the prize will be divided equally between these participants. All other participants receive nothing.

Target number During the experiment you will learn how to determine the target number.

Information In between the rounds you will get no information about the outcome of the previous round.

Timeline of one round On the monitor you will see

1. *the instructions.* After reading the complete instructions please press any key to continue.
2. *how to determine the target number.* Here the program proceeds automatically.
3. *a representation of the numbers from 0 to 100.* Please think now which number you would like to choose to get as close as possible to the target number. You do not have any time constraints. When you have decided on the number, state the number and then press any key on the keyboard to continue.

Then you will proceed to the next round. In total there are six rounds.

B.2 Part 2

In this part you will play another game[11], which is unrelated to the first part of the experiment. You will get the instructions for this game during the experiment.

C Questionnaire

About the first part of the experiment (target number)[12]

1. What would you expect that the other participants decided?
2. Did you have a strategy? If so, what was it?
3. Do you have any further comments on the first part?

[11] We report this data separately.
[12] We also had similar questions about the second part of the experiment.

General questions:

Please tell us which grades you received in your Abitur (the German Highschool diploma)

- in Math
- in German
- your Average Grade