Fakultät für Wirtschafts- und Sozialwissenschaften der Ruprecht-Karls-Universität Heidelberg

Contributions to Behavioral and Experimental Economics

Dissertation

zur Erlangung des akademischen Grades
Doctor Rerum Politicarum
vorgelegt von

Julia Müller

Heidelberg, Juli 2012

Acknowledgments

This dissertation would not have been possible without the guidance, encouragement and support of many people.

I would like to thank my supervisor Jörg Oechssler for giving me the possibility to do research at the University of Heidelberg. Jörg opened the door to the exciting world of scientific research, he gave me the freedom to develop my own ideas and allowed me to benefit from his comments, advice, experience, and generous support. I am thankful for his help and confidence in me. I am very grateful to Christiane Schwieren who was more than a second supervisor. She affiliated me to her team, and thereby gave me a second home within the institute. Having Christiane as a co-author is a wonderful experience, she is always willing to discuss with me, and I have benefited a lot from her suggestions, advice and her encouraging support.

The interdisciplinary research training group *Goals and Preferences* provided me not only with financial support, and also with the possibility of insightful discussions with my colleagues from the Institute of Psychology.

Many colleagues have helped to make the time at the University of Heidelberg instructional and enjoyable. I would like to thank all members of the Chair of Economic Theory II and the Chair of Behavioral Economics who have accompanied me over the past years. Special thanks goes to my colleague and co-author Peter Dürsch for many fruitful discussions, to Rosa Huhn and Gabi Rauscher for creating a nice and friendly atmosphere within the team, and to Florian Spitzer for his valuable assistance in the execution of experiments, and furthermore his willingness to help whenever it is necessary.

Contents

[1]This chapter comprises the paper co-authored with Peter Dürsch.

[2]This chapter comprises the paper co-authored with Christiane Schwieren.

[3]This chapter comprises the paper co-authored with Christiane Schwieren.
[4]This chapter comprises the paper co-authored with Christiane Schwieren.

List of Figures

List of Tables

Introduction

Economic theory grounds on the assumption of the *homo oeconomicus*, the rational decision-maker who is only maximizing his own utility. Economic experiments have shown many different, large, and persistent deviations form the ideal of the *homo oeconomicus*. Experimental economics[5] has shown that many subjects care about fairness and the payoff or well-being of others, that they behave reciprocal, i.e. that they reward cooperation and punish defection, even if this involves costs to them. Departing from pure-self interest, behavioral economists try to find new models and assumptions and find themselves in the tension between rational decision-making and emotional or motivational factors guiding human behavior.

Camerer, Loewenstein, and Rabin (2004) write in the introductory chapter of their book on behavioral economics that "most of the ideas in behavioral economics are not new; [...]. Adam Smith, who is best known for the concept of the "invisible hand" and *The Wealth of Nations*, wrote a less well-known book, *The Theory of Moral Sentiments*, which laid out psychological principles of individual behavior that are arguably as profound as his economic observations." (p.5 therein). Indeed inspired by a quote from Adam Smith in *The Theory of Moral Sentiments*, we conducted an experiment on punishment which we present in chapter 1. More precisely, we analyze the underlying motivation to punish people that treated us unfair. Research in experimental economics on punishment found that people punish offenders, even in non repeated situations and even if it is costly. We are interested in looking deeper into the motivation, more precisely into the personal component of punishment. Why do people punish? Why do they want to punish? Is it enough that the offender gets punished or is it also relevant *who* punishes? In a punishment experiment, we separate the demand for punishment in general from the demand to conduct punishment personally. Subjects experience an unfair split of their earnings from a real effort task and have to decide on the punishment of the person who determines the distribution. First, it is established whether the allocator's payoff is reduced and, afterwards, subjects take part in a second price auction for the right to (physically) carry out the act of

[5]For history and overview of experimental economic see Kagel and Roth (1995).

payoff reduction themselves. Subjects bid positive amounts and are happier if they get to punish personally.

Early attempts to go beyond the *homo oeconomicus* were made by Herbert Simon; he coined the term *bounded rationality* (Simon (1955)). One model of boundedly rational behavior is the famous level-k-model which is often used to explain behavior in the beauty contest game. In chapter 2 we present an eye-tracking experiment on the beauty contest game. This 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 this low level thinking – they might be highly sophisticated adaptations to beliefs about other people's limited reasoning abilities.

The last two chapters present different research projects linking personality psychology to economics. In chapter 3 we ask the question what a specific personality questionnaire, namely the Big Five factors, can contribute to explain small-scale economic behavior. Whether personality variables can be used in general to understand micro-behavior in economic games is inevident. Growing interest in using personality variables in economic research leeds to the question whether personality is useful to predict economic behavior. Is it reasonable to expect values on personality scales to be predictive for behavior in economic games? It is undoubted that personality can influence large-scale economic outcomes. However, it is less clear, whether this also holds for micro-behavior in economic games. We discuss reasons in favor and against this assumption and test it in our own experiment, whether and if so which personality factors are useful in predicting behavior in the trust game. We can also use the trust game to understand how personality measures fare relatively in predicting behavior when situational constraints are strong. This approach will help economists to better understand what to expect from the inclusion of personality variables in their models and experiments, and where further research might be useful and needed.

In chapter 4 we use personality questionnaires for a more specific hypothesis. Knowing about huge gender differences in some of the five factors we try to explain with this personality factors the gender gap in women's unwillingness to compete. There is ample evidence that women do not react to competition as men do and are less willing to enter a competition than men. In this chapter, we use personality variables to understand the underlying motives of women (and men) to enter a competition or to avoid it. We use the Big Five personality factors, where especially neuroticism has been related to performance in achievement settings. We first test whether scores on the Big Five are related to performance in our experiment, and second how this is related to incentives. We can show that the sex difference in the willingness to enter a competition is mediated by neuroticism and further that neuroticism is negatively related to performance in competition. These results raise the possibility that those women who do not choose competitive incentives “know” that they should not.

CHAPTER 1

Taking Punishment into Your Own Hands: An Experiment on the Motivation Underlying Punishment[1]

If the person who had done us some great injury, who had murdered our father or our brother, for example, should soon afterwards die of a fever, or even be brought to the scaffold upon account of some other crime, though it might sooth our hatred, it would not fully gratify our resentment. Resentment would prompt us to desire, not only that he should be punished, but that he should be punished by our means, and upon account of that particular injury which he had done to us.
(Adam Smith, *The Theory of Moral Sentiments*, p.113)

1. Introduction

The desire for revenge, to punish those who did wrong upon oneself, is a strong motivation for humans. From ancient Greek dramas to modern movies, it is ubiquitous in story-lines. It has also been the focus of extensive research in economics, both in the form of experiments which find that, indeed, subjects are willing to forgo monetary gains to exert punishment, and in the form of theoretical models that seek to explain such behavior. However, both the quote by Adam Smith above and many prominent fictional works[2] feature a very specific form of punishment: According to Adam Smith, humans not only care about punishment being inflicted on the perpetrator of a crime against them, but they also value carrying out that punishment themselves, in person. It is this, personal, characteristic of punishment that we try to isolate in the laboratory. Our experiment is designed to exclude other possible reasons why one would be willing to give up money to punish. In

[1] This chapter comprises the paper co-authored with Peter Dürsch.

[2] To use two well known movies as examples: In Pulp Fiction, after being rescued from a rapist by Butch, Marsellus tells Butch, who is about to kill the rapist, to move aside, so he can shoot the rapist himself. Similarly, in Dogville, Grace, after ordering her father's men to torch the village which enslaved her, kills the man who hurt her most personally, telling her father: "Some things, you have to do yourself".

particular, subjects do not have to spend money to assure punishment is carried out, they only spend money to assure it is carried out by them personally.

Punishment has been documented in various experiments, especially in social dilemma situations where individual and group incentives diverge and free-riding occurs. One of the first experiments of this kind was conducted by Ostrom, Walker, and Gardner (1992), where subjects who played various rounds in a common pool resource game were willing to pay a fee to place a fine on other subjects who over-extracted the resource. Fehr and Gächter (2000a) demonstrate that costly punishment of free-riders who do not contribute occurs in a public goods experiment, with punishment leading to higher levels of cooperation. Nikiforakis and Normann (2008) analyze the effectiveness of such peer-imposed punishment in a public good game, finding that contributions increase in the effectiveness. In contrast, Falkinger, Fehr, Gächter, and Winter-Ebmer (2000) use punishment imposed "automatically" by the experimenter on non-contributers. Both peer-imposed and experimenter-imposed punishment raises contributions. However, subjects are not only motivated by the monetary consequences of punishment. As Masclet *et.al.* (2003) show, even non-monetary punishment, the expression of disapproval by others, leads to the same result. Masclet *et.al.* (2003) are mainly interested in the receiving side of the punishment, but it is also interesting to investigate the decision process of the punishing side.

Direct neuroeconomic evidence that subjects "like" to punish was found by de Quervain *et.al.* (2004), who use PET recordings of brain activation to investigate the mechanisms in the brain involved in punishment. Subjects played a trust game where cooperating players could punish defecting partners. In the punishment condition activation of the dorsal striatum was found, which is well known for its reward processing properties.

This could either be due to the fact that the defecting partner lost money or it could be pleasure derived from the act of punishing. Based on their finding that subjects do not condition the amount of their own punishment onto the punishment already dealt (to the same person) by other subjects, Casari and Luini (2009), speculate in the same vein that "the punisher derives her utility from the act of punishment in itself and not from achieving, in conjunction with other punishers, a total amount of punishment."

Spurred on by the experimental observation that people do not always act purely selfish, new theories of other-regarding preferences have been put forward, capturing phenomena like fairness, altruism, inequity aversion. Levine (1998) uses an adjusted utility which is supplemented by a term which takes into account the opponent's utility weighted by an altruism coefficient. Inequity aversion models add to the utility derived from own income a term that represents concern about the payoff distribution; Fehr and Schmidt (1999) use the difference between the subject's own payoff and the payoffs of the opponents, Bolton and Ockenfels (2000) the proportion of own payoff to total payoffs.

Other theories develop techniques to embed concerns for reciprocity. Rabin (1993) models reciprocity in normal form games by adding psychological payoffs to the material payoffs. This additional term captures intentions via beliefs of the players and defines the kindness of players in relation to his possible actions. Dufwenberg and Kirchsteiger (2004) dilate this approach to sequential games. Falk and Fischbacher (2006) also transform standard games into psychological games. In their model, utility of the players depends not only on the payoffs but also on a reciprocity term which embodies kindness and reciprocation.

All of these theories incorporate the opponent's outcome into the utility of the player, and several can explain reciprocal behavior or punishment. However, we are not primarily interested in the fact that the payoff of an offender is reduced, but especially in *who* will derive satisfaction from punishing. Only the person who conducted the punishment? Or everyone who saw the offender being punished, even if the punishment was not conducted "personally"?

Perhaps the theory closest to our design is the one by Andreoni (1990). He examines private provision of a public good and models the utility of the individuals as a function not only of the amount of the public good but also of the own gift to the public good. This individual donation produces what Andreoni calls a "warm glow", utility derived from the act of giving. If one assumes in almost the same manner that the act of punishing enters the utility function, one would arrive at a theory that could account for a demand to punish personally.

In the next section, we introduce the design we use to investigate personal punishment. Section 3 presents our hypotheses and Section 4 the results. Finally in Section 5, we conclude with a discussion.

2. Experiment

2.1. Design. To test the demand for personal punishment, we use three related experimental designs, 1A (one auction), 2A (two auctions) and NC (no context).[3] Table 1 shows an overview of the features of the different designs. We start by describing 1A.

2.1.1. *Design 1A.* Subjects were matched in groups of four; each group consists of three subjects A and one subject B. The experiment was anonymous, so no subject knew about the other subjects he or she was matched with. Instructions for the experiment, which fully

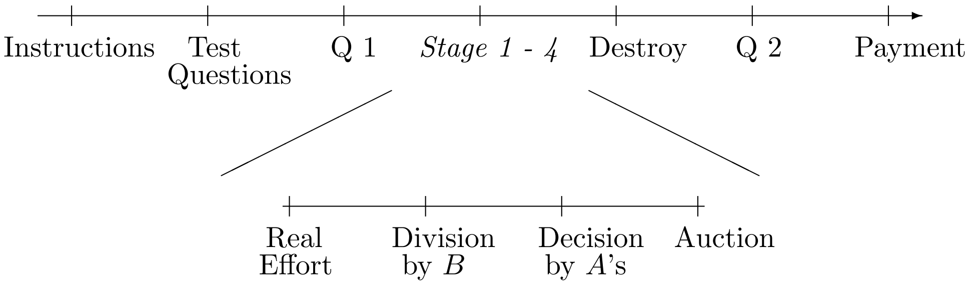
[3]See appendix (section 6) or online-appendix for translations of all instruction material: http://www.uni-heidelberg.de/md/awi/professuren/with2/pdjm-pp-appendix.zip

TABLE 1. Experimental designs

	real effort	punishment opportunity	auction for personal punishment	auction for dummy envelope
1A	yes	yes	yes	no
2A	yes	yes	yes	yes
NC	no	no	no	yes

described the experiment for both type A and type B, were handed to subjects at the start of the experiment. After reading the instructions, subjects had to answer a series of detailed questions in order to make sure that they understood the experimental instructions. Only when all subjects had correctly answered these test questions, did the experiment proceed.[4]

FIGURE 1. Timing



In the first stage,[5] all subjects A participated in a real effort task where they could earn 10 €. They were asked to fill a sheet of graph paper (A5, 148×210 mm, about 1260 squares) with alternating o and + signs. The allocated time frame was 25 minutes. Subjects who did not finish the task in time dropped out of the experiment and received no money apart from the show up fee. We chose this particular task for two reasons: First, it is simple and does not require any special abilities, so all subjects should be equally fit for the task. Second, as we found out in previous tests, the task is considerably more exhausting than it appears. We wanted to induce a feeling of ownership towards the money in those subjects who completed the task. On the other hand, it was to look easy to the non-participating subjects B. During

[4]Subjects who were not able to answer the test questions correctly were replaced by extra participants (who were otherwise dismissed with a flat payment after reading the instructions).

[5]The instructions use a different numbering, since we subdivided some stages for clarity. We also handed to all subjects a flow chart as an overview what happens in each stage. The flow charts are included in the appendix (section 6) and the Online-Appendix.

the task, all subjects B were sitting in the same room as the subjects A, but without any assignment.

After the task, the experimenters collected the sheets and informed each subject B how many subjects A in her group had succeeded. Upon learning that information, in stage two, subject B had to decide on an allocation of the money earned by the subjects A in the previous stage. The only two allocations available were (2,8): 2 for A, 8 for B, or (10,0): 10 for A, 0 for B. Subject B could only implement the same allocation for all three subjects A she was matched with, not different allocations for different subjects A. So in the case of three successful subjects A subject B had to decide between 24 for herself and 2 for each A or 0 for herself and 10 for each A.

Before stage three, the experimenters informed all subjects A about the decision of their matched B. The money that subject B allocated to A was handed to subject A. The money that subject B allocated to herself was *also handed to subject* A, however it was put in an envelope. Then all subjects A had to decide whether they wanted to reduce subject B's payoff by destroying one of the three envelopes designated for B. If all A's decided not to reduce, the envelopes were collected by the experimenters, handed to subject B and stage four did not take place.

If at least one subject A decided to reduce, the entire group entered stage four. Here, all subjects A of the group took part in a sealed bid second price auction. The highest bidder won the right to destroy the envelope lying in front of him. Only the envelope of the winner was destroyed.[6] Note that subjects B's payoff depends entirely on stages 1 to 3. The auction only selected the subject A who would destroy the envelope, it did not affect subject B's payoff. The auction provides a non-arbitrary way to separate the decision to punish from the decision to punish personally. Since, in a second price auction, no participant has a reason to misrepresent his preferences, subjects are incentivized to truthfully state the value they attach to personal punishment.

The auction winner was informed that he won the auction and about the second highest bid he had to pay. He could then proceed to destroy the envelope of subject B. The instructions did not specify any mode of destruction; however a small metal bin was present on the tables of each subject A.[7] The envelopes in front of the non-winning subjects A were collected by the experimenters and delivered to the respective subject B.

[6] The minimum feasible bid was zero, the maximum feasible bid 10 and subjects could bid in increments of 0.01 (one cent). If there was a tie, the experimenters randomly chose a winner. This also applies to the special case of all three subjects A choosing a bid of zero.

[7] The subjects chose different methods to "destroy": Most ripped the envelope apart – some ripped just once, some ripped until only small pieces of paper were left – and deposited the pieces inside the metal bin. Some just folded the envelope.

Between the test questions and the real effort task we asked some demographics from our subjects and two questions about their happiness ("how happy are you in general"/"how happy are you right now"). After stage four and before paying, we presented subjects with a second questionnaire asking their happiness again (only "how happy are you right now"), their perception of subject B's behavior and several attitude questions[8]. All subjects received a 8 € "show up fee" for answering the questionnaires. If a subject A had won the auction and had to pay more money than he earned in the experiment, he had to use a part of those 8 € to pay for his bid.

2.1.2. *Design 2A.* The 2A design is similar to design 1A, with the difference that it uses two auctions instead of one. Stages one to three are identical to 1A. However, in the auction stage, subjects had to make two bids. Bid one was for the auction as described above. For the second auction, the experimenters placed a second envelope in front of the each subject A. The instructions stated that this envelope would be, unless destroyed, collected again by the experimenters and would never have any influence on the payoff of subjects A or subjects B. That is, the second envelope is a dummy, intended to test whether subjects would be willing to pay for destruction of *any* envelope. After bids were made, the experimenters threw a coin in public to determine whether auction one or auction two would be enacted. Only the bids from the chosen auction did count, and only the envelope from the auction chosen was destroyed by the winning subject A. If auction one was chosen, the winner destroyed his envelope from auction one, the other envelopes where handed to subject B, and all three envelopes from auction two were collected by experimenters. If auction two was chosen, the winner destroyed his envelope from auction two, the other envelopes from auction two where collected by the experimenters. In this case, the experimenters also randomly retained one of the envelopes from auction one, such that subject B only received the same amount of envelopes, no matter whether auction one or auction two was chosen by the coin-flip.

2.1.3. *Design NC.* Finally, we used the NC condition to separate the auction stage from the rest of our experiment. To insure that conditions remained comparable, we conducted the control subsequent to another, unrelated and about 1 hour long, experiment, where the subjects earned on average 10.60 €.[9] This money was used to pay for bids in the control auction. After the end of the other experiment, we distributed the instructions for NC. Instructions and test questions

[8]See appendix (section 6) or Online-Appendix.

[9]This is close to the average earnings of 10.81 € that subjects of type A had accumulated in the other conditions (2A and 1A) before the auction was conducted.

were as close as possible to those in the main experiment, but only included the auction stage.[10]

Subjects were placed in groups of three (corresponding to our group size of three subjects A, who did participate in the auction). The highest bidder won the right to destroy an envelope lying in front of him (the envelope was not payoff relevant, as in auction two of 2A). The winner of the auction could destroy the envelope, all others were collected by the experimenters. Auction winners were paid what they earned in the prior experiment minus the second highest bid in their group.

2.2. Procedures. The experiment was conducted in the laboratory of the economics department at the University of Heidelberg and the laboratory of the Sonderforschungsbereich 504 in Mannheim. In total, 149 subjects participated in the experiment (40% male, 60% female). Subjects were students of various fields at the University of Heidelberg and the University of Mannheim. The experiment consisted of nine sessions; no subject participated twice. All recruitment was done via ORSEE (Greiner (2004)).

In total, the experiment lasted slightly less than 2 hours, for which we paid an average of 13.79 €.[11] The full experiment was conducted via pen and paper. During the experiment, we used an experimental currency unit called "Thaler". Thaler were a printed play money handed to subjects during the experiment. At the end of the experiment, we exchanged all Thaler into Euro at a rate of 1:1.[12] All subjects were paid in cash and private.

3. Hypotheses

According to the classic and outcome based social preference theories subjects should not care about the way in which subject B's payoff is reduced. On the other hand, following the reasoning put forward by Adam Smith, subjects should care about punishing personally, so we formulate our main hypothesis:

Hypothesis 1. *Personal punishment: Subjects A bid more in the personal punishment auction than in the dummy auction.*

Connected to hypothesis 1 we would also expect those subjects who punish personally to have some emotional payoff from doing so that makes their monetary loss worthwhile.

[10] We used both envelopes filled with paper money and empty envelopes (the unrelated prior experiment did not use paper money), but did not find any difference and pooled the data.

[11] Only averaging over subjects in 1A and 2A.

[12] The main reason for using play money was that we did not want subjects to worry about destroying legal tender.

TABLE 2. Bids

	Avg.(SD)	Max	> 0	$= 0$
Bid punishment auction	0.43 (1.11)	5.5	36%	64%
Wanted auction	0.51 (1.24)	5.5	46%	54%
Did not want auction	0.32 (0.93)	4.0	22%	78%
Bid dummy auction	0.03 (0.10)	0.5	17%	83%
Bid in NC auction	0.67 (1.78)	6.5	52%	48%

Row 1 to 3: subjects A's bids in 1A and in the punishment auction in 2A. Row 4: bids in dummy auction in design 2A. Row 5: bids in design NC.

HYPOTHESIS 2. *Happiness: Subjects A who punish personally are happier than those who do not.*

4. Results

Stages one and two of our auction designs were constructed to produce a large number of observations where punishment could possibly occur. A first look at the data confirms that this goal is achieved. All 87 subjects A in 1A and 2A did complete the real effort task, therefore all 29 subjects B had to make their decision for 3 matched successful subjects A. Out of the 29 subjects B, all but 3 did implement the allocation $(2, 8)$, which was worse for subjects A. All three subjects B implementing $(10, 0)$ played in design 1A.[13]

Trying to find personal punishment is only viable if there is some punishment in the first place. Given the allocation of their matched subject B, all subjects A could chose to have the auction in stage four implemented. Demanding the auction is equivalent to subject B being punished, since this ensures that subject B's payoff will be reduced by 8. In line with our expectations, subjects A who faced the "bad" (2,8) split demand the auction significantly more often than those who got the "nice" $(10, 0)$ allocation ($p = 0.040$, one-sided Fisher-exact test). In total, 55% of subjects A demanded the auction. Since the auction is implemented if at least one subject A demands it, this translates into the auction happening in 26 out of 29 groups.[14]

Table 2 shows the percentage of subjects A who bid a positive amount in the ensuing auction - split into those who demanded punishment in the previous stage (that is, who wanted the auction to happen) and those who did not. Recall that bids in the auction are not

[13]The distribution choice of subjects B is similar to the one in a dictator game (Kahneman, Knetsch, and Thaler (1986)) or ultimatum game (Güth, Schmittberger, and Schwarze (1982)) with a restricted choice domain.

[14]The three groups missing are not equivalent with the three groups where the $(10, 0)$ decision was implemented. While demand for the auction was lower in these three groups, 2 subjects still wanted the auction to happen. The third $(10, 0)$ group, as well as 2 out of the 26 $(2, 8)$ groups did not see the auction happen.

payoff relevant for subject B, only whether the auction happens or not influences the payoff of subject B. Subjects A who are either strict money maximizers or only interested in the monetary consequences of punishment for the matched subject B have no incentive to bid larger than zero. In contrast to that, we find that 36% of our subjects bid positive amounts of money. So a substantial minority of subjects is interested enough in punishing personally to be willing to sacrifice some of their own money to achieve this. While it is somewhat surprising that we also find some positive bids of subjects who did not want the auction to happen in the previous stage, the average bid by subjects who wanted the auction is significantly higher ($p = 0.021$, one-sided MWU test).

The positive bids in the punishment auction indicate that our subjects want personal punishment, but a better test for the existence of personal punishment is to compare the results for the two auctions in design 2A. Here, within subject, are two identical auctions, leading to a similar results (an envelope gets destroyed and subject B loses a payoff of 8). The only difference is whether subjects get to destroy an unrelated envelope or the envelope belonging to subject B. A Wilcoxon Signed Ranks test shows that bids are significantly higher in the punishment auction compared to the dummy auction ($p = 0.016$, one-sided).[15] Therefore we can not reject hypothesis 1.

RESULT 1. *Subjects bid more in an auction for personal punishment than in a dummy auction, in line with a demand for personal punishment.*

Another interesting comparison is between the auctions in 1A and 2A, where the auctions are embedded in a comprehensible context, to the auction in NC, where we remove the context. The bids in NC are not different from those in the auction for personal punishment (p=0.235, MWU test, two-sided), but significantly higher than those in the dummy auction of 2A ($p = 0.006$, MWU test, two-sided). Ex ante, we would have expected the differences to be the other way round. This points out the importance of giving subjects a context in which to evaluate the auction. Without the proceeding stages, the auction must

[15]While very infrequent, there is some bidding in the dummy auction. The answers from the subject with the highest bid in the dummy auction to an open ended question about motivation for bidding are perhaps revealing:
(personal punishment auction): "Even though subject B is in no way affected (since he always gets 2 envelopes), it feels good to release some pressure this way"
(dummy auction): "To erase the feeling of anger, that, even though I did the whole work, candidate B will earn 3x as much"

TABLE 3. Regression on happiness difference

	regression 1 punishment auctions (1A, 2A)		**regression 2 dummy auction (NC)**	
	coefficient	p-value	coefficient	p-value
age	-0.006	0.889	-0.003	0.938
	(0.044)		(0.036)	
female	-0.301	0.269	0.494	0.210
	(0.270)		(0.385)	
$(10, 0)$ distribution	1.150	0.019		
	(0.479)			
wanted auction	-0.108	0.674		
	(0.256)			
bid	0.048	0.689	-0.162	0.244
	(0.120)		(0.136)	
auction winner	0.659	0.022	0.755	0.092
	(0.281)		(0.432)	
constant	0.020	0.985	0.086	0.934
	(1.039)		(1.024)	
N	78		33	
R^2	0.195		0.119	
adj. R^2	0.127		-0.007	

Notes: Dependent variable: happiness difference. Standard errors in parentheses. Bid: Regression 1 uses the bids from the punishment auction (1A and 2A), regression 2 from the dummy auction (NC). In both sessions of 2A, the coin flip chose the punishment auction, therefore the punishment auction was resolved and the data is used in regression 1.

have made little sense to subjects in NC.[16] Perhaps they were confused, perhaps they (incorrectly) rationalized the existence of the auction with some not-yet-announced price that would be revealed afterwards, or maybe they felt forced to bid in the absence of any explanation. If this experimenter demand effect (see Zizzo (2010)) exists, it is present in our no context treatment, but not in the (similarly non-consequential) dummy auction of treatment 2A.

Finally, we look at the result of the physical destruction carried out by the winners of the auction. Do they enjoy the act of destroying subject B's money? We asked all participants for their subjective

[16] In all designs, subjects had to correctly answer a set of test questions before the experiment proceeded. However, the test questions only related to the mechanism of the auction (and the previous stages for 1A and 2A), not any possible rational behind holding it.

happiness on a seven point scale at the start and at the end of the experiment.[17] While the absolute level might depend on a number of causes we can not control, we can use the difference in happiness between the start and end of the experiment. Let the *happiness difference* be the amount of happiness reported at the end of our experiment minus the amount reported at the start. So subjects with a positive happiness difference felt better after our experiment than before. Table 3 reports two regressions on happiness difference. Not surprisingly, subjects A who encountered the allocation (10,0) felt happier compared to those who received only 2 € from allocation (2,8). Additionally, subjects A who went on to win the auction are happier than those who did not win. So despite being paid less money in the end, subjects who personally destroyed subject B's money leave the experiment happier that those who do not, in line with hypothesis 2. The right side of table 3 reports a similar regression, now run for the subjects in design NC. Here, winning the auction only has a weakly significant effect,[18] but the coefficient is larger.

Result 2. *Subjects who won the auction for personal punishment are happier than those who did not.*

5. Discussion

In an experiment designed to separate the decision to punish personally from the more general decision to punish, we find that many subjects bid positive amounts in a second price auction that auctions off the right to punish personally. Some of these subjects bid substantial amounts.

The experimental designs are constructed to eliminate a range of other effects, which might have an influence on subjects decisions in more general settings. Due to the one-shot nature of the experiment, it is not possible to use bids as a signaling device for future play. Furthermore, seats in the experiment were separated by blinds, so the act of punishing was hard to use to express disapproval as in Masclet *et.al.* (2003). Since punishment is the physical act of destroying (paper) money, it might be a worry that subjects like to destroy money. However, the results of our questionnaire let us discard that thought.[19] The act of destroying the envelope is a punishment of subject B, not money burning as in Zizzo (2003), where no strategic component was involved.

[17]See appendix (section 6) or Online-Appendix for the translated questionnaires.

[18]Obviously, subjects in NC did not see allocations and did not decide on conducting the auction either.

[19]The final questionnaire included the question "Do you like destroying money?". Not one of the subjects answered with yes. Additionally, subjects were given the opportunity to destroy some of their own remaining money during the final questionnaire. Again, none took this opportunity.

Most importantly, the bids in the auction, and thus the willingness to pay for personal punishment, have no influence on the payoff of the offending subject B. Subject B's payoff is completely determined in stages 1 to 3. One of the mandatory test questions covered this point to make it clear to every subject. Our decision to use a second price sealed bid auction stems from the previous considerations. It is a fast and incentive compatible method that lets us elicit a very fine grained willingness to pay for personal punishment. Since the auction always has a winner, it emphasizes the point that punishment will always occur, regardless of the bids of subjects A.

Using an auction might introduce a motivation to bid due to a "desire to win". Van den Bos *et.al.* (2008) find evidence for this in a sealed bid first price auction. In one of their treatments, the opponents are other human subjects (similar to our NC design), while in two other treatments, subjects bid against computerized agents. Furthermore, all subjects are taught to calculate the (risk-neutral) Nash-equilibrium strategy, to rule out a winner's curse effect stemming from limited cognitive ability. They find that subjects playing against humans overbid significantly more often than those playing against computers. There is also evidence from a fMRI experiment by Delgado *et.al.* (2008) who compare subjects' reactions to losing a lottery versus losing an auction to conclude that "The fear of losing the social competition inherent in an auction may lead people to pay too high a price for the good for sale". It is possible that, in a similar vein, our subjects did not want to "lose" the auction and therefore bid positive amounts. Our results in NC can be seen as further evidence for such an effect. However, in 2A, we directly compare the results of two auctions. If a desire to win exists, it should influence both auctions in a similar way, yet we find a significant difference between the two.

We further find that subjects who win the personal punishment auction are becoming happier during the experiment compared to those who do not win. A similar result for the dummy auction is only weakly significant. While we can not exclude the possibility that subjects happiness is only due to winning the auction, the result is also consistent with subjects enjoying the personal punishment they achieved.

The personal punishment we address in this paper differs from antisocial punishment as in Herrmann, Thöni, and Gächter (2008), which is punishing people that behaved pro-socially. In our case, when subjects B decided on the distribution, they (mostly) chose the unfair $(2, 8)$-split; they therefore do not behave prosocial. When we look for antisocial punishment in our data, we find that only 2 out of 9 subjects (22.2%), who were confronted with the fair or prosocial $(10, 0)$-split, voted for punishment.

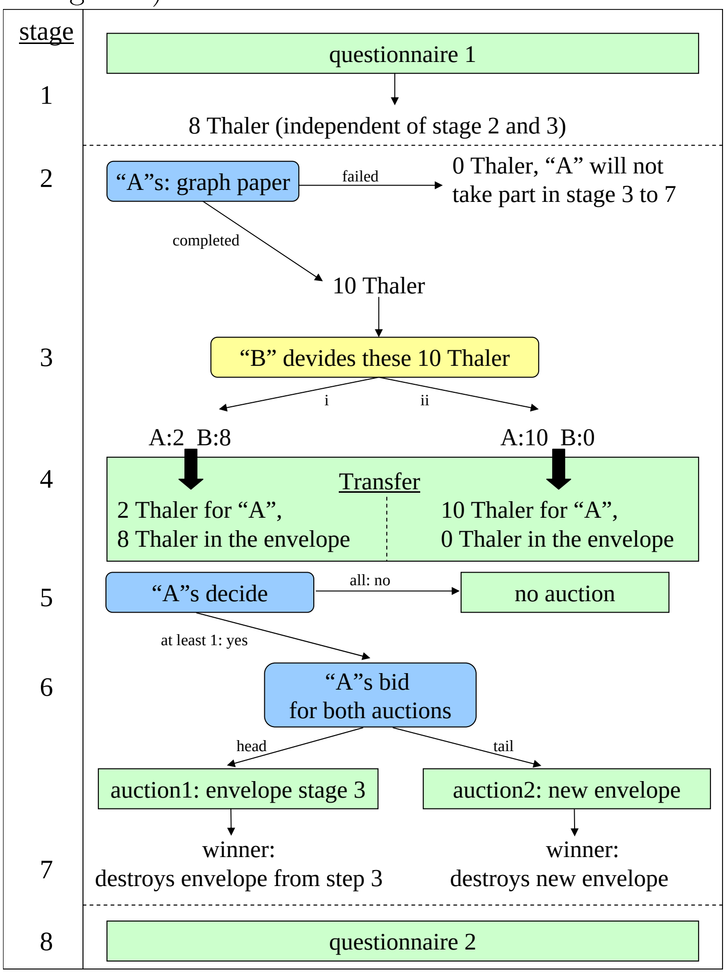
Overall, the effects we observe are significant, but not huge. This is not surprising, since we exclude many other effects which would otherwise work in a similar direction. In many real life examples, the demand for punishment and the demand for personal punishment will be measured simultaneously. Additionally, the personal punishment, as Adam Smith describes it, is punishment for a grave offense. For obvious reasons, laboratory experiments can only implement minor offenses.

Yet modern justice systems might be one of the causes of unfulfilled demand for personal punishment. By moving all aspects of punishment into the hands of state employees and professionals, they remove part of the satisfaction from punishment on part of those who were done wrong. The many advantages of modern justice are obvious, but our paper might point out a hidden disadvantage.

6. Appendix

6.1. Instructions. These instructions have been translated into English from the Original German.

FIGURE 2. Overview Design 2A (handed to all subjects in design 2A)



6.1.1. *Instructions 2A.* Welcome to this experiment! Please read these instructions[20] carefully. From now on, do not talk to your neighbors. Please switch your mobile phone off and keep it switched off until the end of the experiment. If you have any questions, raise your hand and we will come to your seat.

In this experiment, each participant will be assigned one of two roles: A or B.

Three A and one B will be matched with each other. No participant will ever learn the identity of the other participants.

In the experiment, *Thaler* will be used as the experimental currency unit. At the end of the experiment, all paper Thaler will be exchanged into Euro at a rate of 1 Thaler = 1€. Each participant will be paid privately and in cash. Your payout is determined by your decisions during the experiment and by the decisions of the participants you are matched with.

Details of the experiment

Stage 1: Questionnaire Please complete the distributed questionnaire. In return, you will earn 8 Thaler.

Stage 2: Graph paper Each participant A is handed one sheet of graph paper and a pen. The task is to fill that sheet with alternating + and − signs, as shown in the first lines. Each A will have 25 minutes to complete this task. If the sheet is completely filled, A receives 10 Thaler. These Thaler will be handed out by the experimenters. If the sheet is not completely filled, A earns 0 Thaler, and does not take part in stages 3 to 7.

Stage 3: Decision of B Now, the experimenters place in front of B the 10 Thaler of each of the three matched A. B decides on an allocation of the Thaler received by the A in stage 2 between himself and the A. There are two possible allocations. If an allocation is chosen, it is implemented for all A. It is not possible for B to choose different allocations for different A.

Allocation (2,8): 2 Thaler for A and 8 Thaler for B
Allocation (10,0): 10 Thaler for A and 0 Thaler for B

If B chooses allocation (2,8), he places 8 Thaler in each envelope and leaves 2 Thaler outside of the envelope. In this case, each A receives those 2 Thaler. The 8 Thaler in the envelope are designated for B. If B chooses allocation (10,0), he places 0 Thaler in each envelope and leaves 10 Thaler outside of the envelope. In this case, each A receives those 10 Thaler. No Thaler are designated for B.

Stage 4: Transfer Now, the envelope and the Thaler outside the envelope are placed in front of each A – as per the allocation chosen by B.

[20]Subjects in the experiment received the instructions in another, clearly arranged format and page layout than the one displayed here.

- Each A can take the Thaler outside the envelope and combine them with the 8 Thaler he earned in stage 1.
- The envelope must not be opened. The envelope contains those Thaler, which B designated for himself.

Stage 5: Decision of As Each A decides on the following question: "Should one of the envelopes of stage 3 be destroyed?"

If B chose the allocation (2,8) in stage 3 and if at least one A answers with yes, then Bs payout will be reduced by 8 Thaler. If all A answer the question with no, then stage 6 (auctions) does not take place, and the payout of B is not reduced.

Stage 6: Auctions The experimenters distribute a new envelope to each A. This envelope contains the same number of Thaler as are in the envelope of stage 3. Neither A, nor B will ever receive any Thaler out of the new envelope.

There are two auctions, auction1 and auction2. Auction1 relates to the envelope of stage 3, auction2 relates to the new envelope.

Each A has to make a bid for both auctions. Only one of the two auctions will be implemented. A coin toss will decide which auction will be implemented. The coin toss will be made after bids have been collected.

This is how both of the auctions work:
Each A states the number of Thaler he is willing to bid (you can also bid in Cent). The lowest possible bid is 0 Thaler, the highest possible bid is 10 Thaler.
The A who entered the highest bid wins the auction. However, the winner only has to pay the second highest bid in this auction. This cost will be deducted during the payout at the end of the experiment. If there are several, equally high, highest bids, the winner will be randomly determined. This means that there is always a winner in the auction.
Note: In this type of auction, it is optimal to bid just the amount that is equivalent to your valuation of the good (here: the right to destroy the according envelope) that is auctioned off.

In each of the two auctions, the winner earns the right to destroy one envelope:

In auction1, each A bids for the right to destroy the envelope in front of them from stage 3. Only the winner of the auction may destroy the envelope. B does not receive any Thaler out of the destroyed envelope of the winner.

In auction2, each A bids for the right to destroy the newly distributed envelope. Only the winner of the auction may destroy the envelope. Newly distributed envelopes which are not destroyed are collected again by the experimenters. Nobody does ever receive any

Thaler out of the destroyed or not destroyed newly distributed envelopes.

Stage 7: Result of the auction In this stage, a coin toss determines whether auction1 or auction2 will be resolved. The winner of the chosen auction may destroy his envelope and the Thaler in that envelope.

If the coin toss chooses auction1, the two envelopes of stage 3 which are not destroyed are handed to B. B may open the envelopes and take the Thaler within them. All newly distributed envelopes are collected by the experimenters.

If the coin toss chooses auction2, two of the envelopes of stage 3 are handed to B. B may open the envelopes and take the Thaler within them. One randomly chosen envelope of stage 3 is retained by the experimenters and not handed to B. All newly distributed envelopes which are not destroyed are collected by the experimenters.

Stage 8: Questionnaire Finally, please complete the second distributed questionnaire.

Payment Now, all Thaler are exchanged into Euro.

Payout A: All A own the Thaler placed in front of them. That is, the 8 Thaler received in stage 1 as well as the Thaler received in stage 4. The winner of the auction that was chosen by the coin toss has to pay the second highest bid in this auction.

Payout B: B owns the 8 Thaler received in stage 1 and all Thaler out of the envelopes of stage 3, with exception of any destroyed envelope or any envelope retained by the experimenters.Nobody receives Thaler out of the newly distributed envelopes from stage 6.

6.1.2. *Instructions 1A.* Welcome to our experiment! Please read these instructions carefully. From now on, do not talk to your neighbors. Please switch your mobile phone off and keep it switched off until the end of the experiment. If you have any questions, raise your hand and we will come to your seat.

In this experiment, each participant will be assigned one of two roles: A or B.

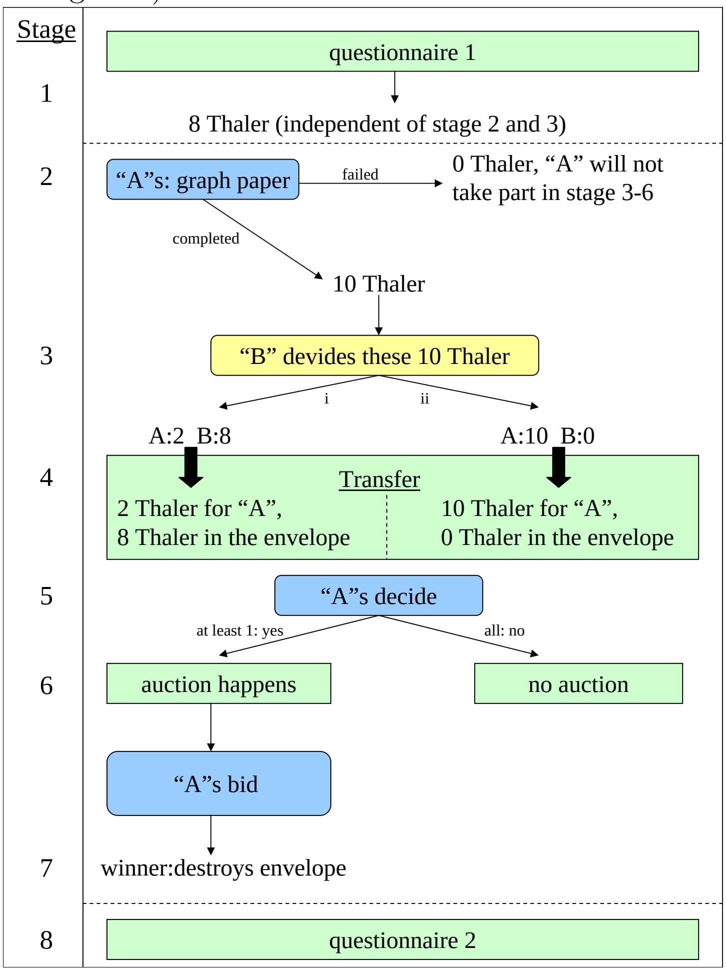
You are in the role of A [B][21] *for the entire experiment.*

Three A and one B will be matched with each other. No participant will ever learn the identity of the other participants.

In the experiment, *Thaler* will be used as the experimental currency unit. At the end of the experiment, all paper Thaler will be exchanged into Euro at a rate of 1 Thaler = 1€. Each participant will be paid privately and in cash. Your payout is determined by your decisions during the experiment and by the decisions of the participants you are matched with.

[21] Subjects A and B received the same instructions, only at this position stating their role.

FIGURE 3. Overview Design 1A (handed to all subjects in design 1A)



Details of the experiment

Stage 1: Questionnaire Please complete the distributed questionnaire. In return, you will earn 8 Thaler.

Stage 2: Graph paper Each participant A is handed one sheet of graph paper and a pen. The task is to fill that sheet with alternating $+$ and $-$ signs, as shown in the first lines. Each A is allowed to spend maximal 25 minutes to complete this task.

If the sheet is completely filled, A receives 10 Thaler. These Thaler will be handed out by the experimenters. If the sheet is not completely filled, A earns 0 Thaler, and does not take part in stages 3 to 7.

Stage 3: Decision of B B decides on an allocation of the Thaler received by the A in stage 2 between himself and the A. There are two possible allocations. If an allocation is chosen, it is implemented for all A.

i) 2 Thaler for $A$and 8 Thaler for B

ii) 10 Thaler for $A$and 0 Thaler for B

In the first case each A receives 2 Thaler. B receives 8 Thaler for each A who still takes part in stage 3 (that means 24, 16, 8 or 0 Thaler when 3, 2, 1, 0 A's are still participating).

In the second case, each A receives 10 Thaler and B receives 0 Thaler in total.

Stage 4: Transfer The experimenters allocate the Thaler according to the decision of B. Each A receives:

- The Thaler that B allocated to him.
- The envelopes must not me opened. The envelope contains those Thaler, which B designated for himself.

Stage 5: Decision of A's Each A decides on the following question and notes this on decision sheet A:

"Should one of the envelopes be destroyed?"

In case of allocation i) this will reduce Bs payout by 8 Thaler.

If all A answer the question with no, then stage 6 (auctions) does not take place, and the payout of B is not reduced.

Stage 6: Auction All three, with the exception of those who dropped out in stage 2, take part in this auction. Out of the three envelopes exactly one will be destroyed, two will remain. Each A can bid for the right to destroy his own envelope with the included money which B would receive from him. Only the winner of the auction may destroy the envelope. B will not receive any Thaler out of the envelope of the winner.

This is how the auction works: Each A states the number of Thaler he is willing to bid on decision sheet A (minimum 0 Thaler, maximum 10 Thaler, step size 0.01 Thaler). The A who entered the highest bid wins the auction and obtains the right to destroy his envelope. However, the winner has to pay the second highest bid. This cost will be deducted during the payout at the end of the experiment. There will always be a winner of the auction. If there are several , equally high, highest bids, the winner will be randomly determined.

Note: In this type of auction, it is optimal to bid just the amount that is equivalent to your valuation of the good (here: the right to destroy the envelope) that is auctioned off.

Stage 7: Result of the auction The winner of the auction may now destroy his envelope and the Thaler in that envelope in arbitrary manner. Afterwards, the envelopes of those A who did not win the auction will be handed to B. B may open the envelopes and take the Thaler within them.

Stage 8: Questionnaire Finally, please complete the second distributed questionnaire.

Payment Now all Thaler are exchanged into Euro.

All A own the 8 Thaler received in stage 1 as well as the Thaler received in stage 4. The winner of the auction has to pay the second highest bid in the auction.

B owns the 8 Thaler received in stage 1 and all Thaler out of the envelopes of stage 3, with exception of the destroyed envelope.

6.1.3. *Instructions NC.* In this experiment, you are, together with two other participants, in a group of three people. No participant will ever learn the identity of the other participants.

In front of every participant is an envelope. In this experiment, the right to destroy this envelope is auctioned off.

Auction All three participants take part in this auction. Out of the three envelopes exactly one will be destroyed, two will remain. Each participant can bid for the right to destroy *his own* envelope. Only the winner of the auction may destroy the envelope.

This is how the auction works: Each participant states the number of Euro he is willing to bid on the decision sheet (minimum 0 Euro, maximum 10 Euro, step size 0.01 Euro). The participant who entered the highest bid wins the auction and obtains the right to destroy his envelope. However, the winner has to pay the second highest bid. This cost will be deducted during the payout at the end of the experiment. There will always be a winner of the auction. If there are several, equally high, highest bids, the winner will be randomly determined.

Note: In this type of auction, it is optimal to bid just the amount that is equivalent to your valuation of the good (here: the right to destroy the envelope) that is auctioned off.

Result of the auction The winner of the auction may now destroy his envelope in arbitrary manner. Afterwards, the envelopes of those participants who did not win the auction will be collected by the experimenters.

Payment The winner of the auction has to pay the second highest bid. All other participants pay nothing.

6.2. Test Questions.

6.2.1. *Test Questions 2A.* **Question 1:** What payment will you receive at the end of the experiment, if you are A and you do not manage to fill out the complete graph paper.

Question 2: As A, you are bidding 2 Thaler in one of the auctions. The second A bids 0 Thaler and the third A bids 3.40 Thaler.

a) Who wins the auction and may destroy an envelope?
b) How much does the winner have to pay?

Question 3: Assuming all As were successful in stage 2 and B did decide on the allocation $(2, 8)$ in stage 3. Look at stage 5 and 6. What is the only case in which the payout of B is not reduced by 8 Thaler?

Question 4: You are B. All A did fill out the complete paper in stage 2 and you did decide on the allocation $(2, 8)$. The As decide they want the auction. In the auction chosen by the coin flip, the As are bidding exactly as in question 2. What is your payout at the end of the experiment?
Question 5: You are B. 2 out of 3 As did fill out the complete paper in stage 2 and you did decide on the allocation $(10, 0)$. All As decide against the auction.

a) What payout will you receive at the end of the experiment?
b) What payout will those As who completed the paper receive?
c) What is the payout of the A who did not complete the entire paper?

Question 6: Assume that auction 1 is chosen by the coin flip. Which A has to pay something? Which bid does this A have to pay?

6.2.2. *Test Questions 1A.* **Question 1:** What payment will you receive at the end of the experiment, if you are A and you do not manage to fill out the complete graph paper.
Question 2: As A, you are bidding 2 Thaler in the auction. The second A bids 0 Thaler and the third A bids 5 Thaler.

a) Who wins the auction and may destroy the white envelope?
b) How much does the winner have to pay?

Question 3: Assuming all As were successful in stage 2 and B did decide on the allocation 2 Thaler for each A and 8 Thaler for B in stage 3. Look at stage 5 and 6. What is the only case in which the payout of B is not reduced by 8 Thaler?
Question 4: You are B. All A did fill out the complete paper in stage 2 and you did decide on the allocation (2,8). The As decide they want the auction. In the auction chosen by the coin flip, the As are bidding exactly as in question 2. What is your payout at the end of the experiment?
Question 5: You are B. 2 out of 3 As did fill out the complete paper in stage 2 and you did decide on the allocation 10 Thaler for each A and 0 Thaler for B. All As decide against the auction.

a) What payout will you receive at the end of the experiment?
b) What is the payout those As who completed the paper will receive?
c) What is the payout of the A who did not complete the entire paper?

6.2.3. *Test Questions NC.* **Question 1:** You are bidding 2 Euro in the auction. A second participant bids 0 Euro and the third participant bids 5 Euro.

a) Who wins the auction and may destroy the envelope?

b) How much does the winner have to pay?

Question 2: Can it happen that in a group no one of the three participants destroys his or her envelope?

Question 3:

a) Assume you bid 0 Euro in the auction. What payout will you receive for this part of the experiment?
b) Assume you bid 1,50 Euro in the auction and your bid is the highest bid, the second highest bid is 1 Euro What payout will you receive for this part of the experiment?

6.3. Questionnairs.

Questionnaire 1[22]

How happy are you in general?

(very unhappy) o o o o o o o (very happy)

How happy are you at the moment?

(very unhappy) o o o o o o o (very happy)

How old are you?

What is you gender?

Are you a student?

If yes: What is you major?

Questionnaire 2[23]

How happy are you at the moment?

(very unhappy) o o o o o o o (very happy)

[ONLY ROLE A] How did you perceive Bs behavior in stage 3?

(not fair)	o	o	o	o	o	o	o	(fair)
(not nice)	o	o	o	o	o	o	o	(nice)
(not comprehensible)	o	o	o	o	o	o	o	(comprehensible)
(not rational)	o	o	o	o	o	o	o	(rational)
(not selfish)	o	o	o	o	o	o	o	(selfish)

In general, do you like destroying money?

o true o not true

I am always fair to others, even if I am at a disadvantage because of it.

o true o not true

I think fairness is an exceptionally important characteristic of humans.

o true o not true

I dislike taking responsibility.

o true o not true

[22]Questionnaire 1 and 2 were identical in designs 1A and 2A. The questionnaires in NC were nearly identical, we only dropped some questions, because they were not relevant (the questions about behavior of player B).

[23]In design 2A we had some additional questions regarding the bidding behavior. We asked subjects to explain why they choose to bid or not to bid, how they chose their bid and finally, to judge the likelihood to win the auction with their bid.

I rarely hit back, even if someone else hits me first.

o true o not true

If someone hits me first, I'll show him.

o true o not true

If I am angry I occasionally bang doors shut.

o true o not true

If someone angers me, I tend to tell him what I think about him.

o true o not true

Even if I don't show it, I am sometimes consumed with envy.

o true o not true

If someone does not treat me right, I do not let it get at me.

o true o not true

Before we pay out the money, you have the possibility to destroy an arbitrary amount of your own Thaler lying in front of you. Do you want to destroy Thaler?

o Yes, Thaler o No, I don't want to destroy my own Thaler.

CHAPTER 2

More than Meets the Eye: An Eye-tracking Experiment on the Beauty Contest Game[1]

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 getting 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, Ben-Ami, and Rubinstein (2011), Knoepfle, Wang, and Camerer (2009), Wang, Spezio, and Camerer (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

[1]This chapter comprises the paper co-authored with Christiane Schwieren.

the beauty contest game. Monitoring the sequence of numbers considered in the thinking process gives us a deeper insight into the strategies used.

We find that the evidence with respect to levels of reasoning is less clear than has so far been assumed. We find 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).[2]

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 equilibrium: all players announce zero.[3]

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 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

[2] For the history of the beauty contest game see Bühren, Frank, and Nagel (2009).

[3] 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 multiple equilibria; see López (2001) for a characterization of these.

have inferred 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 $[100p^{(k+1)}, 100p^k]$.[4]

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, Crawford, and Broseta (2001), Camerer, Ho, and Chong (2004), Ho, Camerer, and Weigelt (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 hierarchy 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 interval 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 (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.[5] 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[6] 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

[4] 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.

[5] 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.

[6] 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.

one step of reasoning, choosing 33 with $p = 2/3$, or 2 steps, choosing 22. Ho, Camerer, and Weigelt (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 (2003a), Crawford, Costa-Gomes, and Iriberri (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", conducts his or her own experiments with friends to find out what they are doing. Additionally 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, Crawford, and Broseta (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, Huang, and Wang (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 Eye-tracking 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.

Table 1. Values of p used in the experiment

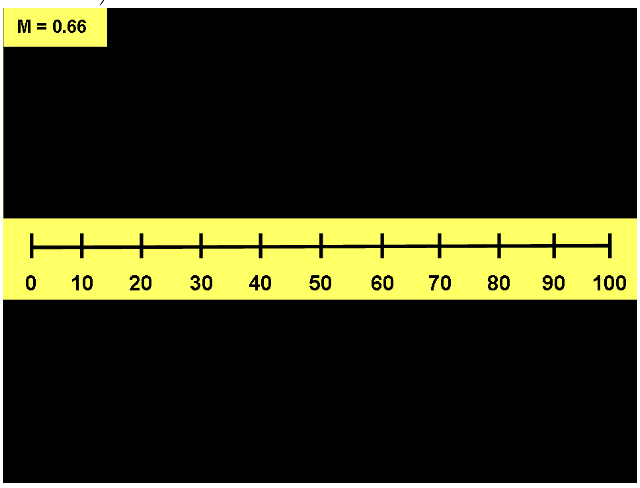
round	1	2	3	4	5	6
p	0.125	0.2	0.33	0.5	0.66	0.75

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

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



playing with nine other players who either had already played or would play later, up to reaching ten players in total.[7] The design of our experiment differs from most other beauty contest games in laboratories, but 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 head-mounted system and chose the respective camera. After fixing the system comfortably on subjects' head the experiment started with a calibration phase. Only after reaching a good fit we proceeded 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.

[7]Translated instructions of the experiment can be found in the appendix 6.2

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[8]. 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[9] 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.

Table 2. 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

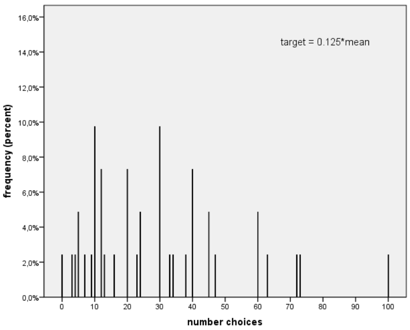
If we compare the chosen numbers with the typical number choices in other experiments, we find that our subjects chose somewhat higher numbers. For example our mean of 46.41 (for p = 0.66) is definitely larger (t-test two-tailed,t =2.839, p=.007) than the mean of 36 reported in Nagel (1995). Figure 2 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 equilibrium - and if the Nash equilibrium is chosen then it usually happens in later rounds, after learning from feedback.

For analyzing how many subject choose a weakly dominated strategy we are going to look for chosen numbers larger than $100p$. The frequency of the choices greater than $100p$ for the different parameters p can be found in table 3. For the larger parameters we find the usual

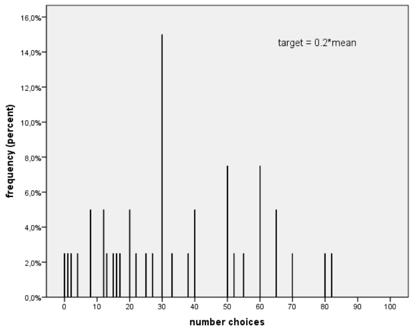
[8]Originally we had eye-tracking data of 40 subjects, but we excluded one subject from the analysis who was familiar with the beauty contest game.

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

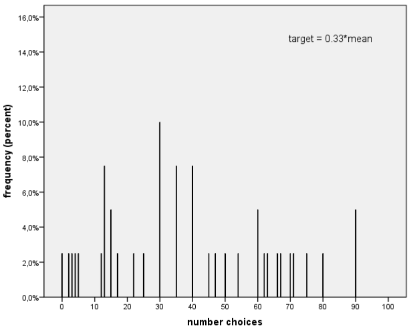
Figure 2. Number choices



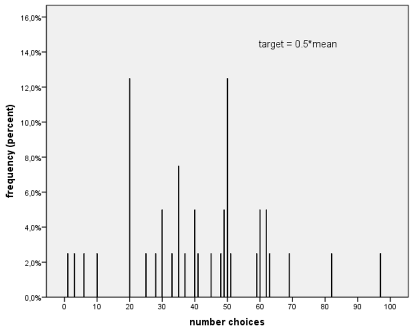
(a) $p = 0.125$



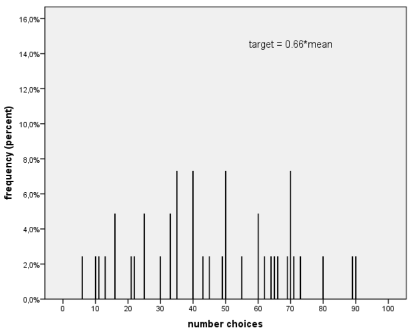
(b) $p = 0.2$



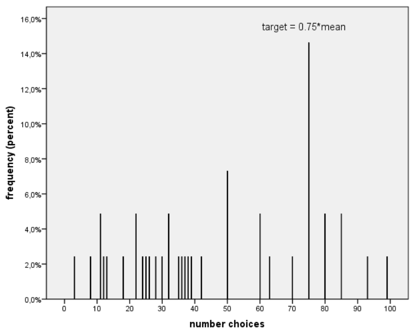
(c) $p = 0.33$



(d) $p = 0.5$



(e) $p = 0.66$



(f) $p = 0.75$

percentage, around 20 to 25% choose dominated strategies, while for the smaller parameters we have more subjects choosing a dominated strategy.

Table 3. Choice of weakly dominated strategies

	$p = 0.125$	$p = 0.2$	$p = 0.33$	$p = 0.5$	$p = 0.66$	$p = 0.75$
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%

4.1.1. *Levels.* 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 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[10] 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.

[10]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.

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	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 5. Levels via level-k-model using 50

	$p = 0.125$		$p = 0.2$		$p = 0.33$		$p = 0.5$		$p = 0.66$		$p = 0.75$		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	

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).

Table 6. Levels per subject

by majority		**tie**		**learner**	
level	frequency	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				

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[11] in the literature.

4.1.2. *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)).

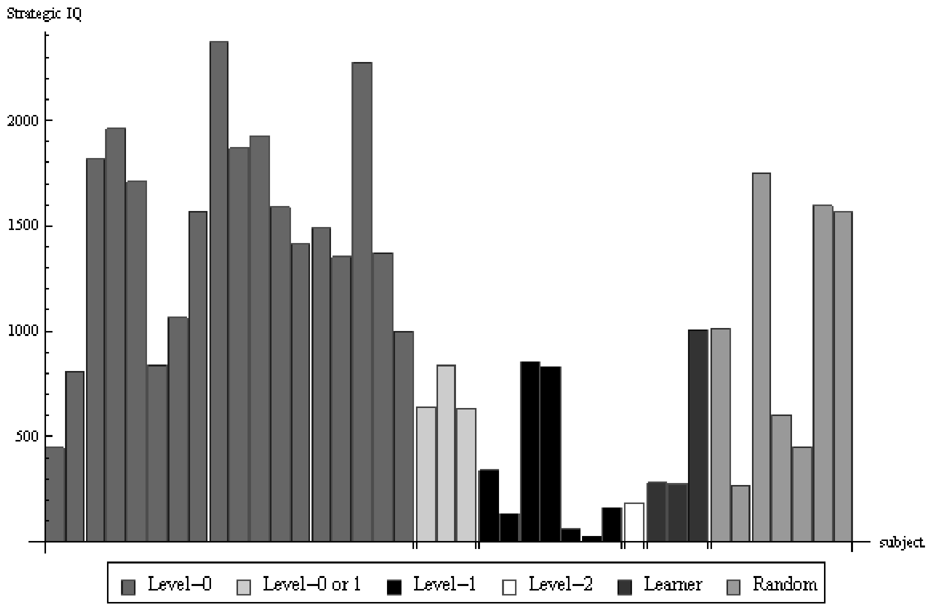
We have a measure of strategic IQ for each round and take the average over the rounds to generate one aggregate measure of strategic IQ.

Additionally we asked our participants, as a proxy for intelligence, to provide their grade in the "Abitur" (the German High-school 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.

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

FIGURE 3. Strategic IQ for each subject, ordered by the level assignment as listed in table 6



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.

4.2.1. *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 definitions of the Eyelink II System.

4.2.2. *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.

TABLE 7. Duration for the different 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

4.2.3. *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.

TABLE 8. Fixations in the interest area *parameter*

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

4.2.4. *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.

4.2.5. *Hot Spots.* Hot spots were analyzed 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).

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)

	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

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.

4.2.6. *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).

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.

Table 10. Fixations only below 50

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%

4.2.7. *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 definition, 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 need 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 over-classification with a simple definition 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 interfered 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 exactly 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 for example Bosch-Domenech *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.

6. Appendix

6.1. 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 eye-tracked.

6.2. 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.

6.2.1. *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 chooses 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.

6.2.2. *Part 2.* In this part you will play another game[12], which is unrelated to the first part of the experiment. You will get the instructions for this game during the experiment.

6.3. Questionnaire. About the first part of the experiment (target number)[13]

(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?

General questions:

Please tell us which grades you received in your Abitur (the German High-school diploma)

- in Math
- in German
- your Average Grade

[12]We report this data separately.

[13]We also had similar questions about the second part of the experiment.

CHAPTER 3

What Can the Big Five Personality Factors Contribute to Explain Small-scale Economic Behavior?[1]

1. Introduction

Recently, a growing interest among (behavioral) economists in personality variables can be observed (e.g., Almlund *et.al.* (2011), Dohmen *et.al.* (2010), Borghans *et.al.* (2008)). In most published studies involving personality measures so far, the Big Five personality factors are used. Usually, correlations of the personality measures with some real-world aspects of economic behavior are reported and interpreted, for example with earnings or performance on the job. [2] Researchers in experimental economics recently also started to include personality measures in experiments, hoping to be able to explain part of the behavioral heterogeneity found. Many studies relate some kind of Big Five personality variables, although measured by different instruments, to behavior in games like the Prisoner's dilemma, dictator, or ultimatum games (e.g., Brandstätter and Güth (2002), Ben-Ner, Kong, and Putterman (2004), Ben-Ner *et.al.* (2004), Swope *et.al.* (2008)). Other studies use more specific scales, as locus of control, self-monitoring and sensation seeking (Boone, De Brabander, and Van Witteloostuijn (1999)), or the Myers-Briggs Type Indicator (Schmitt *et.al.* (2008)). Results of these exercises so far are not very conclusive.

One reason for this might lie in a methodological concern: Is it reasonable to expect values on personality scales to be predictive of microbehavior in economic games? It is undoubted that personality can influence economic outcomes at large (Ozer and Benet-Martínez (2006)), such as occupational attainment (Filer (1985)) or occupational performance and success (Barrick and Mount (1991), Seibert and Kraimer (2001)). Whether personality variables can also be used to understand "micro"-behavior in economic games is however less clear.

In this paper, we discuss reasons in favor and against this assumption and test in our own experiment, whether and which personality factors are useful in predicting behavior in the trust game (Berg, Dickhaut, and McCabe (1995)). We can also use the trust game to

[1]This chapter comprises the paper co-authored with Christiane Schwieren.

[2]See for example Barrick and Mount (1991); Mueller and Plug (2006).

understand how personality measures fare relatively in predicting behavior when situational constraints are strong. This approach will help economists to better understand what to expect from the inclusion of personality variables in their models and experiments, and where further research might be useful and needed.

The aim of this paper is exploratory, and due to this, our method is somewhat non-standard: We use the NEO-PR-I (Costa and McCrae (1992)) to measure the Big Five personality factors and link scores with behavior in a trust game, both of trustor and trustee. To find the relevant predictors we use in our regressions the method of backward stepwise elimination (Eid, Gollwitzer, and Schmitt (2010)). We do this on two levels – first, on the level of the five factors, and then also on the level of sub-scales. Here we follow an argument by Paunonen and Ashton (2001) who propose to look at sub-scales (facets) as well for predicting behavior, because they are more specific and therefore more apt to explain small-scale behavior.

To preview our results, first, we can show that behavior of player 1 is more strongly determined by personality than behavior of player 2. Second, our analysis of subscale-correlations can tell us something about the trust-game in general. We discuss these results on the background of our aims, to get an idea of when personality matters and whether using personality as an additional explanatory variable is recommendable for (experimental) economists, and how this could be done.

The remainder of the paper is structured as follows: In section 2, we give an overview of the literature on personality measurement. Then, we describe our experimental design (section 3) and the personality measures (section 3.2) used in more detail. Section 5.2 presents the results for player 1's behavior and section 5.3 those for player 2's behavior. Section 6 discusses the results and concludes.

2. Measurement of Personality

Personality psychology provides a large set of specific measures of potential interest for economists. On the one hand, there are general models of personality, comprising usually between four and seven general factors of personality (e.g., Goldberg (1981); Cloninger, Svrakic, and Przybeck (1993); Cattell and Schuerger (2003)). These are measured with different scales, varying in the content of the factors and the sub-factors measured. The most famous example is the NEO-PR-I measuring the so called Big Five Personality Factors (Costa and McCrae (1992)). On the other hand, there are more specific measures, capturing certain aspects of personality like anxiousness or aggressiveness. Here, we focus on the general measures and use the NEO-PI-R (Costa and McCrae (1992)), German: (Ostendorf and Angleitner (2004)) to measure the Big Five personality factors.

Researchers in personality psychology discuss whether personality factors can be expected to correlate strongly with real life outcomes and behavior, and whether it would be problematic if this were not the case. Since Mischel (1968), many personality psychologists argue that there is a ceiling of a correlation of .3 between personality variables and real life outcomes, the so called *.3 barrier* (Mischel (1968); see also McCrae (1982) for exceptions). Researchers that adhere to this ceiling argument put forward that the situation is at least as or more important in determining behavior and important life outcomes as is personality. Others (e.g., Ozer (1985)) however argue that .3-correlations are not so small and can have important practical effects and that most social, psychological (and even medical) variables, like socioeconomic status or cognitive ability do on average not correlate any stronger with important life outcomes. It is noteworthy that usually the outcomes studied are larger-life outcomes, such as divorce, occupational or educational attainment, and not "micro"-behaviors as trust-game behavior. An exception to this is research in organizational behavior that links, for example, locus of control or conscientiousness to individual performance, turnover decisions etc. (e.g., Judge and Bono (2001); Allen, Weeks, and Moffitt (2005); Dudley *et.al.* (2006)). Most researchers argue that personality influences outcomes in life not in a direct way, but rather affects general tendencies to act, e.g., to continue an education or to be persistent despite failures, which then influences the developmental path over the life span.

We therefore do not expect to be able to explain behavior in the trust game by a single personality factor. We do however think that if personality is indeed something important influencing behavior, it should at least somewhat contribute to an explanation also of small-scale behavior, especially when the situation does not provide much guidance on how to behave.

3. Experimental Design and Procedure

The experiment was conducted in the experimental laboratory of SFB 504 in Mannheim. We had 138 subjects in total (57 male, 70 female, the remaining failed to indicate their sex). All subjects participated in two sessions with one week in-between. The experiment consisted of 12 independent sessions in the first week and 12 sessions in the second week. In total, the experiment lasted for about one hour in the first and one hour in the second week. Subjects had filled in the personality questionnaires before our experimental sessions, which took them about 2 hours. We paid subjects at the very end of the experiment, i.e. after the session in the second week. Part of these earnings were performance-based, and part was fixed: both in week one and two they received a show-up fee of €5, and they received a fixed amount of €14 for filling in the personality questionnaires. Parts of the experiment (the questionnaires) were conducted via pen and paper and parts (the games) were programmed and conducted with the software z-tree (Fischbacher (2007)). All subjects were paid in cash and private. Subjects knew about the whole timing in advance, at the beginning of each session they received instructions containing the curse of events of the session. For each of the games and the decisions instructions were distributed and also read aloud in each part by the experimenter and participants had a chance to ask questions.

3.1. The Trust Game. Players were randomly assigned to be either player 1, the trustor, or player 2, the trustee. Two players were randomly matched together. Both players got 10 units of an experimental currency. The trustor could first decide whether or not to sent units to player 2, if he sent x units ($x \leq 10$), these units got tripled. Then player 2 got informed about the amount she received and she could decide to send an amount y ($y \leq 10+3x$) back to player 1 (these units were not tripled). Therefore the payoff for both players were determined by

$$\text{player } 1: 10 - x + y \qquad \text{player } 2: 10 + 3x - y. \tag{1}$$

At the end of the experimental currency was transformed into Euro with an exchange rate of 1 Euro =0.3 ECU.

3.2. The Big Five. To measure personality we use the *five-factor model* or the *"Big Five"* (Goldberg (1981), McCrae and Costa JR (2003))).

This model organizes personality traits in five basic dimensions: neuroticism, extraversion, openness to experience, agreeableness and conscientiousness.[3] A list of the personality dimensions and their facets measured by the Big Five model can be found in table 1.

We use the NEO PI-R (Costa and McCrae (1992)), German version (Ostendorf and Angleitner (2004)) to measure the Big Five personality factors. It consists of 241 items which have to be rated on a 5-point-Likert-scale.

4. Behavioral Predictions

As this paper is exploratory in character, we do not test specific hypotheses but rather explore how personality is related to behavior in the trust game. We did however formulate predictions that we will explain in the following. The basis for our predictions is on one hand the analysis of the situation both players in the trust game are in.

[3] There are other labels for the five factors, we use the names by Costa and McCrae (1992).

TABLE 1. The five factors and their facets (NEO-PR-I), acronyms in parenthesis

Factor	**Facets**
Neuroticism (N)	Anxiety ($N1$), Angry Hostility ($N2$), Depression ($N3$), Self-Consciousness ($N4$), Impulsiveness ($N5$), Vulnerability to Stress ($N6$)
Extraversion (E)	Warmth ($E1$), Gregariousness ($E2$), Assertiveness ($E3$), Activity ($E4$), Excitement-Seeking ($E5$), Positive Emotions ($E6$)
Openness to Experience (O)	Fantasy ($O1$), Aesthetics ($O2$), Feelings ($O3$), Actions ($O4$), Ideas ($O5$), Values ($O6$)
Agreeableness (A)	Trust ($A1$), Straightforwardness ($A2$), Altruism ($A3$), Compliance ($A4$), Modesty ($A5$), Tender-Mindedness ($A6$)
Conscientiousness (C)	Competence ($C1$), Order ($C2$), Dutifulness ($C3$), Achievement-Striving ($C4$), Self-Discipline ($C5$), Deliberation ($C6$)

On the other hand, we rely on the literature in personality psychology to derive predictions about which personality factors should be most important for behavior of player 1 and of player 2 in the trust game.

First, we derive hypotheses for the link between personality factors and behavior, i.e. we hypothesize in what way a subject with a certain personality will behave in the trust game.

Neuroticism refers to the tendency to experience negative emotions and feelings, especially anxiety and general distress. Therefore we would expect that a person with high *neuroticism*-scores to be anxious and to avoid the risk of not getting the money back.

PREDICTION 1. *Higher levels on neuroticism will correlate with lower amounts sent by player 1.*

With respect to *extraversion* and *openness to experience*, we do not have clear-cut predictions regarding behavior in the trust game.

Agreeableness is defined as being compassionate and cooperative, the names of the facets are rather self-explanatory. *Agreeableness* is linked to cooperative behavior (Volk, Thöni, and Ruigrok (2011); LePine and Van Dyne (2001)). This leads to the following intuitive prediction:

PREDICTION 2. *Higher levels on agreeableness will correlate with higher amounts sent by player 1 and with higher amounts returned by player 2.*

People high on *conscientiousness* act planned not spontaneous, are dutifully and self-disciplined. Therefore we could imagine that high levels*conscientiousness* will lead to higher amounts sent by player 1 (being dutiful) if a norm for sending is salient. As *conscientiousness* is also linked to rationality (D'Zurilla, Maydeu-Olivares, and Gallardo-Pujoi (2011), Witteman *et.al.* (2009)) and high levels of could *conscientiousness* as well lead to lower amounts sent by player 1 (being rational). For player 2, we assume the norm of reciprocity to be salient and thus, controlling for player 1's sending we expect trustors that are high on *conscientiousness* to follow this norm dutifully, and thus to send back more.

PREDICTION 3. *Higher levels on conscientiousness of player 1 might lead to more or less sending. For player 2, we assume that high conscientiousness-scores lead to higher returns.*

The reason to select the trust game for our research is that it contains two different situations (for player 1 and 2 respectively) that can be described in terms of a distinction often made in personality psychology: the distinction between *weak* and *strong* situations (Mischel (1977)). In *weak situations*, the behavioral triggers stemming from the situation are weak, and therefore personality variables should contribute significantly to an explanation of behavior. In *strong situations* on the contrary, situational triggers of behavior are strong and therefore personality variables should not contribute much to an explanation of behavior if player 1's behavior has been controlled for.

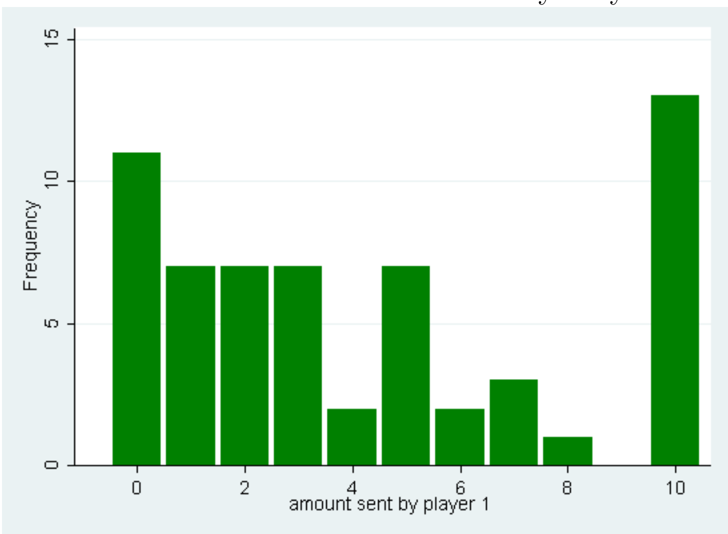
For player 2, the situation she finds herself in is relatively clearly determined: Player 1 has either trusted her with a certain amount of money and now she has to decide how to react to this. As is known from the experimental literature, reciprocity is a strong norm prevailing in this context (e.g., Berg, Dickhaut, and McCabe (1995); McCabe, Rassenti, and Smith (1998); Fehr and Gächter (2000b); McCabe, Rigdon, and Smith (2003)). Player 1 however faces a situation where norms or guidances for behavior are not that clear. Personal tendency to trust or to take risks will determine how much of the money he will send to player 2.

PREDICTION 4. *First players find themselves in rather* weak *situation, therefore personality variables should contribute significantly to an explanation of behavior. Second players are in a rather* strong *situation, therefore personality variables should not contribute much to an explanation of behavior if player 1's behavior has been controlled for.*

5. Results

5.1. Behavior in the Trust Game. Figure 1 shows the distribution of the amount sent by player 1 in the trust game.

FIGURE 1. Trust – Amount Sent by Player 1



60 subjects played the trust game in the role of player 1, and the mean amount sent by player 1 is 4.3. This is slightly below what is usually reported. Usual results are that player 1 sends on average half of his endowment and trust is not repaid by player 2 (e.g., Camerer (2003b)).

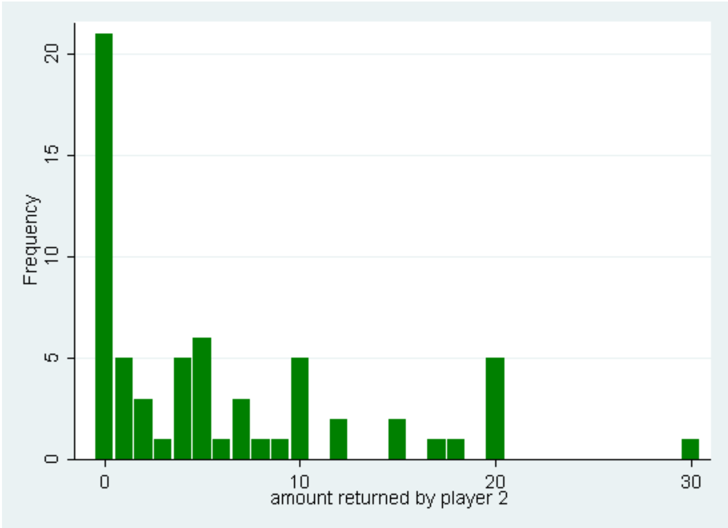
Returns show an absolute average of 5.9, and are strongly correlated with offers ($r = .736$). Figures 3(a) shows absolute returns. The relation between the amount sent by player 1 and the amount returned by player 2 can be found in 3(b), the red line indicates where the amount sent is equal to the amount returned. Above the red line, trust is repaid by player 2.

5.2. Personality measures and trustors behavior. Generally, we find reasonable variance in our personality scales[4], even though one might assume at least with respect to some of the scales that a student population might be comparably homogeneous. Scores on all five of the personality measures are normally distributed (Kolmogorov-Smirnov test of normality).

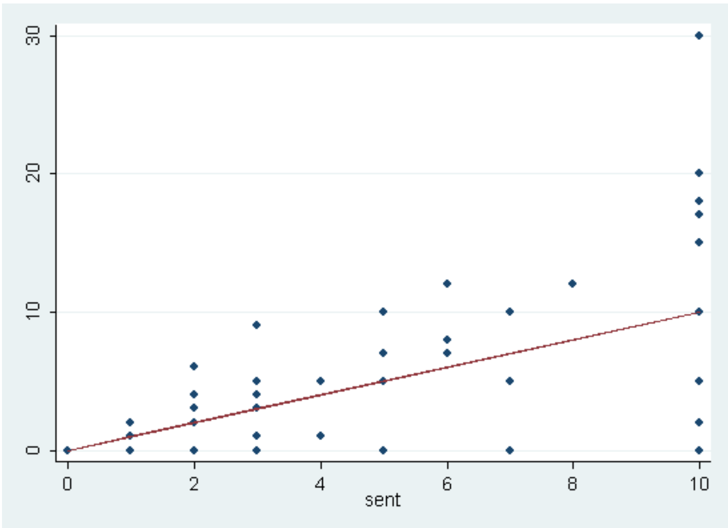
As a first step to test our predictions we calculate correlations of first player behavior (and later second player behavior) with the personality factors. We report correlations in the first column of table 2. This table also shows intercorrelations of the personality measures.

[4]For descriptive statistics of the personality scales see table 7 in appendix 7.2.

FIGURE 2. Trustworthiness – Player 2



(a) Amount Returned by Player 2



(b) Trustworthiness – Amounts sent and returned

As conjectured in predictions 1 and 2, sending of player 1 correlates significantly negative with *neuroticism*, and significantly positive with *agreeableness*.

Next we look at subscales (facets) and also calculate correlations here. We only report significant correlations.

We analyze the facets of the two factors that correlate with trustor behavior, *neuroticism* and *agreeableness*. *Anxiety* ($N1$), *angry hostility* ($N2$), and *depression* ($N3$) correlate significantly (negative) with sender behavior among the subscales of *neuroticism*. Of the subscales

TABLE 2. Correlations between x, the amount sent by player 1, and the personality factors

	x	N	E	O	A	C
N	-0.339**	1.000				
	0.009					
E	-0.052	-0.331**	1.000			
	0.697	0.011				
O	0.199	-0.102	0.404**	1.000		
	0.134	0.446	0.002			
A	0.284*	-0.071	0.146	0.133	1.000	
	0.031	0.596	0.274	0.318		
C	-0.258	-0.210	0.233	0.010	-0.078	1.000
	0.050	0.113	0.078	0.938	0.561	

*, **, *** indicate significance at the 5%, 1% and .1% level respectively. Abbreviations: N = *neuroticism*, E = *extraversion*, O = *openness to experience*, A = *agreeableness*, C = *conscientiousness*

of *agreeableness*, *trust* ($A1$) correlates significantly positive with the amount sent by player 1, and so does *straightforwardness* ($A2$).

TABLE 3. Correlations between x, the amount sent by player 1, and the personality facets of *neuroticism* and *agreeableness*

	x
$N1$	−0.377**
	0.003
$N2$	−0.319*
	0.015
$N3$	−0.280*
	0.033
$N4$	−0.211
	0.111
$N5$	−0.078
	0.559
$N6$	−0.123
	0.358

	x
$A1$	0.370**
	0.004
$A2$	0.314*
	0.016
$A3$	0.167
	0.210
$A4$	0.150
	0.262
$A5$	−0.007
	0.958
$A6$	0.172
	0.197

Although having an intuitive appeal, *altruism* ($A3$) does not correlate with behavior of the trustor. There is a discussion about other motives than trust that are included in the trust game; Cox (2004) points out that not only trust and trustworthiness, but also altruistic preferences can account for sending by player 1 or returning by player 2. From the personality variables involved in player 1's decision, only *trust*, but not *altruism* has an influence on the amount sent by the first player.

After this first results, which personality factors and facets have an influence on the amount sent by player 1, we now turn towards the regression analysis.

Here, we use a modeling approach that is often used in exploratory studies (e.g., Eid, Gollwitzer, and Schmitt (2010)): Backward stepwise elimination of insignificant predictors. In the first step we include all predictors that correlate with behavior of player 1 and also the control variables in a regression. We then stepwise eliminate always the least significant predictor until we get a model that consists only of significant predictor variables. This exploratory way of modeling is indicated in our case as most of the personality variables we study are inter-correlated. A model including all potentially relevant personality variables therefore underestimates the explanatory power of each of the variables, due to multicollinearity. By doing a step-wise elimination of insignificant predictors, we reach a model where only the most inclusive and important personality variables remain.

Table 4. Regression on x, the amount sent by player 1

Variable	**model I**	**model II**	**model III**	**model IV**
N	-0.292^{**}	-0.320^{**}		
	(0.024/0.032)	(0.022/0.011)		
N1			-0.327^{*}	-0.316^{**}
			(0.121/0.063)	(0.082/0.011)
N2			0.047	
			(0.124/0.782)	
N3			0.032	
			(0.126/0.859)	
A	0.261^{**}	0.261^{**}		
	(0.025/0.042)	(0.024/0.037)		
A1			0.238	0.213^{*}
			(0.114/0.112)	(0.099/0.099)
A2			0.243^{*}	0.235^{*}
			(0.109/0.075)	(0.100/0.061)
age	−0.071		−0.102	
	(0.113/0.572)		(0.111/0.410)	
sex	−0.061		−0.041	
	(0.989/0.650)		(0.999/0.760)	
n	58	58	58	58
R^2	0.1902	0.1827	0.2830	0.2719
adj. R^2	0.1291	0.1530	0.1826	0.2314

Note: beta, SE/p-value in parenthesis; *, **, *** indicate significance at the 10%, 5% and 1% level respectively.

For the trustor we explain the amount sent in two different ways using the Big Five personality variables: in the first approach we use the factors correlating individually with the behavior of the first player (table 2) and in the second approach we use the facets of this factors, and again only the facets correlating individually with the behavior of the first player (table 3).

Table 4 shows all four models (model I and II – factor-approach, model III and IV – facet-approach). In model I, the factors *neuroticism* and *agreeableness*, together with the controls are included, stepwise elimination leads to model II. In both models *neuroticism* has a significant negative impact on the amount sent and *agreeableness* a significant positive impact. These models explain 13% to 15% of the variance in the amount sent by the first player. Using facets, significant predictors are *anxiety* ($N1$) and *straightforwardness* ($A2$), again signed as before: a negative impact of the *neuroticism*-facet and a positive impact of the *agreeableness*-facet. The explained variance is 18%. In model VI, all insignificant predictors have been excluded, and the remaining predictors are those that had been significant before, and *trust* ($A3$) having also a positive impact. This model explains an even larger part of the variance in player 1's behavior, around 23%.

Anxiety is defined[5] as the level of free floating anxiety and has been linked to risk averse behavior in different domains[6], so it is intuitive that this facet has a negative impact on the amount sent by player 1. The positive influence of *trust* highlights again that the trust game is indeed about trust: *trust* being defined as the belief in the sincerity and good intentions of others has a positive impact on the amount sent by player 1. Finally, *straightforwardness* is defined as frankness in expression, in general a person high on *straightforwardness* is rather frank, sincere, and ingenuous, than manipulative or deceptive.

5.3. Trustee Behavior. We now turn to the behavior of the second player. As described before, there is one clear difference between predicting first player's behavior and predicting second player's behavior: behavior of player 2 will most probably be guided by reciprocal incentives, i.e., what the first player has sent to the second player will matter. We thus have a *strong situation* here, as opposed to the *weak situation* in which first players find themselves in. In line with the general search for interactions of personality variables and the environment in personality psychology, our main question is whether personality variables predict beyond "material", situational characteristics, or whether it is only player 1's behavior that predicts the responses of player 2.

We start again with correlations. If we take data of all trustees, the only and highly significant predictor of player 2's behavior is the amount player 1 sent to her (0.731, $p = 0.000$), we find no correlation

---

[5] All facet-definitions following Costa and McCrae (1992).

[6] Nicholson, O'Creevy, Soane, and Willman (2005) analyze personality and risk propensity in different domains and find *anxiety* being linked to less risk taking in recreation, career, and safety.

at all between behavior of player 2 and any of the personality factors or of the sub-factors, i.e., the situation determines behavior stronger than do personality variables.

TABLE 5. Regression on y, the amount returned by player 2

Variable	**model I**	**model II**
sent	0.737***	0.731***
	(0.169/0.000)	(0.160/0.000)
Sex	−0.56	
	(1.322/0.546)	
Age	0.049	
	(0.280/0.592)	
n	61	64
R^2	0.540	0.534
adj. R^2	0.516	0.527

Note: beta, SE/p-value in parenthesis; *, **, *** indicate significance at the 10%, 5% and 1% level respectively.

Including these variables in a regression on the return of player 2, the results in table 5 again highlight that only the amount that player 1 has sent to player 2 explains the amount player 2 returns.[7]

In an attempt to give personality variables the best opportunity to have an effect, we look at only part of our sample, namely those second players who received high offers in the trust game, where high offers are defined[8] as offers of at least five. We are aware of the fact that analyzing this sub-sample leads to a small number of observations, but due to the exploratory character of this study we still give personality variables this chance.

As expected, personality factors have some influence on behavior in the case where player 1's sending has been reasonable and fair: On the factor-level we still find no correlations, but on facet-level *modesty* ($A5$) correlates negatively with the amount returned (-0.415, $p = 0.035$), while *competence* ($C1$) correlates positive (0.399, $p = 0.044$).

A positive relationship between a *conscientiousness*-facet is in line with prediction 3. This is also generally coherent as *contentiousness* is defined as the degree of organization, control and goal directed behavior, the facet *competence* measures the belief in own self efficacy

[7]Sometimes it is argued that one should only analyze those second players that received positive amounts from the first player, because players receiving zero are forced to also sent back zero. Repeating our analysis only with those subjects that received strictly positive amounts we find structurally the same results as shown in table 5.

[8]This definition of *high* conveys the definition of Blanco, Engelmann, and Normann (2011) for high offers in an ultimatum game.

which could intuitively be linked to higher amounts returned. In contrast, a negative relationship between *modesty* and the amount return is very unintuitive. *Modesty* is defined as a tendency to play down own achievements and be humble. Because of the small sample size we had a closer look at this result. A visual inspection and detection of extremes demonstrated that some subjects with low scores on *modesty* returned very high amounts. Analyzing this observations more detailed we declared two of the observations as outliers. After dropping this two observations there is no longer a significant correlation between *modesty* and the amount returned. Therefore, *modesty* is not included into the following regression.

TABLE 6. Only for high amounts received: regression on y, the amount returned by player 2

Variable	**model I**	**model II**
sent	0.449**	0.497**
	(0.776/0.048)	(0.561/0.004)
$C1$	0.222	
	(0.436/0.311)	
Sex	−0.063	
	(3.140/0.756)	
Age	0.090	
	(0.507/0.646)	
n	25	31
R^2	0.310	0.247
adj. R^2	0.172	0.221

Note: beta, SE/p-value in parenthesis; *, **, *** indicate significance at the 10%, 5% and 1% level respectively.

Including the facet *competence* in a regression together with the amount sent by the first player and controls for age and sex, in the full model again only the amount sent is significant, and 17% of the variance is explained, while after step-wise exclusion of insignificant predictors, the amount sent by the first player remains the only significant predictor in the model, and the model explains 22% of the variance.

So even in this case where we gave personality its best chance: the situation determines behavior.

6. Discussion

To answer the question to what extent personality can contribute to explain small-scale economic behavior we decided to use the trust game as an example.

There are other studies relating the trust game or trust in general to personality. Two studies relate the Machiavellian personality test to the trust game: Gunnthorsdottir, McCabe, and Smith (2002) use a modified trust game and Burks, Carpenter, and Verhoogen (2003) the standard trust game. Having hypotheses about both trust and trustworthiness, related to scoring high on Machiavellism, Gunnthorsdottir, McCabe, and Smith (2002) find that subjects high on Machiavellism are less trustworthy, where Burks, Carpenter, and Verhoogen (2003) find that high Machiavellism predicts lack of trust, but not trustworthiness.

Fahr and Irlenbusch (2008) use the Big Five personality model, measured by Catell's 16 PF-R, to analyze trust between representatives of organizations. To study this question they use a modified trust game. To implement their organizational setting players were in groups of four and had to decide as a representative of their own group. They found a link between *anxiety*, being linked to risk averse behavior, and trustor behavior and *anxiety*, on the other hand being liked to cooperative behavior, to trustee's decision. Our study strengthens the result that *anxiety* is linked to distrust. Using another measure of the Big Five and the standard trust game we find that trust is negatively related to *anxiety* (see table 4).

The research focus of Ben-Ner and Halldorsson (2010) concentrates on understanding trusting and trustworthiness. They use many different measures, and among others the Big Five factors (measured by the NEO-FFI), but to define trust and trustworthiness they use on the one hand survey questions and on the other hand a modified trust game (a repeated variant).

Our exploratory study had two main aims: First, we wanted to test whether personality variables can be used to predict "micro"-level behavior in economic games, where we use the example of the trust game. Next, we hypothesized that *strong situations* allow for less influence of personality factors than *weak situations*, and that first players in a trust game are in a weak situation, while second players face a strong situation.

Our results confirm most of our general and some of the more specific predictions: First, we do find that personality variables contribute to an explanation of behavior. Trustor behavior can be explained to a large extent using personality variables. This is good news especially for personality psychologists, who so far seldom validate their personality scales with the help of clear-cut behavioral experiments. It is also good news for all those experimental and behavioral economists that now start to use personality measures in their experiments. But, we also confirm the notion of strong and weak situations found in personality psychology: First player's behavior can be explained to a large extent (up to 23% of the variance) using personality variables, while second player's behavior is explained by the situation. This is essentially good news for standard economics, as this means that if incentives or behavioral norms are clear and strongly point into a specific direction, most people, independent of their personality, will react to these incentives, and predictably so.

7. Appendix

7.1. Trust Game.

Instructions. These instructions have been translated into English from the Original German.

In this game you will play together with one other person in the laboratory. You are either player A or player B. This will be randomly determined by the computer. The other person (A or B) you will play together will also be randomly determined by the computer.

Both player A and B receive 10 experimental currency units (ECU). Player A can decide whether he would like to send taler to player B and if so, how many (only integer amounts are possible). The amount of ECU that player A sents to player B is tripled. Therefore player B receives 3 units for each unit sent by player A. Player B can then decide whether she wants to return ECU to player A and if so, how many. These units will not be tripled. This is the end of this game.

The experimental currency is converted into Euros as follows: 1 ECU = 0.30 Euro.

If you now got questions regarding these instructions, raise your hand and one of the experimenters will come to answer your questions.

7.2. Descriptive Statistics of the Personality Scales. In table 7 we show basic descriptive statistics of the Big Five factors.

TABLE 7. Descriptive Statistics of the Personality Scales

Variable	**n**	**Mean**	**SD**	**Min**	**Max**
Neuroticism	126	91.985	23.666	27	155
Extraversion	126	116.020	21.564	32	158
Openness	126	124.478	16.781	72	180
Agreeableness	126	109.925	18.700	67	152
Conscientiousness	126	116.294	21.636	58	166

CHAPTER 4

Can Personality Explain what is Underlying Women's Unwillingness to Compete?[1]

1. Introduction

There is ample evidence that women do not react to competition as men do and are less willing to enter a competition than men. In their first laboratory study in this field Gneezy, Niederle, and Rustichini (2003) found that men increase performance in competitive environments more than women do. Following this experiment there are by now many studies replicating this gender gap in performance and also the gender gap in entering into competitive environments, e.g., Niederle and Vesterlund (2007); Croson and Gneezy (2009); Cason, Masters, and Sheremeta (2010); Dohmen and Falk (2011); Niederle and Vesterlund (2010); Niederle, Segal, and Vesterlund (2010); Datta Gupta, Poulsen, and Villeval (2011). Trying to understand this gender gap researchers studied hormones (Wozniak, Harbaugh, and Mayr (2010)) or social imprint (Booth and Nolen (2011)). There are replications of the findings with different age and cultural groups (Gneezy and Rustichini (2004); Gneezy, Leonard, and List (2009); Dreber, Von Essen, and Ranehill (2011); Sutter and Rützler (2010); Andersen *et.al.* (2011); Cárdenas *et.al.* (2011)).

To the best of our knowledge no paper thus far has tried to understand the underlying motives of women to enter a competition less willingly than men do. One way of doing this is to study personality variables that are related to performance and achievement. This is the focus of the current paper.

A recent paper by Almlund *et.al.* (2011) gives a good overview of the literature of personality psychology of relevance for economics. Fietze, Holst, and Tobsch (2011) attempt to explain the gender career gap by personality. They use SOEP-Data and find evidence that the personality traits have an influence on the career gap. While the direct impact is rather small, they discuss the possibility that there is also an indirect effect.

We study the Big Five personality factors (Goldberg (1981), McCrae and Costa JR (2003)). We test whether the Big Five are related

[1] This chapter comprises the paper co-authored with Christiane Schwieren.

to performance in our experiment, and whether this depends on incentives. We then relate gender differences in personality to the choice of an incentive system.

We replicate the experiment by Niederle and Vesterlund (2007) with a sample of participants who have filled in various personality questionnaires before coming to the laboratory. In this experiment, subjects can earn money by solving real-effort tasks (summing up two-digit numbers). They start out with a piece-rate payment scheme, followed by a winner-takes-all competition in groups of four. In a third round, subjects can choose whether they prefer a competitive incentive scheme or a piece-rate incentive scheme for this round. Niederle and Vesterlund (2007) report a clear gender difference in choice. Women are less willing to enter a competition than men, such that — based on performance -– too few women, but too many men enter the competition.

We can show that the sex difference in our sample can be explained by a difference in neuroticism. We further show that neuroticism is negatively related to performance in a competitive setting. This raises the possibility that those women who do not choose competition "know" that they should not do so, even though their piece rate performance is high. Our results are a first step towards a clarification of the determinants of the gender difference in preferences for competitive environments.

In the remainder of the paper section 2 describes our experimental design and procedures, in section 3 we explain our research question. Section 4 reports the results and section 5 concludes with a discussion.

2. Experimental Design

The experiment[2] was conducted in the experimental laboratory at Mannheim University. We had 138 subjects in total (57 male, 70 female, 11 failed to indicate their sex and thus are not part of the analyses reported here). We paid subjects at the very end of the larger study

[2]This experiment is part of a larger study. All subjects participated in two experimental sessions with one week in-between. A total of 24 sessions were run; twelve in each week, consisting of different experimental games. We also elicited the risk attitude of the participants, using the method by Holt and Laury (2002). The order of the experimental games remained fixed in both weeks over all sessions. In total, the experiment lasted for about one hour in the first and one hour in the second week. Subjects had spent about two hours on average for filling in the personality questionnaires previous to our experimental sessions. Questionnaires were never filled in directly before or after the experimental sessions. Subjects knew about the whole timing in advance. At the beginning of each session they received instructions containing the curse of events of the session. The aim of this large study is to link personality to decisions in economic games. Because subjects had to come at least two times to the laboratory, we decided to get the data for all sub-projects of the study together. We only report data of one of these sub-projects, namely the tournament game, here.

(i.e. after the session in the second week). Earnings from the experiments were performance-based, and a fixed fee was paid for filling in the questionnaires.

The questionnaires were filled in with pen and paper, while the games were programmed and conducted with the software z-tree (Fischbacher (2007)). For each of the experimental games additional individual instructions were distributed and read aloud by the experimenter. Participants had a chance to ask questions before each new game.

2.1. Timing. This project is part of our larger study on personality. The data used in this paper was generated in the first week, and the tournament game was the first game in the session. At the end of the session we elicited subjects risk attitude with the method of Holt and Laury (2002).

The personality questionnaires were filled in the week before the first session in the laboratory.

2.2. The tournament game. For the tournament game[3] we followed the set-up by Niederle and Vesterlund (2007): participants had to do a real-effort task which was to add up two-digit numbers. The allocated time was five minutes and subjects could solve as many problems as they were able to. Subjects got absolute feedback only after each sum.

Subjects were told that the game consisted of four parts and that one of the parts would be randomly chosen for payment.

In each of the four rounds subjects had to do the same real-effort task. In the first round participants were paid with a piece-rate payment scheme, and in the second round in a competitive incentive scheme (winner-takes-it-all-tournament), then they had to choose the incentive scheme they preferred for the third round. In a final step they could decide to submit their performance in the piece-rate part to competitive pay. For competitive pay subjects played randomly matched[4] in groups of four. For our analyses, we focus only on the first decision, whether to enter a competition in round three or not.[5]

The exact rewards were 0.5€ for each correct answer in the piece-rate compensation scheme, and in the tournament compensation scheme

[3]For translated instructions see appendix 6.1.

[4]Here is a difference to the design of Niederle and Vesterlund (2007) were subjects also competed in groups of four, but there were always two male and two female participants in a group. As we made sure that we always had mixed-gender groups in the lab, subjects could believe that they might compete against both sexes.

[5]Very few subjects submitted their piece-rate performance to competition.

the subject who solved most sums received 2€ for each correct answer, all others got nothing.

There was no relative feedback given during the game, but subjects learned for each sum they calculated whether it was wrong or correct and could furthermore always track the number of correct and incorrect answers they had given so far.

After all rounds had been played, participants were asked to indicate how they would rank themselves compared to their group of four in part 1 (piece-rate) and part 2 (forced competition). They had to give their exact position in their group, that is first, second, third or fourth. The accuracy of this ranking was incentivized, subjects received 1€ for each correct ranking.

2.3. Risk measure. To measure risk aversion we use the method developed by Holt and Laury (2002). Subjects had to make a series of ten choices between two risky lotteries. The consequences of the two lotteries were always the same: 2.00€ or 1.60€ for lottery A, and 3.85€ or 0.10€ for lottery B. The probabilities for these consequences were always the same for both lotteries, ranging from 1/10 probability for the higher payoff in the first choice up to a probability of 1 in the last choice, so the probability shifts from the lower to the higher payoff. While subjects should choose lottery A in the first choices, a more risk averse person should switch to lottery B later than a less risk averse person. Payoffs for each subject were determined by randomly choosing one of the ten choices to be paid out.

We use the switching point from Lottery A to lottery B to measure risk aversion – where later switching points correspond to higher risk aversion.

Most of our subjects[6] are consistent, in the sense that they have a unique switching point from lottery A to lottery B. Inconsistency, that is multiple switching, happens only in 6 out of our 127 cases (4.7%).

2.4. Measurement of personality: The Big Five. To measure personality we use the *five-factor model* or the *"Big Five"* (Goldberg (1981), McCrae and Costa JR (2003)). This model organizes personality traits in five basic dimensions: neuroticism, extraversion, openness to experience, agreeableness and conscientiousness.[7] A list of the personality dimensions and sub-dimensions measured by the Big Five scale we use can be found in table 1.

[6]Unfortunately, there was a computer problem, therefore we do not have the risk measure for all subjects: for 127 of our subjects we have the switching point, for 12 we do not.

[7]There are other labels for the five factors, we use the names by Costa and McCrae (1992).

TABLE 1. The five factors (Costa and McCrae (1992))

Neuroticism	Anxiety, Hostility, Depression, Self-Consciousness, Impulsiveness, Vulnerability to Stress
Extraversion	Warmth, Gregariousness, Assertiveness, Activity, Excitement-Seeking, Positive Emotion
Openness to Experience	Fantasy, Aesthetics, Feelings, Actions, Ideas, Values
Agreeableness	Trust, Straightforwardness, Altruism, Compliance, Modesty, Tender-Mindedness
Conscientiousness	Competence, Order, Dutifulness, Achievement-Striving, Self-Discipline, Deliberation

In general, psychological scales to measure personality are derived by various approaches. One consists of using lists of adjectives that are thought to be able to describe personality (Allport and Odbert (1936)). Another approach uses questionnaire data and, with the help of factor-analytic methods, derives traits from the collection of questionnaire items (see Norman (1963), Funder (2001)). Both approaches are rather method-driven than theory-driven. The Big Five Personality Factors that have been derived using a combination of both of these methods (McCrae and Costa Jr. (1987), John and Srivastava (1999)) have been proven to be useful in an empirical sense.

We use the NEO PI-R (Costa and McCrae (1992)), German version (Ostendorf and Angleitner (2004)) to measure the Big Five personality factors. It consists of 241 items rated on a 5-point-Likert-scale.

From these items one constructs a measure for each of the five factors, which we use for our analysis. In the following, we refer to these variables with the name of the factor, *neuroticism*, *extraversion*, *openness*, *agreeableness*, *conscientiousness*.

3. Research question

To explain the gender difference found in competitive environments Niederle and Vesterlund (2007) thought about a personal characteristic, risk attitude, as a predictor. The idea was that gender differences in risk attitude could at least partly explain gender differences in competition. In a review Eckel and Grossman (2008) show that most studies find that women are more risk averse than men. But Niederle and Vesterlund

(2007) find gender differences in risk aversion to play a negligible role in the explanation of the effect that women avoid competition.

In a similar vein we consider a more general concept of personality, measured by the five factor model. Research in personality psychology revealed gender differences in some of the Big Five factors. In a meta-analysis Feingold (1994) found that women score higher than men on *extraversion*, *anxiety* (sub-factor of *neuroticism*), *trust* and *tender-mindedness* (sub-factors of *agreeableness*). In a cross-cultural study Costa Jr., Terracciano, and McCrae (2001) conclude that women score higher on *neuroticism*, *agreeableness*, *warmth* (sub-factor of *extraversion*) and *openness to feelings* (sub-factor of *openness*). Schmitt *et.al.* (2008) report gender differences in personality variables in 49 nations. They also asses the five personality factors, using the Big Five Inventory (BFI). They find that women score higher on *neuroticism*, *agreeableness*, *extraversion* and *conscientiousness*. Regarding *neuroticism* they state that in no country men reported significantly more *neuroticism* than women.

In this paper we address the question whether part of the gender differences in competition can be explained by personality. Knowing that men and women differ in personality variables raised the conjecture that maybe personality mediates the gender difference in the choice to compete. If personality as a broader concept has an influence on behavior in competitive environments it is moreover interesting to know which of the Big Five factors affects behavior most.

4. Results

4.1. Gender Differences in Competitive Settings – Replication. To relate our paper to the literature in the field, we first test whether we can replicate the basic results of Niederle and Vesterlund (2007).

Table 2. Gender differences in choice of incentives

	choice	frequency	percent
Women	piece rate	52	74.3
	tournament	18	25.7
	total	70	100
Men	piece rate	33	57.9
	tournament	24	42.1
	total	57	100

The first question is whether we also find gender differences in the choice of competitive incentives.[8] Table 2 shows that 25.7% of women

[8]For an overview please find a table with variable explanations in appendix 6.2.

choose to compete in round 3, compared to 42.1% of men. This gender difference is (marginally) significant (chi square test $\chi^2 = 3.813$, $p = .051$).

We have to keep in mind however that it might be rational for women not to choose competition, if they indeed perform worse than men in the task. Table 3 shows that the baseline performance in the task is the same for both genders: Performance in the piece-rate payment scheme is not significantly different between men and women. But women perform significantly worse than men in forced competition, even though they improve their performance from piece-rate to competition just as men do (two sample Mann-Whitney-U-tests on *performance PR* $z = 1.518$, $p = .129$, *performance FC* $z = 2.169$, $p = .028$, *improvement* $z = 1.139$, $p = .254$).

Table 3. Performance of men and women

	sex	n	mean	SD
Performance PR	female	70	9.96	3.78
	male	57	11.16	4.51
Performance FC	female	70	12.14	4.57
	male	57	14.02	5.10
Improvement (FC – PR)	female	70	2.19	3.24
	male	57	2.86	4.11

Using the same kind of simulation as Niederle and Vesterlund do, we determine at which performance level a subject should rationally enter the competition. We do not distinguish between men and women here, because our set-up was a bit different from that of Niederle and Vesterlund (2007).[9]

Our simulation indicates that someone solving 14 sums correctly should be (nearly) indifferent between entering the competition or not (having a 24.54% chance to win when entering the competition), while someone solving 15 sums should always enter the competition (having a 32.74% chance to win).

In contrast to Niederle and Vesterlund's findings, in our sample, there are as many women as men who do not enter the competition while they should (65.4% of women vs. 55.2% of men, chi square test $\chi^2 = .596$, $p = .440$). For those who enter while they should not we do however find the same sex difference Niederle and Vesterlund found:

[9] We had randomly composed groups, ensuring that always both men and women were in the laboratory. In Niederle and Vesterlund (2007) there were always equal numbers of women and men in the laboratory and two women and two men competing in a group. We had groups of at least 8 subjects in the laboratory with a random composition of sex. Except once, there were always at least 25% of the minority sex in the lab (in one session only 22%), and competing groups were composed randomly.

Marginally significantly more men than women enter a competition while they should not (20.5% of women vs. 39.3% of men, chi square test $\chi^2 = 3.025$, $p = .082$).

4.2. Performance and choice. We now test whether those choosing competition differ in "substantial" variables from those not choosing competition, i.e., we test for differences in performance in piece rate and forced competition and the difference in improvement from piece rate to competition. As subjects also indicated which rank they believe to hold in forced competition, we can test for differences in performance beliefs between those who choose competition in round three and those who do not. We do this first for all participants together, and then split the data by gender to study differences between the sexes.

TABLE 4. Differences in performance by choice of incentive scheme

	choice	n	mean	SD
performance PR	piece-rate	85	10.16	3.99
	competition	42	11.17	4.44
performance FC	piece-rate	85	12.45	4.55
	competition	42	14.07	5.39
improvement	piece-rate	85	2.28	3.94
	competition	42	2.90	3.01

Table 4 shows performance in piece-rate and forced competition and improvement between those two treatments for those who do and those who do not choose competition. Even though those who choose competition perform slightly better on average than those who choose piece rate in both treatments, and improve slightly more, only the difference in forced competition is marginally significant (two sample Mann-Whitney-U-tests piece-rate: $z = -1.219$, $p = .223$; forced competition: $z = -1.757$, $p = .079$; improvement: $z = -1.272$, $p = .272$).

The most important difference we do find is the belief subjects hold about their performance in forced competition. Those subjects who later do choose competition have significantly more "positive" performance beliefs than those who do not choose competition (univariate ANOVA, F-test: $F = 6.886$, $p = .000$).

We now turn to the analysis of gender differences. First, we look at the performance variables and compare separately for men and women performance in piece rate and forced competition, and improvement between those who do and who do not choose competition in round three. Then, we look at performance beliefs of men and women who choose/do not choose competition. Even though those who choose competition perform slightly better both in piece-rate and in forced competition, and also improve more from piece-rate to forced competition, neither

TABLE 5. Differences in performance by gender

	choice	performance piece rate	tournament	improvement
Women	piece rate	9.56	11.67	2.11
		(3.71)	(4.16)	(3.57)
	tournament	11.11	13.50	2.39
		(5.10)	(5.49)	(2.09)
Men	piece rate	11.12	13.67	2.54
		(4.92)	(4.94)	(4.50)
	tournament	11.21	14.50	3.29
		(4.00)	(5.39)	(3.54)

Note: averages with standard deviation in parenthesis.

for men nor for women separately any of these differences is significant (see table 5).

TABLE 6. Self-Ranking in forced competition and performance

measure	self ranking FC	n	mean	SD	min	max
performance FC	1	42	16.19	4.71	7	29
	2	43	12.44	4.01	4	19
	3	33	11.73	3.29	6	20
	4	9	5.22	2.22	3	10
	total	127	12.98	4.89	3	29
improvement from	1	42	4.24	3.03	-1.00	11.00
piece rate to	2	43	2.30	3.42	-6.00	11.00
competition	3	33	1.76	3.53	-8.00	8.00
	4	9	-2.11	3.14	-8.00	1.00
	total	127	2.49	3.66	-8.00	11.00

Performance beliefs however differ both for men and women significantly between those who do and those who do not choose competition (univariate ANOVA: Women: $F = 12.936$, $p = .001$; Men: $F = 4.325$, $p = .042$). In table 6 one can see that these performance beliefs (self-ranking in forced competition) are overall related to real performance: For each performance measure applied, those ranking themselves highest indeed perform best, while those ranking themselves lowest indeed perform worst.

4.3. The Impact of the Big Five Factors. We now turn to our main research question, whether personality can explain (choice) behavior in the tournament game.

4.3.1. *Personality and Performance.* We first analyze whether personality factors are linked to performance. To test whether this is the case for our sample, we correlate performance in all three rounds with the values in the personality variables we study. Table 8 shows the correlations.

We can see in table 8 that *openness to experience* is negatively related to performance in the piece-rate setting, but not to performance in the forced competition setting. *Neuroticism* is marginally significantly negatively related to performance in the forced competition setting and highly significantly negatively related to performance in the choice setting, while *openness to experience* is marginally significantly negatively related to performance in the choice setting. A relationship between performance and some of the Big Five factors, especially *openness to experience* for piece-rate and the choice setting and *neuroticism* for forced competition and the choice setting could thus be established.

TABLE 7. Gender differences for personality factors

	sex	n	mean	SD	SE Mean
neuroticism	female	66	99.17	21.31	2.62
	male	54	84.76	24.93	3.39
extraversion	female	66	120.39	17.58	2.16
	male	54	111.78	24.54	3.34
openness	female	66	128.44	15.45	1.90
	male	54	119.98	17.70	2.41
agreeableness	female	66	109.76	19.94	2.45
	male	54	109.68	17.49	2.38
conscientiousness	female	66	117.30	21.83	2.69
	male	54	114.70	22.08	3.01

We now test whether we can replicate the gender differences in the personality variables reported in the literature. Two-sample t-tests[10] show that there are indeed gender differences in some of the personality variables in our sample. Women score significantly higher on *neuroticism* ($t = 3.424$, $p = .001$), significantly higher on *extraversion* ($t = 2.235$, $p = .027$), and significantly higher on *openness to experience* ($t = 2.795$, $p = .006$). These differences have also been mentioned in the literature (see 3).

4.3.2. *Personality and Choice.* It is noteworthy here that those personality factors where women, on average, score higher than men, have a negative impact on performance in a competitive (*neuroticism*) or a piece-rate (*openness*) setting. Therefore, in the following we test

[10]Using visual tests and also the Kolmogorov-Smirnov test of normality, we can show that for all five factors scores in the Big Five personality variables are normally distributed. Therefore we use t-tests for our analysis.

TABLE 8. Correlations

	perf PR	*perf FC*	*perf R3*	*N*	*E*	*O*	*A*
perf FC	.684***						
	(.000)						
	127						
perf R3	679***	.855***					
	(.000)	(.000)					
	127	127					
N	-.052	-.154*	-.230**				
	(.570)	(.094)	(.011)				
	120	120	120				
E	-.029	.024	-.013	-.279***			
	(.755)	(.756)	(.892)	(.002)			
	120	120	120	126			
O	-.217**	-.149	-.156*	.024	.341***		
	(.017)	(.104)	(.090)	(.789)	(.000)		
	120	120	120	126	126		
A	-.105	-.123	-.125	-.158*	.173	.064	
	(.252)	(.181)	(.174)	(.077)	(.053)	(.477)	
	120	120	120	126	126	126	
C	-.069	.046	.073	-.259**	.167*	-.067	-.037
	(.452)	(.616)	(.430)	(.003)	(.062)	(.456)	(.677)
	120	120	120	126	126	126	126

Note: coefficients, significance and number of observations. *, **, *** indicate significance at the 10%, 5% and 1% level respectively. Abbreviations: *perf = performance*, *PR = piece-rate*, *FC = forced competition*, *R3 = round 3 (choice)*, *N = neuroticism*, *E = extraversion*, *O = openness to experience*, *A=agreeableness*, *C = conscientiousness*.

whether women "know"[11] that they have certain characteristics that do not help them in a competitive environment and therefore stay out of a competition; i.e., we test whether the gender difference in personality variables disappears for those women who chose to compete.

Table 9 shows that this is indeed the case: While for women and men who do not choose to compete in round three, the gender difference in *neuroticism* is highly significant (two-sample t-test $t = 2.787$, $p = .008$), there is no significant difference between men and women who do choose to compete. This does not hold for *openness*, where there is a marginally significant difference for those who do not choose to compete and a significant difference for those who do compete. Remember, however, that *openness* mainly influenced performance in a piece-rate

[11]With this "know" we refer to knowledge in the sense that a person knows about its own characteristics and personality, and thus knows in what kind of situations he or she is successful and confident or not, and therefore decides to enter or to avoid that kind of situation. Recent research in psychology comes to the conclusion that we are not perfectly aware of our own personality, but at least we know ourselves quite well, see Vazire and Carlson (2010), Wilson (2009).

Table 9. Gender differences by choice of competition

		sex	n	mean	SD	SE mean
no	*neuroticism*	female	50	101.72	22.64	3.20
		male	31	86.77	24.74	4.44
	openness	female	50	127.22	13.93	1.97
		male	31	120.48	19.24	3.45
	risk attitude	female	50	6.64	1.86	.26
		male	32	6.75	1.72	.30
yes	*neuroticism*	female	16	91.19	14.28	3.57
		male	23	82.04	25.47	5.31
	openness	female	16	132.25	19.49	4.87
		male	23	119.30	15.78	3.29
	risk attitude	female	15	6.40	1.30	.33
		male	24	5.87	1.39	.28

setting negatively and thus, it might be rational to avoid piece-rate settings when scoring high on *openness*.

For comparison we include risk attitude here, as this has been studied by Niederle and Vesterlund (2007). We measured risk attitude with the method developed by Holt and Laury (2002). We see that no significant differences in risk attitude exist between the sexes and for those choosing or avoiding competition. Also, for the following regression results, in line with the findings in the literature, including risk aversion does not change our results.

4.3.3. *Neuroticism and the Rationality of the Decision to Enter.* As we calculated whether subjects should rationally enter the competition we use this measure to compare the scores in *neuroticism* for subjects that should rationally enter the competition, but in reality do not enter, and also subjects that entered the competition, but rationally should not enter. For the latter subjects we find no difference in the *neuroticism* score. Subjects who should rationally enter the competition but do not, score significantly higher on *neuroticism* (two-sample t-test $t = -2.081$, $p = .043$). This again confirms our result that high *neuroticism* scores have a negative influence on the willingness to compete, even for potential high performers and both for women and men.

4.3.4. *The Influence of Personality on the Choice to Compete.* In the following, we run regressions to test the robustness of the results so far reported and to test whether we can establish that *neuroticism* mediates the gender difference both in performance and in choice.

We explain the choice of competitive incentives in round 3, using a binary logistic regression, where we enter as explanatory variables in a first step sex alone, in a second step additionally all five personality factors, and in a third step more "substantial" measures: the number of correct answers in piece-rate (round 1) as a baseline performance measure, self-ranking in forced competition and improvement from piece-rate to tournament. The coefficients can be found in table 10.

In all the following regressions we will use *performance PR*, the number of correct answers in the first round with the piece-rate incentive scheme, as basic performance measure. We decided to use this measure as our baseline measure, because the measure *performance FC* (forced competition) is already in a competitive setting and thus is at least potentially influenced by the same factors that influence performance in round 3. Therefore, especially for the regression on performance in round 3, we decided to use *performance PR*.[12]

TABLE 10. Logistic regression (I) on the choice to enter a competition

	(I-1)	(I-2)	(I-3)
sex	-.182**	-.100	-.131
	(.085/.033)	(.097/.306)	(.099/.184)
neuroticism		-.005**	-.004*
		(.002/.032)	(.002/.072)
extraversion		-.003	-.003
		(.002/.151)	(.002/.130)
openness		.002	.005*
		(.003/.460)	(.003/.060)
agreeableness		.000	.000
		(.002/.877)	(.002/.886)
conscientiousness		-.002	.001
		(.002/.447)	(.002/.938)
performance PR			.004
			(.012/.708)
self-ranking FC			-.260***
			(.064/.000)
improvement			-.022
			(.015/.134)
Pseudo R^2	.03	.06	.21
n	120	120	120

Note: marginal effects at the mean with standard errors/p-values in parenthesis. *, **, *** indicate significance at the 10%, 5% and 1% level respectively.

One can see in table 10 that *sex* alone does predict the choice (I-1), but it is mediated by *neuroticism*: The effect of sex disappears when we include *neuroticism* (and the other personality factors) in the regression (I-2). In step I-3, when we include more variables, *neuroticism* remains

[12]But our results are robust, nothing substantial changes when we use performance in forced competition instead of performance in piece-rate as the baseline performance.

marginally significant, *openness* is as well marginally significant, and the *self-ranking* of the subject in forced competiton becomes the main and highly significant predictor of choice of competitive incentives.

To be more precise about the mediating effect of *neuroticism* we further did a mediation analysis following Baron and Kenny (1986). The first two steps are already included in table 10: in the first regression model (I-1) we see that *sex* is a significant predictor for the choice to compete, but is not significant in the second model (I-2) when the mediator *neuroticism* and the other personality factors are added in. Then we calculate the indirect effect which is the product of the coefficient for *neuroticism* on *sex* and the coefficient for *choice* on *neuroticism*. Using bootstrapping with 5,000 replications we find a significant indirect effect ($p = 0.042$).

4.3.5. *Performance Beliefs.* To analyze the influence of performance beliefs on the choice to compete and on our results regarding *neuroticism* we correlate the five personality factors with self-ranking in piece-rate and self-ranking in forced competition. But neither *neuroticism* nor any other personality factor is significantly correlated to any of our measured beliefs.

Using actual performance in forced competition and the belief about performance in forced competition instead of the combination of performance in piece-rate and self-ranking in forced competition, and running the same regression as the third model in table 10 leads structurally to the same results: self-ranking FC is highly significant.

We define a subject as *overconfident* when the self-estimated rank is higher than the actual rank and *underconfident* when the self-estimated rank is lower than the actual (we do not categorize those subjects who guessed the correct rank). We find that overconfident subjects have significant lower *neuroticism* scores than underconfident subjects (two-sample t-test $t = 2.371$, $p = .020$). This fits well with the fact that overconfidence is connected with more entry into the competition and with the fact that subjects scoring high on *neuroticism* avoid the competitive situation.

4.3.6. *The Influence of Personality on Performance in Competition.* We now turn to analyze the influence of personality on performance, where we examine performance in both types of competition: forced competition in round two and self-selected competition in round three. Beginning with forced competiton we run a regression with performance FC (the number of correct answers in forced competition) as dependent variable, and we enter sex in a first step as explanatory variable, then additionally all five personality factors, in a third step we add performance in piece rate (round 1), and finally we also include interaction terms. The coefficients can be found in table 11.

TABLE 11. Regression (II) on performance in forced competition

	(II-1)	(II-2)	(II-3)	(II-4)
sex	-.215**	-.176*	-.123	-.131*
	(.899/.020)	(.974/.078)	(.765/.117)	(.771/.097)
neuroticism		-.089	-.062	.150
		(.021/.388)	(.015/.396)	(.031/.331)
extraversion		.101	.039	.051
		(.022/.302)	(.019/.632)	(.019/.550)
openness		-.130	.023	.011
		(.028/.180)	(.024/.783)	(.024/.898)
agreeableness		-.146	-.067	-.074
		(.023/.106)	(.017/.310)	(.017/.271)
conscientiousness		.002	.076	.069
		(.021/.983)	(.018/.354)	(.018/.403)
performance PR			.665***	.960***
			(.089/.000)	(.229/.000)
performance PR *				-.366*
neuroticism				(.002/.095)
n	120	120	120	120
R^2	.046	.091	.497	.505
adj. R^2	.038	.043	.466	.469

Note: standardized coefficients β with robust standard errors/p-values in parenthesis. *, **, *** indicate significance at the 10%, 5% and 1% level respectively.

We find a gender effect on performance in forced competition (II-1), which reduces to marginally significant when the Big Five factors are included. None of the five factors gets significant separately, but *neuroticism* does have a (nonsignificant) negative effect on performance. Including performance in the piece-rate payment scheme into the regression (II-3) explains performance in the competitive payment scheme. In (II-4) we also include the interaction term between performance in piece-rate and *neuroticism*. This is additionally significant, and gender again is marginally significant.

We finally run a regression with the number of correct answers in round three as dependent variable. We enter again as explanatory variable in a first step sex, then additionally all five personality factors, in the third step we include the number of correct answers in piece-rate and the choice to enter the competition, and again include interaction terms. The coefficients can be found in table 12.

When we analyze performance in round three we again find that the effect of *sex* disappears when we include the Big Five factors. Here, *neuroticism* gets significant and *agreeableness* marginally significant. Having chosen a payment scheme seems to impact performance

Table 12. Regression (III) on performance in round 3

	(III-1)	(III-2)	(III-3)	(IV-4)
sex	.206**	.117	.057	-.071
	(.877/.027)	(.960/.248)	(.757/.475)	(.748/.367)
neuroticism		-.210**	-.167**	.153
		(.020/.049)	(.014/.025)	(.029/.291)
extraversion		.011	-.040	-.025
		(.019/.896)	(.018/.625)	(.018/.759)
openness		-.116	.030	.013
		(.024/.188)	(.019/.659)	(.019/.840)
agreeableness		-.152*	-.075	-.084
		(.021/.069)	(.016/.238)	(.016/.183)
conscientiousness		.010	.088	.077
		(.019/.907)	(.017/.257)	(.017/.325)
performance PR			.658***	1.113***
			(.082/.000)	(.216/.000)
*performance PR * neuroticism*				-.560***
				(.002/.005)
choice to compete			.068	.051
			(.681/.315)	(.698/.459)
n	120	120	120	120
R^2	.042	.110	.525	.541
adj. R^2	.034	.063	.491	.504

Note: standardized coefficients β with robust standard errors/p-values in parentheses. *, **, *** indicate significance at the 10%, 5% and 1% level respectively.

of those negatively who are highly neurotic, independent of the payment scheme chosen. When performance in piece-rate and the choice to enter the competition are included, *performance* in the piece-rate setting together with *neuroticism* remain significant predictors of performance. Wen we finally include the interaction between *neuroticism* and *performance*, we find than *neuroticism* is no longer significant, while *perfomance PR* still is and also the interaction has a significant negative influence.

Overall this confirms the intuition we got earlier: It is generally not women who do not self-select in the competitive treatment, but those (women) who score high on *neuroticism* - maybe knowing that this will negatively impact their performance in a competitive setting.[13]

5. Discussion

We study gender differences with respect to the choice of competitive incentive schemes and to performance in competiton in relation to

[13]One could in principal test this with the choice in round 4, but we had hardly anybody choosing to submit his or her piece-rate performance to competition.

personality variables on a behavioral level. By and large, we succeed in replicating the findings by Niederle and Vesterlund (2007), even if our setting is slightly less controlled in terms of gender composition of the competing groups. While in their case, subjects could see that there were always two women and two men in a group, in our case subjects only knew the gender composition of the whole group in the lab, with considerable variance thereof. Even though, we do find that women enter the competition less frequently than men do, and men enter the competition significantly more often if they should not than women do. In contrast to Niederle and Vesterlund, we do find an overall sex difference in performance in the competitive part and in part three of the game, and we do not find a difference between men and women with respect to not entering the competition when they ought to.

Our focus is however not on the choice of an incentive scheme per se, but on personality factors underlying this choice. Our results show that there is one of the Big Five personality variables, neuroticism, that is related to performance in and choice of a competitive context.

Neuroticism represents the tendency to be anxious, insecure and emotionally unstable in general, and to be susceptible to be stressed or depressed (McCrae and John (1992), Hogan and Johnson (1997)). In a meta-analysis looking at the link between personality and psychological disorders, Kotov*et.al.* (2010) found neuroticism to be related to post-traumatic stress disorder and major depression. High neuroticism is the key characteristic of burnout (Langelaan *et.al.* (2006), Kim, Shin, and Swanger (2009)). Neuroticism is, among others linked with difficulties in coping with conflicts and distress (Bolger and Schilling (1991), Bolger and Zuckerman (1995)). It has also been associated with impaired academic performance (e.g., Chamorro-Premuzic and Furnham (2003), Heaven *et.al.* (2002)). In a large study with about 1,000 truck drivers Andersson *et.al.* (2010) found that in the integration between personality theory and economic preferences intelligence and neuroticism play the major role.

So it seems intuitive, that people scoring high on neuroticism perform worse in a competitive setting than more emotionally stable subjects, and that they fear the stress involved and rather stay out of competitive settings.

As women on average score higher on neuroticism than men, one should expect women to enter a competition less often than men do, and to perform worse when they are forced into a competitive setting. Our findings corroborate this: Those women who do enter a competition score lower on neuroticism than women who do not enter a competition, and equal to average men. Low neuroticism women thus self-select in competitive environments, while the others stay out. Men seem to be less influenced by these factors, maybe scoring just “low enough" in general (they indeed score (non-significantly) lower than even women who do chose competition).

In our study we used as real-effort task a math task, adding two-digit numbers. Regarding performance differences of men and women, research on gender differences could establish only a few tasks were real differences in the performance of men and women could be observed (see e.g. Kimura (1999), Kimura (2004)). Summation is not among them, even if other mathematical tasks are.

While performance differences are not always and unequivocal found, stereotypes about performance of women and men in different tasks exist and have shown to affect behavior (see, e.g., Günther *et.al.* (2010), Grosse and Riener (2010), Dreber, Von Essen, and Ranehill (2011)). The notion *stereotype threat* captures the phenomenon that activation of a stereotype impairs perfomance of subjects of the negatively stereotyped group (Steele and Aronson (1995); Steele (1997)). We could imagine that this stereotype threat works intensified for women with high neuroticism scores.

We can however only speculate about the influence of neuroticism if we would use another task. As we do find some relation between underconfidence and neuroticism, it might be that a stereotypically female task would have led to other results, but we cannot say whether underconfidence is task-specific or not.

Since we also found that neuroticism has a negative impact on performance, and thereby on the choice to enter a competition we might equally well expect the same picture with a more stereotypically female task, as neuroticism could also be an obstacle in competitive situations per se. But as this is highly speculative, further research should address this question!

What do our results imply in a more general sense? It seems to be not being male or being female per se that influences whether someone likes to enter a competition or not. Rather, there are certain individual characteristics influencing performance in and preference for competitive settings which are stronger related with one gender than the other. Those scoring high on these characteristics rationally avoid competitive settings and those scoring low enough seek such settings. If we understand how these characteristics can be influenced, we might, rather than simply encouraging women to be more competitive, try to focus on these characteristics during education. Developing them in women equally as in men should be the more successful approach to achieving gender equality. Personality traits can change during lifetime, as has been shown by Roberts (2009); Roberts and Mroczek (2008). For women, there is evidence that personality changes through work experience (Roberts (1997)).

Encouraging women to enter a competition despite them being high on emotional instability might just provoke failure and thus reinforce the stereotype and discourage other women to follow suit.

Analyzing occupational attainment and relative wages Cobb-Clark and Tan (2010) conclude "our results document that women are much more likely to enter some and avoid other occupations than men with the same cognitive and non-cognitive skills. To what extent is this the result of differences in either preferences or skills that have we failed to measure?". With our results we can explain part of the gender gap by differences in personality, namely neuroticism.

Our paper represents just a first step towards a deeper understanding of the causes for women's lower willingness to compete. It shows in our view that looking for personality factors underlying the gender differences in economic behavior is a promising avenue, asking for more studies in the future.

6. Appendix

6.1. Tournament game.

Instructions. [14] In this game you will get math problems where you have to add numbers. You will receive money only for correct answers to these problems. For your calculations you are not allowed to use a calculator, but you can use scratch paper which lies on your desk.

You will be in a group with three other participants in the laboratory. We will randomly build these groups of four. You will at no point in time be informed, with whom you are in a group.

This game is divided in four parts. For each part you will get the instructions for that part at the monitor.

Payment. At the end of the experiment you will get paid for one of the four parts of the game. We will randomly determine which part is to be paid and tell you at the end of the experiment.

Generally there are two different kinds of payment: piece-rate payment and tournament. If the payment is piece-rate payment you will receive €0.50 per correct answer. In a tournament the winner in a group is the participant who solves the largest number of correct answers. The winner receives €2 per correct answer, all other participants in the group get no payment. In case of a tie, the profit is equally split between the winners.

In each part of the game you will be informed at the monitor which kind of payment there is in that part.

If you now got questions regarding these instructions, please raise your hand. One of the experimenters will come to answer your question.

[14] These instructions have been translated into English from the Original German.

6.2. Variables Used. Table 13 gives an overview of the variables used in the section 4; you can find name and a very short description.

TABLE 13. Variable Explanation

Name	**Description**
Neuroticism	score in NEO-PR-I (Big Five)
Extraversion	score in NEO-PR-I
Openness	score in NEO-PR-I
Agreeableness	score in NEO-PR-I
Conscientiousness	score in NEO-PR-I
Risk	switching point (Holt & Laury)
Self-ranking PR	estimated rank in piece-rate (round 1) $\in \{1, 2, 3, 4\}$
Self-ranking FC	estimated rank in forced competition (round 2) $\in \{1, 2, 3, 4\}$
Performance PR	number of correct answers in piece-rate
Performance FC	number of correct answers in forced competition
Improvement	improvement from piece-rate to competiton
Choice to compete	whether or not a subject choose to compete in round 3

Bibliography

Allen, D. G., K. P. Weeks, and K. R. Moffitt (2005): "Turnover intentions and voluntary turnover: the moderating roles of self-monitoring, locus of control, proactive personality, and risk aversion.," *The Journal of Applied Psychology*, 90(5), 980–90.

Allport, G. W., and H. S. Odbert (1936): "Trait-names: A psycho-lexical study.," *Psychological Monographs*, 47(1), 1–171.

Almlund, M., A. L. Duckworth, J. J. Heckman, and T. D. Kautz (2011): "Personality psychology and economics," Nber working Papers 16822.

Andersen, S., S. Ertac, U. Gneezy, J. A. List, and S. Maximiano (2011): "Gender, Competitiveness and Socialization at a Young Age: Evidence from a Matrilineal and a Patriarchal Society," `http://web.ics.purdue.edu/~smaxim/kids\_jan2011\_SE\%5B1\%5D.pdf`.

Anderson, J., S. Burks, C. DeYoung, and A. Rustichini (2011): "Toward the integration of personality theory and decision theory in the explanation of economic behavior," .

Andreoni, J. (1990): "Impure altruism and donations to public goods: a theory of warm-glow giving," *The Economic Journal*, 100(401), 464–477.

Arieli, A., Y. Ben-Ami, and A. Rubinstein (2011): "Tracking Decision Makers under Uncertainty," *American Economic Journal: Microeconomics*, 3(4), 68–76.

Baron, R. M., and D. A. Kenny (1986): "The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations," *Journal of personality and social psychology*, 51(6), 1173–82.

Barrick, M. R., and M. K. Mount (1991): "The Big Five personality dimensions and job performance: A meta-analysis," *Personnel Psychology*, 44, 1–26.

Ben-Ner, A., and F. Halldorsson (2010): "Trusting and trustworthiness: What are they, how to measure them, and what affects them," *Journal of Economic Psychology*, 31(1), 64–79.

Ben-Ner, A., F. Kong, and L. Putterman (2004): "Share and share alike? Intelligence, socialization, personality, and gender-pairing as determinants of giving," *Journal of Economic Psychology*, 25(5), 581–589.

Ben-Ner, A., L. Putterman, F. Kong, and D. Magan (2004): "Reciprocity in a two-part dictator game," *Journal of Economic Behavior & Organization*, 53(3), 333–352.

Berg, J., J. Dickhaut, and K. A. McCabe (1995): "Trust, reciprocity, and social history," *Games and Economic Behavior*, 10(1), 122–142.

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

Blanco, M., D. Engelmann, and H. T. Normann (2011): "A within-subject analysis of other-regarding preferences," *Games and Economic Behavior*, 72, 321–338.

Bolger, N., and E. A. Schilling (1991): "Personality and the Problems of Everyday Life: the Role of Neuroticism in Exposure and Reactivity to Daily Stressors," *Journal of personality*, 59(3), 355–86.

Bolger, N., and A. Zuckerman (1995): "A Framework for Studying Personality in the Stress Process," *Journal of Personality and Social Psychology*, 69(5), 890–902.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 90(1), 166–193.

Boone, C., B. De Brabander, and A. Van Witteloostuijn (1999): "The impact of personality on behavior in five Prisoner's Dilemma games," *Journal of Economic Psychology*, 20, 343–377.

Booth, A., and P. Nolen (2011): "Choosing to Compete : How Different are Girls and Boys?," *Journal of Economic Behavior & Organization*.

Borghans, L., A. Duckworth, J. Heckman, and B. Ter Weel (2008): "The economics and psychology of personality traits," *NBER Working Paper*, (December).

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.

——— (2010): "A finite mixture analysis of beauty-contest data using generalized beta distributions," *Experimental Economics*, 13(4), 461–475.

Brandstätter, H., and W. Güth (2002): "Personality in dictator and ultimatum games," *Central European Journal of Operations Research*, 10(3), 191–215.

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

BÜHREN, C., B. FRANK, AND R. NAGEL (2009): "A historical note on the Beauty Contest," https://cms.uni-kassel.de/unicms/fileadmin/groups/w_030516/BC/A_historical_note_on_the_Beauty_Contest.pdf.

BURCHARDI, K. B., AND S. P. PENCZYNSKI (2011): "Out of your mind: eliciting individual reasoning in one shot games," http://personal.lse.ac.uk/burchark/research/BurchardiPenczynski2011.pdf.

BURKS, S. V., J. P. CARPENTER, AND E. VERHOOGEN (2003): "Playing both roles in the trust game," *Journal of Economic Behavior & Organization*, 51(2), 195–216.

CAMERER, C. F. (2003a): *Behavioral Game Theory*. Princeton University Press.

——— (2003b): *Behavioral Game Theory*. Princeton University Press, Princeton, New Jersey.

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.

CAMERER, C. F., G. LOEWENSTEIN, AND M. RABIN (eds.) (2004): *Advances in Behavioral Economics*. Princeton University Press, Princeton, New Jersey.

CÁRDENAS, J.-C., A. DREBER, E. VON ESSEN, AND E. RANEHILL (2011): "Gender differences in competitiveness and risk taking: comparing children in Colombia and Sweden," SSE/EFI Working Paper Series in Economics and Finance, No. 730.

CASARI, M., AND L. LUINI (2009): "Cooperation under alternative punishment institutions: An experiment," *Journal of Economic Behavior & Organization*, 71(2), 273–282.

CASON, T. N., W. A. MASTERS, AND R. M. SHEREMETA (2010): "Entry into winner-take-all and proportional-prize contests: An experimental study," *Journal of Public Economics*, 94(9-10), 604–611.

CATTELL, H. E. P., AND J. M. SCHUERGER (2003): *Essentials of 16PF Assessment*. Wiley.

CHAMORRO-PREMUZIC, T., AND A. FURNHAM (2003): "Personality Predicts Academic Performance: Evidence from Two Longitudinal University Samples," *Journal of Research in Personality*, 37(4), 319–338.

CHEN, C.-T., C.-Y. HUANG, AND J. T.-Y. WANG (2009): "A Window of Cognition: Eyetracking the Reasoning Process in Spatial Beauty Contest Games," http://homepage.ntu.edu.tw/~josephw/SpatialBeautyContest_09July14.pdf.

CLONINGER, C. R., D. M. SVRAKIC, AND T. R. PRZYBECK (1993): "A Psychobiological Model of Temperament and character," *Archieves of General Psychatry*, 50(12), 975–990.

COBB-CLARK, D. A., AND M. TAN (2010): "Noncognitive skills, occupational attainment, and relative wages," *Labour Economics*, 18(1), 1–13.

CORICELLI, G., AND R. NAGEL (2009): "Neural correlates of depth of strategic reasoning in medial prefrontal cortex.," *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), 9163–8.

COSTA, P. T., AND R. R. MCCRAE (1992): *Revised NEO Personality Inventory (NEO-PI R) and Neo Five Factor Inventory (NEO-FFI)*. Psychological Assessment Inventories, Odessa.

COSTA-GOMES, M. A., AND V. P. CRAWFORD (2006): "Cognition and behavior in two-person guessing games: An experimental study," *The 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.

COSTA JR., P. T., A. TERRACCIANO, AND R. R. MCCRAE (2001): "Gender Differences in Personality Traits across Cultures: Robust and Surprising Findings," *Journal of Personality and Social Psychology*, 81(2), 322–331.

COX, J. C. (2004): "How to identify trust and reciprocity," *Games and Economic Behavior*, 46(2), 260–281.

CRAWFORD, V. P., M. A. COSTA-GOMES, AND N. IRIBERRI (2010): "Strategic Thinking," `http://dss.ucsd.edu/~vcrawfor/CGCI27Dec10.pdf`.

CRAWFORD, V. P., AND N. IRIBERRI (2007a): "Fatal attraction: Salience, naivete, and sophistication in experimental 'Hide-and-Seek' games," *The American Economic Review*, (5), 1731–1750.

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

CROSON, R., 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 (2011): "Gender Matching and Competitiveness: Experimental Evidence," *Economic Inquiry*.

DE QUERVAIN, D. J.-F., U. FISCHBACHER, V. TREYER, M. SCHELLHAMMER, U. SCHNYDER, A. BUCK, AND E. FEHR (2004): "The neural basis of altruistic punishment.," *Science*, 305(5688), 1254–8.

DELGADO, M. R., A. SCHOTTER, E. Y. OZBAY, AND E. A. PHELPS (2008): "Understanding overbidding: using the neural circuitry of reward to design economic auctions," *Science*, 321(5897), 1849–1852.

DOHMEN, T., AND A. FALK (2011): "Performance Pay and Multidimensional Sorting : Productivity, Preferences, and Gender," *The*

American Economic Review, 101(2), 556–590.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2010): "Are risk aversion and impatience related to cognitive ability?," *The American Economic Review*, 100(3), 1238–1260.

Dreber, A., E. Von Essen, and E. Ranehill (2011): "Outrunning the Gender Gap–Boys and Girls Compete Equally," *Experimental Economics*, 14(4), 567–582.

Dudley, N. M., K. A. Orvis, J. E. Lebiecki, and J. M. Cortina (2006): "A meta-analytic investigation of conscientiousness in the prediction of job performance: examining the intercorrelations and the incremental validity of narrow traits.," *The Journal of Applied Psychology*, 91(1), 40–57.

Dufwenberg, M., and G. Kirchsteiger (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47(2), 268–298.

D'Zurilla, T. D., A. Maydeu-Olivares, and D. Gallardo-Pujoi (2011): "Predicting social problem solving using personality traits," *Personality and individual Differences*, 50(2), 142–147.

Eckel, C. C., and P. J. Grossman (2008): "Men, Women and Risk Aversion: Experimental Evidence," *Handbook of experimental economics results*, (0316), 1–16.

Eid, M., M. Gollwitzer, and M. Schmitt (2010): *Statistik und Forschungsmethoden: Lehrbuch.* Beltz Verlag, Weinheim, Basel.

Fahr, R., and B. Irlenbusch (2008): "Identifying personality traits to enhance trust between organisations: an experimental approach," *Managerial and Decision Economics*, 29(6), 469–487.

Falk, A., and U. Fischbacher (2006): "A theory of reciprocity," *Games and Economic Behavior*, 54(2), 293–315.

Falkinger, J., E. Fehr, S. Gächter, and R. Winter-Ebmer (2000): "A simple mechanism for the efficient provision of public goods: Experimental evidence," *American Economic Review*, 90(1), 247–264.

Fehr, E., and S. Gächter (2000a): "Cooperation and punishment in public goods experiments," *American Economic Review*, 90(4), 980–994.

——— (2000b): "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, 14(3), 159–182.

Fehr, E., and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly journal of Economics*, 114(3), 817–868.

Feingold, A. (1994): "Gender Differences in Personality: A Meta-Analysis," *Psychological Bulletin*, 116(3), 429–56.

Fietze, S., E. Holst, and V. Tobsch (2011): "Germany's Next Top Manager: Does Personality Explain the Gender Career Gap?," *Management Revue*, 22(3), 240–273.

FILER, R. K. (1985): "The role of personality and tastes in determining occupational structure," *Industrial & Labor Relations Review*, 39(3), 412–424.

FISCHBACHER, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10(2), 171–178.

FUNDER, D. C. (2001): "Personality," *Annual Review of Psychology*, 52, 197–221.

GNEEZY, U., K. L. LEONARD, AND J. A. 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.

GNEEZY, U., AND A. RUSTICHINI (2004): "Gender and Competition at a Young Age," *The American Economic Review Papers and Procedings*, 94(2), 377–381.

GOLDBERG, L. R. (1981): "Language and Individual Differences: The Search for Universals in Personality Lexicons," *Review of personality and social psychology*, 2, 141–165.

GREINER, B. (2004): "An online recruitment system for economic experiments," in *Forschung und wissenschaftliches Rechnen 2003*, ed. by K. Kremer, and V. Macho, chap. GWD Berich, pp. 79–93. Gesellschaft für Wissenschaftliche Datenverarbeitung, Göttingen.

GROSSE, N. D., AND G. RIENER (2010): "Explaining Gender Differences in Competitiveness: Gender-Task Stereotypes," Jena Economic Research Papers 2010-017.

GUNNTHORSDOTTIR, A., K. MCCABE, AND V. SMITH (2002): "Using the Machiavellianism instrument to predict trustworthiness in a bargaining game," *Journal of Economic Psychology*, 23(1), 49–66.

GÜNTHER, C., N. A. EKINCI, C. SCHWIEREN, AND M. STROBEL (2010): "Women can't jump?—An experiment on competitive attitudes and stereotype threat," *Journal of Economic Behavior & Organization*, 75(3), 395–401.

GÜTH, W., R. SCHMITTBERGER, AND B. SCHWARZE (1982): "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior & Organization*, 3(4), 367–388.

HEAVEN, P. C., A. MAK, J. BARRY, AND J. CIARROCHI (2002): "Personality and Family Influences on Adolescent Attitudes to School and Self-rated Academic Performance," *Personality and Individual Differences*, 32(3), 453–462.

HERRMANN, B., C. THÖNI, AND S. GÄCHTER (2008): "Antisocial punishment across societies," *Science*, 319(5868), 1362–7.

HO, T.-H., C. F. CAMERER, AND K. WEIGELT (1998): "Iterated dominance and iterated best response in experimental" p-beauty contests"," *American Economic Review*, 88(4), 947–969.

Hogan, R., and J. Johnson (1997): *Handbook of Personality Psychology*. Academic Press.

Holt, C. A., and S. K. Laury (2002): "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), 1644–1655.

John, O. P., and S. Srivastava (1999): "The Big Five Trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research*, ed. by L. A. Pervin, and O. P. John, pp. 102–138. Guilford Press, New York, NY, US, 2nd edn.

Judge, T. A., and J. E. Bono (2001): "Relationship of core self-evaluations traits–self-esteem, generalized self-efficacy, locus of control, and emotional stability–with job satisfaction and job performance: A meta-analysis.," *Journal of Applied Psychology*, 86(1), 80–92.

Kagel, J. H., and A. E. Roth (eds.) (1995): *The Handbook of Experimental Economics*. Princeton University Press, Princeton, New Jersey.

Kahneman, D., J. L. Knetsch, and R. Thaler (1986): "Fairness as a constraint on profit seeking: Entitlements in the market," *The American economic review*, 76(4), 728–741.

Keynes, J. M. (1936): *The general theory of employment, interest and money*, vol. VII. Macmillan.

Kim, H. J., K. H. Shin, and N. Swanger (2009): "Burnout and Engagement: A Comparative Analysis Using the Big Five Personality Dimensions," *International Journal of Hospitality Management*, 28(1), 96–104.

Kimura, D. (1999): *Sex and Cognition*. MIT Press.

——— (2004): "Human sex differences in cognition, fact, not predicament," *Sexualities, Evolution & Gender*, 6(1), 45–53.

Knoepfle, D. T., J. T.-y. Wang, and C. F. Camerer (2009): "Studying learning in games using eye-tracking," *Journal of the European Economic Association*, 7(2-3), 388–398.

Kotov, R., W. Gamez, F. Schmidt, and D. Watson (2010): "Linking "big" Personality Traits to Anxiety, Depressive, and Substance Use Disorders: A Meta-Analysis," *Psychological bulletin*, 136(5), 768–821.

Langelaan, S., A. B. Bakker, L. J. Van Doornen, and W. B. Schaufeli (2006): "Burnout and work engagement: Do individual differences make a difference?," *Personality and Individual Differences*, 40(3), 521–532.

LePine, J. A., and L. Van Dyne (2001): "Voice and cooperative behavior as contrasting forms of contextual performance: Evidence of differential relationships with Big Five personality characteristics and cognitive ability," *Journal of Applied Psychology*, 86(2), 326–336.

Levine, D. K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of economic dynamics*, 1, 593–622.

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

Masclet, D., C. Noussair, S. Tucker, and M.-C. Villeval (2003): "Monetary and nonmonetary punishment in the voluntary contributions mechanism," *American Economic Review*, 93(1), 366–380.

McCabe, K. A., S. J. Rassenti, and V. L. Smith (1998): "Reciprocity, trust, and payoff privacy in extensive form bargaining," *Games and Economic Behavior*, 24(1-2), 10–24.

McCabe, K. A., M. L. Rigdon, and V. L. Smith (2003): "Positive reciprocity and intentions in trust games," *Journal of Economic Behavior & Organization*, 52(2), 267–275.

McCrae, R. R. (1982): "Consensual validation of personality traits: Evidence from self-reports and ratings.," *Journal of Personality and Social Psychology*, 43(2), 293–303.

McCrae, R. R., and P. T. Costa Jr. (1987): "Validation of the five-factor model of personality across instruments and observers.," *Journal of Personality and Social Psychology*, 52(1), 81–90.

McCrae, R. R., and P. T. Costa JR (2003): *Personality in Adulthood, a Five-Factor Theory Perspective.* Guilford Press, New York.

McCrae, R. R., and O. P. John (1992): "An Introduction to the five-factor Model and its Applications.," *Journal of personality*, 60(2), 175–215.

Mischel, W. (1968): *Personality and assessment.* Wiley, New York.

——— (1977): "The interaction of person and situation," in *Personality at the Crossroads: Current Issues in Interactional Psychology*, ed. by D. Magnussen, and N. Endler, pp. 333–352, Hillsdale, NJ. Lawrence Erlbaum.

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

Moulin, H. (1986): *Game Theory in the Social Sciences.* New York University Press.

Mueller, G., and E. Plug (2006): "Estimating the Effect of Personality on Male and Female Earnings," *Industrial and Labor Relations Review*, 60(1), 3–22.

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

Nagel, R. (1993): "Interactive Competetive Guessing," Bonn Working Paper.

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

——— (1999): "A survey of experimental guessing games: a study of bounded rationality and learning," in *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, ed. by D. V. Budescu, I. Erev, and R. Zwick, pp. 105–145.

Nicholson, N., M. F. O'Creevy, E. Soane, and P. Willman (2005): "Personality and domain-specific risk taking," *Journal of Risk Research*, 8, 157–176.

Niederle, M., C. Segal, and L. Vesterlund (2010): "How Costly is Diversity ? Affirmative Action in Light of Gender Differences in Competitiveness," NBER Working Paper No. 13923.

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.

——— (2010): "Explaining the Gender Gap in Math Test Scores : The Role of Competition," *Journal of Economic Perspectives*, 24(2), 129–144.

Nikiforakis, N., and H.-T. Normann (2008): "A comparative statics analysis of punishment in public-good experiments," *Experimental Economics*, 11(4), 358–369.

Norman, W. T. (1963): "Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings," *The Journal of Abnormal and Social Psychology*, 66(6), 574–583.

Ohtsubo, Y., and A. Rapoport (2006): "Depth of reasoning in strategic form games," *Journal of Socio-Economics*, 35, 31–47.

Ostendorf, F., and A. Angleitner (2004): *NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (NEO-PR-I)*. Hogrefe, Göttingen.

Ostrom, E., J. Walker, and R. Gardner (1992): "Covenants With and Without a Sword: Self-Governance is Possible," *The American Political Science Review*, 86(2), 404–417.

Ozer, D. J. (1985): "Correlation and the Coefficient of Determination," *Psychological Bulletin*, 97(2), 307–315.

Ozer, D. J., and V. Benet-Martínez (2006): "Personality and the prediction of consequential outcomes.," *Annual review of psychology*, 57(8), 401–21.

Paunonen, S. V., and M. C. Ashton (2001): "Big Five factors and facets and the prediction of behavior," *Journal of Personality and Social Psychology*, 81(3), 524–539.

Rabin, M. (1993): "Incorporating fairness into game theory and economics," *The American Economic Review*, 151(3712), 867–868.

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

Roberts, B. W. (1997): "Plaster or plasticity: are adult work experiences associated with personality change in women?," *Journal of Personality*, 65(2), 205–32.

——— (2009): "Back to the future: Personality and assessment and personality development," *Journal of research in personality*, 43(2), 137–145.

Roberts, B. W., and D. Mroczek (2008): "Personality Trait Change in Adulthood," *Current directions in psychological science*, 17(1), 31–35.

Sbriglia, P. (2008): "Revealing the depth of reasoning in p-beauty contest games," *Experimental Economics*, 11, 107–121.

Schmitt, D. P., A. Realo, M. Voracek, and J. Allik (2008): "Why can't a Man be More like a Woman? Sex Differences in Big Five Personality Traits Across 55 Cultures," *Journal of personality and social psychology*, 94(1), 168–82.

Seibert, S. E., and M. L. Kraimer (2001): "The Five-Factor Model of Personality and Career Success," *Journal of Vocational Behavior*, 58(1), 1–21.

Simon, H. A. (1955): "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69(1), 99–118.

Smith, A. (1759): *The Theory of Moral Sentiments*. Edinburgh.

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.

Steele, C. M. (1997): "A threat in the air: How stereotypes shape intellectual identity and performance.," *American psychologist*, 52(6), 613–629.

Steele, C. M., and J. Aronson (1995): "Stereotype threat and the intellectual test performance of African Americans.," *Journal of personality and social psychology*, 69(5), 797–811.

Sutter, M., and D. Rützler (2010): "Gender Differences in Competition Gender Differences in Competition Emerge Early in Life," IZA Discussion Paper Series 5015.

Swope, K. J., J. Cadigan, P. M. Schmitt, and R. Shupp (2008): "Personality preferences in laboratory economics experiments," *Journal of Socio-Economics*, 37(3), 998–1009.

van den Bos, W., J. Li, T. Lau, E. Maskin, J. D. Cohen, P. R. Montague, and S. M. McClure (2008): "The value of victory: social origins of the winner's curse in common value auctions.," *Judgment and decision making*, 3(7), 483–492.

Vazire, S., and E. N. Carlson (2010): "Self-Knowledge of Personality: Do People Know Themselves?," *Social and Personality Psychology Compass*, 4(8), 605–620.

Volk, S., C. Thöni, and W. Ruigrok (2011): "Personality, personal values and cooperation perferences in public goods games: A longitutinal study," *Personality and individual Differences*, 50(6), 810–815.

Wang, J. T.-Y., M. Spezio, and C. F. Camerer (2010): "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *American Economic Review*, 100(3), 984–1007.

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

Wilson, T. D. (2009): "Know thyself," *Perspectives on Psychological Science*, 4(4), 384–389.

Witteman, C., J. van den Bercken, L. Claes, and A. Godoy (2009): "Assessing Rational and Intuitive Thinking Styles," *European Journal of Psychological Assessment*, 25(1), 39–47.

Wozniak, D., W. T. Harbaugh, and U. Mayr (2010): "Choices About Competition: Differences by gender and hormonal fluctuations, and the role of relative performance feedback," .

Zizzo, D. J. (2003): "Inequality and Procedural Fairness in a Money Burning and Stealing Experiment," *Research on Economic Inequality*, 11(155), 215–247.

——— (2010): "Experimenter demand effects in economic experiments," *Experimental Economics*, 13(1), 75–98.