Essays on Social Preferences in Experimental Economics

Dissertation

zur Erlangung des akademischen Grades

Doctor Rerum Politicarum

An der

Fakultät für Wirtschafts- und Sozialwissenschaften

der Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Hannes Rau

Heidelberg, Februar 2019

Acknowledgements

During my studies and research at the University of Heidelberg, I met and worked with many people, to whom I am very grateful. Without their support this dissertation would not have been possible. They contributed towards this thesis and inspired me in many different ways. I would like to take this opportunity to individually thank everyone.

First of all, I am immensely grateful to my supervisor Christoph Brunner, who constantly supported and encouraged me during my thesis. He always took the time to help me, by providing perceptive feedback and suggestions as well as moral support. Our discussions and his deep knowledge greatly improved the quality of my research projects. I am also very grateful to Christiane Schwieren, who kindly agreed to be my second supervisor. She gave me the freedom to develop my own ideas, but always offered her help and advice when needed. We had very productive discussions on a regular basis and an excellent atmosphere at the chair.

Furthermore, I would like to thank Jürgen Eichberger, at whose chair I have been employed in the past years. He provided valuable advice and especially encouraged and supported me during the final stage of my dissertation. I am also very thankful to Graciela Kuechle, Jörg Oechssler and Stefan Trautmann with all of whom I have had the opportunity to collaborate. Working with them was a very positive and valuable experience, which helped me to become a better researcher. I would like to give special thanks to Florian Kauffeldt, who co-authored a project of this thesis and with whom I had a great time when we worked together at the Department. I would like to extent my thanks to my other co-authors Dietmar Fehr, Alex Roomets and Yilong Xu. We had a very friendly and productive collaboration.

I am indebted to Christoph Becker for providing important help with formatting the thesis, to Christian König for sharing his programming skills and to Andis Sofianos, who gave valuable comments and helped improve my English writing style and grammar.
My thanks extent to my past and current colleagues from the institute, Peter Dürsch, Sara Kettner, Christopher Kops, Marco Lambrecht, Illia Pasichnichenko, Gert Pönitzsch, Robert Schmidt, and Martin Vollmann. We had very stimulating discussions, a very friendly working atmosphere and a lot of fun together.
I would also like to thank Ulrike Arnold, Barbara Neef, Marcus Padberg and Gabi Rauscher, whose efforts made my time at the institute both productive and very pleasurable. Financial support from the University of Heidelberg and Heilbronn University for conducting experiments and presenting our work at conferences is gratefully acknowledged.
Finally and above all, I would like to thank my family for everything.

Contents

Introduction

The dissertation consists of four experimental studies where social preferences can act as a motivational factor for individual decisions. "Social preferences" or "other-regarding preferences" refer to a class of preferences, in which the decision maker does not only take the consequences for herself into account, but also those for other persons. Additionally, the interaction with other parties might play a role for how the consequences are evaluated by the decision maker. We distinguish social preferences from "selfish preferences", where the decision maker only cares about her own outcomes.[1]

There is abundant evidence for social preferences in the field of experimental economics: Across various experimental paradigms and examples including Dictator Games, Ultimatum Games, Trust Games and Prisoner's Dilemmas subjects frequently do not chose their payoff maximizing strategy. Furthermore, in studies on the distribution of financial resources, subjects often prefer allocations, in which they do not receive the highest possible payoff for their own.

Initially, models trying to explain these type of preferences mainly used an *outcome-based* approach. As the name suggests, these approaches focus on the resulting distribution of payoffs of the decision maker and other involved parties. They do not explicitly model, how the distribution is generated. The most influential examples of this class are the models of E. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). In these kind of models, the utility of a decision maker is derived from his own payoff as well as how it compares to the other players' payoffs. These models are fairly tractable and can explain a variety of choices, where players forgo own payoffs in order to achieve a more egali-

[1] As in the context of experimental economics the outcomes usually correspond to monetary payoffs, we will use these terms synonymously.

tarian distribution. Beyond critiques on the functional form[2] and the parametrization, other studies, as for example Engelmann and Strobel (2004), question the assumption, that difference aversion is the driving force behind the behaviors we observe. They provide evidence for behavior being motivated to a greater extent by concerns for efficiency (maximization of total payoffs) and "maximin preferences" (increasing the welfare of the least well-off player). In response to this critique, both Bolton and Ockenfels (2006) and E. Fehr, Naef, and Schmidt (2006), argue that it boils down to a tradeoff between equity and efficiency, which players are facing.

In reaction to evidence, that behavior is not solely motivated by the resulting payoffs, a second class of models, so called *intention-based* models, have emerged. The basic concept is that players reward kind actions and punish unkind actions. The idea of examining players' intentions is already found in papers of Rabin (1993) and Blount (1995). Rabin proposes an alternative equilibrium concept, named "fairness equilibrium". This concept allows to take other players' actions and beliefs into account and can model reciprocal behavior. Blount studys intentions in the context of an Ultimatum Game. The author shows, subjects do not penalize other players for unequal offers, if the proposer is not responsible for those. The most influential model in this class is Charness and Rabin (2002). In their model, the decision maker assigns different weights to the payoffs of other persons, which depend on their behavior towards the decision maker. In that way, the model provides a richer framework as previous models, because it is both able to capture a variety of distributional preferences and to account also for motives such as reciprocity. There are though some open questions left, as for example, which kind of behavior is seen as kind or unkind. An encompassing definition is hardly possible. Usually social norms and beliefs of the interacting persons play an important role here. Following this concept, a lot of different extensions emerged, as for example Dufwenberg and Kirchsteiger (2004), who apply the concept of reciprocity to sequential games and Falk and Fischbacher (2006), who incorporate both players' actions and intentions into their model.

[2]For example in E. Fehr and Schmidt (1999), the amount of disutility a decision maker faces is constant for each monetary unit lagging behind the reference group.

Despite various attempts to explain behavior, there are still many further relevant factors, which are difficult to incorporate into a single model. In the following, the more important ones are briefly discussed:

There is clear evidence that facing the same situation repeatedly can lead to changes in behavior, thus *learning effects* might exist. For example, even selfish players increase their offers in Ultimatum Games or Bargaining Games, if their previous offers were mostly rejected (see Cooper & Dutcher, 2011 for a meta-analysis). Another evidence is how cooperative players switch to defection in Prisoners' Dilemma Games, after they are faced with several repetitions of defection. Further aspects are *procedural fairness* and *diffusion of responsibility*. In a sequential Battle-of-Sexes Game, Bolton, Brandts, and Ockenfels (2005) find that even unfair outcomes are accepted if they are the result of a fair procedure. Thus, a lot depends on how the action of a person is perceived by the other player(s), which links to the role of intentions. We will examine the consequences of procedural fairness in more detail in chapter 1.

Another question, which is usually ignored, is to specify exactly whose payoffs are decision-relevant. By convention, in experiments one often refers to the persons, with whom the player is directly interacting. But this reasoning is not entirely convincing. One could also argue that the whole group of participants forms the reference group or the reference group is even constituted by persons outside the lab.[3] To assess this factor, there are many studies about how behavior is influenced by group identity. They provide a rather mixed picture of results (see for example Chen & Li, 2009 for an overview). There are many more aspects to consider, which refer to the *context of the decision-making situation*, such as framing effects, the influence of socioeconomic variables, the choice set, but we will not discuss them in more detail here.

In all of the upcoming studies the focus is on situations in which groups of two players interact with each other in a way, such that their behavior mutually can influence each others' payoffs. Given the evidence from previous studies, social preferences play an

[3]One participant in the study about antisocial behavior justified his decision to destroy money of some other participant with the argument, that in his point of view the money is not lost, as it is belonging to the taxpayer and can be used for better purposes than paying students for this study.

important role for subjects' decisions in this type of situations.

The structure of the thesis and research questions of each projects are as follows: In chapter 1 we report results of an experiment examining how the fairness of the endowment generating process influences subjects' propensity to costly destroy money of other participants. Chapter 2 is co-authored with Graciela Kuechle. There we analyze results of a study about gender differences in bargaining behavior in an environment in which subjects either have or do not have knowledge about the gender of the involved parties. In chapter 3 we investigate if subjects more often play a Nash equilibrium strategy, if they are mutually informed about each other's preferences over payoff tuples which correspond to the monetary outcomes of the games. This chapter is based on collaborative work with Christoph Brunner and Florian Kauffeldt. Finally, chapter 4 deals with the question of how the degree of "social capital", measured by subjects' behavior in a trust game, is affected by being exposed to different payment schemes beforehand, which vary in their degree of competitiveness. This chapter is co-authored with Dietmar Fehr, Stefan Trautmann and Yilong Xu. The last chapter summarizes the results and offers some concluding remarks on the impact social preferences have on decision making.

Chapter 1

Antisocial Preferences

Do people exhibit more antisocial behavior if the income allocating process has been unfair?

Abstract

We examine whether an unfair process of income allocation leads to a higher degree of antisocial behavior. In order to test this hypothesis, we run an experiment where we vary the way players' endowments are determined: by a fair, random, or unfair process. The initial distribution has a certain degree of inequality, which is held constant across treatments. After receiving their income, subjects can anonymously reduce the income of another player at a cost. The overall frequency and percentage of destruction is similar and not significantly different across treatments. Surprisingly, even if money is allocated in an unfair manner, subjects do not destroy more. We furthermore elicit subjects' perceptions about the fairness of the income-generating process. They are in line with the intended treatment effect, but we find almost no correlation between subjects' fairness evaluation and the propensity to burn money. The findings indicate that the degree of antisocial behavior is rather constant in this context and independent of the fairness of the income-allocating process. Subjects' justifications of their decision and insights of related studies suggest that the decision to destroy other's income depends a lot on whether other subjects can be held responsible for the initial (unfair) distribution.

1.1 Introduction

Many economic experiments have demonstrated that subjects exhibit antisocial behavior. The term *antisocial behavior* in this context is used in the sense that subjects destroy each other's income without any material benefits to themselves, even if it is costly. The amount burned in this way varies a lot, and depends on the exact framework in which subjects are acting. The percentage of destruction in such experiments ranges from below 5% to over 60% of total endowment/income.[4] Previous studies focus on the effect of variables like *the costs of burning* (Zizzo & Oswald, 2001), *the degree of anonymity* (Abbink & Sadrieh, 2009) and *relative income positions* (Abbink, Masclet, & van Veelen, 2011 as well as Grossman, Komai, et al., 2013), just to name a few. For some factors such as anonymity and the price of burning, the results seem expectable and quite clear: Abbink and Sadrieh (2009) find that a higher degree of anonymity leads to more money burning. In the experiment of Zizzo and Oswald (2001) subjects burn significantly more money, when costs are very low (2.5% and 5% of the amount destroyed) compared to a higher price (25%). However, for other factors such as the relative income position, there are contradictory results so far concerning the amount of destruction: Abbink et al. (2011) show that subjects with a similar rank in the income distribution destroy more, while Dawes et al. (2007) find an opposite effect.

These studies suggest that the degree of destruction depends a lot on the context of the situation. In most studies discussed above, there exists a certain degree of inequality in the initial income distribution. This potentially creates a conflict: Under some circumstances these differences might be accepted, while on others subjects might have the desire to change the initial unequal allocation. Taking into account prominent theories of justice, one key factor is **the fairness of the money allocating process**. This aspect has not been examined in the context of antisocial behavior so far.

In this study we investigate, if the degree of antisocial behavior can depend on "how fair" subjects view the money allocating process. The natural hypothesis we want to test is, if participants burn more money, when the endowment allocating process has been unfair,

[4] In the context of this study we use the terms endowment and income synonymously.

compared to when it has been fair.

If we use the term *fair* in the context of our study, e.g. for labeling the treatments and forming the hypothesis, it mostly corresponds to the concept called *process fairness/procedural fairness*. Concretely this means that subjects' incomes are positively correlated to their time and effort spent during the experiment. We also use the term to describe the subjective fairness evaluations of our participants concerning the endowment-allocating process. These fairness evaluations are also based on the concept of process fairness and the respective question is framed accordingly. Process fairness has to be distinguished from the one of *outcome/distributional fairness*. Both play an important role in assessing a distribution of resources. Distributional fairness emphasizes the resulting distribution, while process fairness focuses on the way how this distribution is generated. So one can say, process fairness takes more an ex ante view, while outcome fairness judges the result more ex post. Consider for example a lottery, which awards the winner a high prize and the rest receives nothing. If everyone has the same probability of winning, this mechanism satisfies the criterion of process fairness, but not that one of outcome fairness.

Generally, in economic experiments, endowment is mostly provided in the form of windfall gains -the most common way- or as earned money. In the latter case participants usually perform a (real-effort-) task according to which their initial income is assigned. Evidence suggests the process determining initial income affects subsequent subjects' behavior: In an Ultimatum Game lower offers are accepted if the proposer has earned his role by winning a logic game (Hoffman et al., 1994). Similarly, Oxoby and Spraggon (2008) observe larger transfers in a Dictator Game, if the recipient has created the amount of money to be divided by his performance in a problem solving task.

Most fairness norms and theories on procedural justice emphasize the role of effort as an important factor justifying differences in the distribution of (financial) resources (e.g., Konow, 2003, p. 1207). In contrast there are differences resulting from luck, or factors which are congenital such as abilities or talents. Usually people favor a more even distribution if initial differences result from external reasons (e.g. a handicap) compared to internal reasons (e.g. lower effort provided or bad decisions made in the past) (see

Faravelli, 2007 and Konow, 2001). Roemer (1998) goes even one step further and argues, that effort can also be partially seen as some characteristic of type, for which people cannot be held responsible completely. For further information on theories of distributional justice see Cappelen et al. (2007) and for an overview of results in empirical social choice see Gaertner and Schokkaert (2012).

A closely related paper is D. Fehr (2018). The author examines in a lab experiment the relationship between increasing inequality and the tendency to burn other's income. He finds that increasing inequality leads to more money burning, but only if the underlying process creating this higher inequality is unfair. In contrast, if higher inequality can be unambiguously attributed to higher effort, subjects do not destroy more income.

Taking all this together, a possible conjecture can be that the perceived fairness of endowment determination is an important factor influencing the decision to change/destroy others' income.

The results of this study are important for the real world. One can think of many situations, in which resources are assigned by different procedures, which vary in their perceived fairness. For example companies provide different remuneration schemes depending on individual performance. Individual performance is though often not easy to measure and its value contribution can be hard to disentangle. Thus, there can be great heterogeneity of the acceptance of the income determining mechanism. If this process is perceived as unfair, it can trigger negative emotions and harm individual productivity. Hence, from society's point of view it is an essential question, if the resulting allocation is approved or not. The main difference to our experiment probably is that antisocial behavior usually is not directly possible in the real world. But in some cases even (costly) sabotage might occur.

1.2 Experimental Design

The experiment is pen and paper based and consists of three treatments:

1. **Treatment "Fair"**
2. **Treatment "Random"**
3. **Treatment "Unfair"**

All treatments have the same basic structure, consisting of two parts: First, the endowment determination phase and in the second part the destruction decision. Finally, subjects fill out a questionnaire and then receive their final payoff. As endowment, half of the participants receive a **high (€10)** and the other half a **low (€5) amount of money**. The only difference between the treatments is the way in which participants are assigned these values:

In the random treatment, the endowment is determined by a lottery. Subjects pick up a sealed envelope from a box containing a note that they either are allocated €5 or €10. They are told that 50% of the participants receive the high amount and the other 50% the low one. This treatment is designed as a "baseline", to be comparable with most of the other money burning experiments with similar parameter values, in which endowment is provided in the form of "windfall gains". In the other two treatments, the subject pool is divided into two groups, the "early group" and the "late group". Members of the early group have to perform a real effort task, involving correcting IQ-tests from another experiment. The subjects of the late group are told to show-up 30min later for the experiment and do not have to do any work. In the fair treatment all members of the early group receive €10 as endowment and participants from the late group are assigned €5. The payment scheme in the unfair treatment is exactly reversed: That means, subjects showing up early only receive €5 and the others, who do not have to do any work, obtain the higher value of €10. The whole procedure is common knowledge. That means, in each treatment all subjects are informed what kind of "task" both groups of the respective treatment have to complete and how they are rewarded for it.

After the phase of endowment determination each subject is given the opportunity to spend a fraction of his or her income to anonymously destroy some or all money of a randomly selected other participant. The second part of the experiment is identical across all treatments. The costs of burning money are 10% of the chosen amount, so for every Euro a subject wants to destroy, he or she has to pay 10 Cents. Subjects can choose any value ranging from zero up to the total income of the other person (either €5 or €10). Furthermore, in the instructions it was pointed out, destruction is only optional and one does not necessarily have to subtract any money. Framing here was as neutral as possible to avoid experimenter-demand effects in any direction. All decisions are made anonymously. For that purpose subjects generate a code, which corresponds to their identity and decisions during the experiment. Destruction decisions are made using strategy method: Every subject indicates which amount she or he would like to reduce, if the other person has an income of €5 or €10. Remember that subjects know which task the other participant (with an endowment of €5 and €10) had to perform beforehand. Afterwards, participants are randomly paired in groups-of-two. In each of these groups, only one of the two decisions is actually carried out ("unilateral destruction"). So in the end only half of the destruction decisions are implemented. The idea is to prevent motives like preemptive retaliation or negative reciprocity. If both decisions were to be implemented, it would be possible, that some players would not want to burn any money at all, but have the belief the others would do so and therefore would want to preemptively retaliate. This is maybe one of the reasons for surprisingly high burning rates in experiments, like for example, in Zizzo and Oswald (2001) or the occurrence of vendettas in repeated money burning games Bolle et al. (2014). After everyone has made his or her choice, the experimenters randomly draw which decisions of the groups are actually implemented and then calculate the resulting payoffs. Meanwhile, subjects fill out a questionnaire, for which they receive additional money (€3). In the questionnaire subjects are asked, as how fair they rate the money allocating process and what are their motives for burning (or not burning) money. At last, subjects receive feedback about which decisions are carried out and their final payoff.

The experiment was conducted in 2013 in the experimental lab of the University of Heidelberg. Overall 119 subjects, mostly students, took part (42% had their major in Economics). Recruitment was carried out with ORSEE (Greiner, 2015). Each treatment included 3 sessions, with an average duration of 45-60min. Further details are summarized in table 1.1 below. Average earnings were around €10 per subject, with payoffs ranging from €3-13.

Table 1.1: Treatment overview

Treatment	Procedure	Sessions	Subjects
Fair	Real-effort task	3	41
Random	Lottery	3	34
Unfair	Real-effort task	3	44

1.3 Hypotheses

The first hypothesis is about whether subjects evaluate the fairness of the process of endowment determination differently across treatments.

Hypothesis 1: *The fair treatment is rated as fairer than the random treatment, which again is rated as fairer than the unfair treatment.*

Hypothesis two compares destruction decisions between the treatments.[5] Considering the discussion above about social norms and theories of distributional justice, it is natural to suppose that people destroy less money, if they perceive the endowment allocating process as rather fair . This hypothesis is also supported by concepts from social psychology such as "Equity Theory" by Adams (1963). Accordingly, people tend to accept income differences as long as the proportion of effort to payoff is similar across all participants. If people perceive this relation as unbalanced, they experience negative emotions, leading to actions restoring a more even situation. In our setting, we would expect subjects viewing the income allocation as unfair to reduce the income of other players to some

[5]The argument in this paragraph is based on the assumption that subjects actually evaluate the endowment-allocating process in the treatments as differently fair. This assumption will be confirmed later on. Especially the process in the unfair treatment is rated as clearly less fair than the one in the other two treatments.

extent until subjective equity is restored. In the unfair treatment of our experiment this relation is clearly more unbalanced than in the fair one. Therefore one would expect higher destruction rates there to reduce differences in the effort to payoff-ratio. In a similar fashion, models of Inequality Aversion (as e.g. E. Fehr & Schmidt, 1999 predict this kind of behavior, if effort is seen as some form of monetary costs).

Hypothesis 2: *Most destruction is chosen in the unfair treatment unfair and least in the fair treatment.*

To explain the composition of overall destruction, one has to look more into detail of the behavior of specific income classes. There are four different cases to distinguish, concerning the endowment of the decision maker and the target of destruction: (low, low), (low, high), (high, low), (high, high). The first value refers to the endowment of the decision maker and the second to the endowment of the target. According to the theories mentioned before, a substantial part of the predicted differences in overall destruction should stem from the combination (low, high). In this combination there are the highest differences in the effort-to-payoff ratio, especially in the unfair treatment. This leads to hypothesis 3:

Hypothesis 3: *Destruction in the combination (low, high) is highest in the unfair treatment and lowest in the fair treatment.*

In the cases, in which both subjects have the same endowment (low, low) and (high, high) the situation is completely symmetric. Both parties have to perform the same kind of task and receive the same reward for it. There might be a certain number of subjects having inherently antisocial preferences. As there is no clear reason why this number should differ depending on the treatment, we expect no treatment effect in these combinations.

Hypothesis 4: *Destruction in the combinations (low, low) and (high, high) is the same across all treatments.*

Finally for the combination (high, low), most theories do not predict subjects with a higher endowment would destroy earnings of those with a lower endowment. Only in the fair treatment it is possible to imagine a subject receiving €10 to feel disadvantaged. This would be the case, if she evaluates the additional €5 of endowment as inferior to the effort

she had to exert beforehand.

Hypothesis 5: *Destruction in the combination (high, low) is highest in the fair treatment and lowest in the unfair treatment.*

One could refine hypothesis 5 by adding, that destruction in the combination (high, low) is lower than in the combination (low, high). But the comparison of these two cases is more complicated, because the amount of endowment of the target differs. Therefore one cannot really compare absolute values in destruction between these two cases. Possible solutions might be, to look at the percentages or frequencies of destruction here.

1.4 Experimental Results

1.4.1 Fairness Evaluations

We first check if the treatments work as intended. In hypothesis 1 we anticipate that participants will find the unfair treatment least fair. To test this hypothesis we ask all subjects at the end of the experiment to evaluate as how fair they perceived the endowment assigning mechanism in their treatment. They can rate the process on a scale ranging from 1-5, where 1 means they perceived the mechanism as "very fair", 3 corresponds to "neutral" and 5 to "very unfair". The results are shown in table 1.2 below.

Table 1.2: Fairness evaluations

Treatment	Subjects	Fairness Evaluations(average)
Fair	41	2.4
Random	34	2.5
Unfair	44	4.3^{***}

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As expected, the participants rate the unfair treatment as clearly less fair than the other two treatments. The scale should be interpreted as ordinal, therefore we use a Wilcox-rank-sum test to compare the fairness perceptions pairwise between treatments. The difference of treatment Unfair compared to each of the other ones is highly significant

($p < 0.001$). While there is no significant difference between treatment Fair and Random ($p = 0.96$). A Kruskal-Wallis rank sum test of comparing all treatments simultaneously leads to the same result. The results (partly[6]) support hypothesis 1:

Result 1: *The fair treatment is rated as similarly fair as the random treatment. Both are rated as much fairer than the unfair treatment.*

1.4.2 Destruction Decisions (Pooled)

In our experiment we measure destruction in two ways: Either as **destruction frequency** or as **percentage of destruction**. The first case corresponds to the number of decisions, in which one subject wants to reduce the payoff of another subject, divided by the total number of decisions. The latter case corresponds to the amount of money intended to burn divided by the endowment of the other subject. The hypotheses in general refer to both measures.

Overall destruction is moderate and less pronounced compared to other money burning experiments.[7] As explained before we measure destruction activity in two ways: Destruction frequency and percentage of destruction. We have two destruction decisions per subject. This amounts to a total of 238 decisions. The average burning frequency over all treatments is 23.5% and the average percentage of destruction is 10.8%. Results for each treatment are summarized in table 1.3 and illustrated in figure 1.1.

For the analysis we focus on the percentage of destruction, as this variable contains additional information compared to the mere frequency. Note again, that subjects make their decisions using the strategy method, specifying for each of the two cases (the other player has either an endowment of €5 or €10) the amount they would like to reduce from her. For the computation of our variables (burning frequency and percentage of destruction) we take both decisions into account, no matter, if they are actually carried out or not. Therefore the percentage of destruction per subject is calculated by adding up

[6] As the difference in evaluations between the fair and the random treatment is not significant, the hypothesis is only partly confirmed. Maybe the difference between these two treatments would have been bigger, if participants also would have been told the allocation mechanisms of the other treatments.

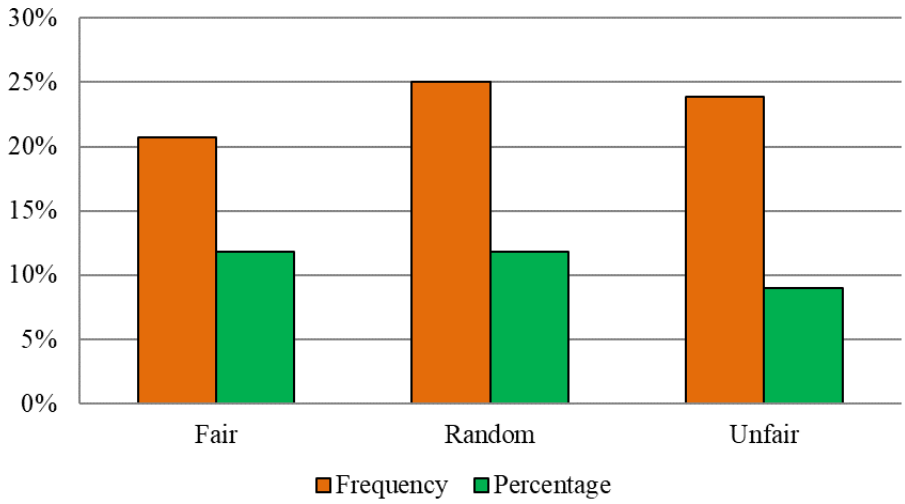
[7] For example in Zizzo and Oswald (2001) about 70% of subjects burn money at a cost of 10% of the chosen amount and in Abbink and Sadrieh (2009) ca. 25% do so, but at higher costs of 20%.

Table 1.3: Destruction decisions

Treatment	Number of decisions	Destruction frequency	Percentage of destruction
Pooled	238	23.1% (0.39)	10.7% (0.21)
Fair	82	20.7% (0.40)	11.8% (0.26)
Random	68	25.0% (0.41)	11.8% (0.21)
Unfair	88	23.9% (0.37)	9.0% (0.15)

Standard errors (absolute values) in parentheses.

Figure 1.1: Overall destruction frequency and percentage of destruction



both intended destruction values and dividing them by the sum of the target endowments (which corresponds to €5 + €10 = €15 in every case).

For the statistical analysis we use a Wilcoxon-rank sum/Mann-Whitney test (for pairwise comparisons of treatments) and a Kruskal-Wallis test to compare destruction decisions simultaneously across all treatments. We first look at the pooled values and later examine the behavior of specific income classes both of the decision maker and the target. We additionally run a linear regression with the amount of destruction as outcome variable and the treatments as dummy variables (+controls). We do this both for pooled values and for controlling for specific income classes. Results of the regressions can be found in section

1.4.5. For the discussion and conclusion we focus on the results of the Wilcoxon-rank sum test, as some of the underlying assumptions of the regression analysis are not fully met (such as e.g. the assumption that the error terms are normally distributed). Looking at the pooled values, there are no significant differences across treatments concerning the amount of destruction (Kruskal-Wallis test, $p = 0.74$). A pairwise comparison of treatments leads to similar results. In the unfair treatment, average values are even slightly lower than in the random one.

Result 2: *There are no significant differences concerning overall destruction rates across treatments.*

There are two potential explanations: Either the treatments do not have a strong effect on individual decisions to destroy somebody else's income, or we have multiple effects going in opposite directions and balancing each other on average.[8] To check the second point we take a closer look at the behavior of specific income classes.

1.4.3 Behavior of Specific Income Classes

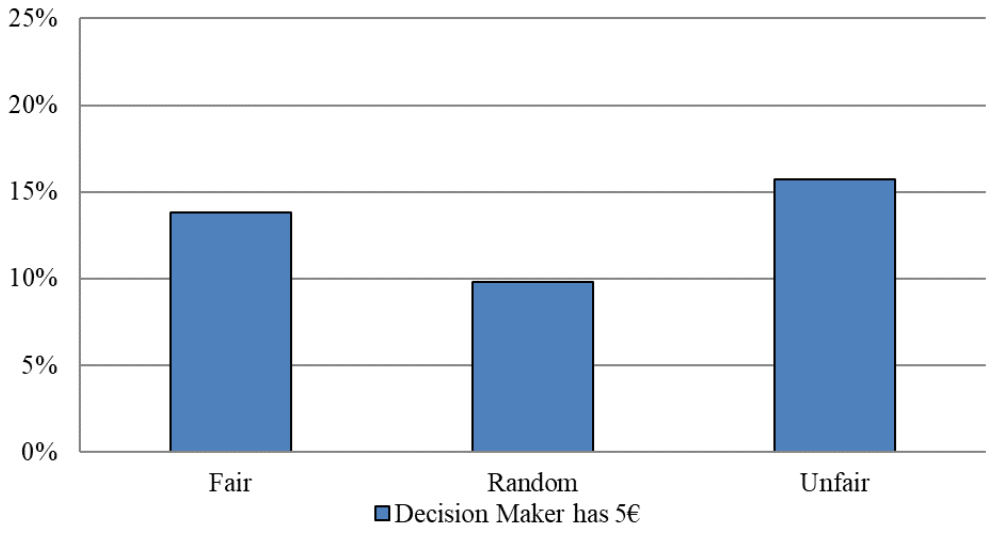
Case 1: Combination (low, high)

In the first case we examine destruction behavior of subjects receiving a low endowment (€5) targeting subjects with a high endowment (€10). Average percentage destruction[9] per treatment is displayed in figure 1.2 below. Destruction rates are highest in the unfair treatment, but they do not differ much across treatments. Comparing treatments pairwise using a Wilcoxon rank-sum test, we find no significant differences in destruction rates.

[8] It could for example be the case that in the unfair treatment subjects burn more money from players with a high endowment of €10 compared to the other treatments, but less from players with the lower value of €5. This would overall also lead to similar destruction rates across treatments.

[9] Note again that these are all intended values as the actual destruction, which is finally implemented is randomly determined.

Figure 1.2: Average percentage destruction if target has €10



Result 3: *There are no significant differences across treatments concerning destruction rates for the combination (low, high).*

Hypothesis 3, in which we expected substantial differences in destruction rates across treatments, is not supported by the evidence. Possible explanations will be discussed later.

Case 2 and 3: Combinations (low, low) and (high, high)

For these two cases average destruction rates are displayed in figures 1.3 and 1.4. In the first case, when both subjects have an endowment of €5, it seems there are some differences across the fair and the unfair treatment. But due to low destruction rates on general these differences are not significant (Wilcoxon rank sum test, $p = 0.36$). The other pairwise comparisons of treatments as well as the case, when both subjects are endowed with €10 also show no significant differences.

Figure 1.3: Average percentage destruction if target has €5

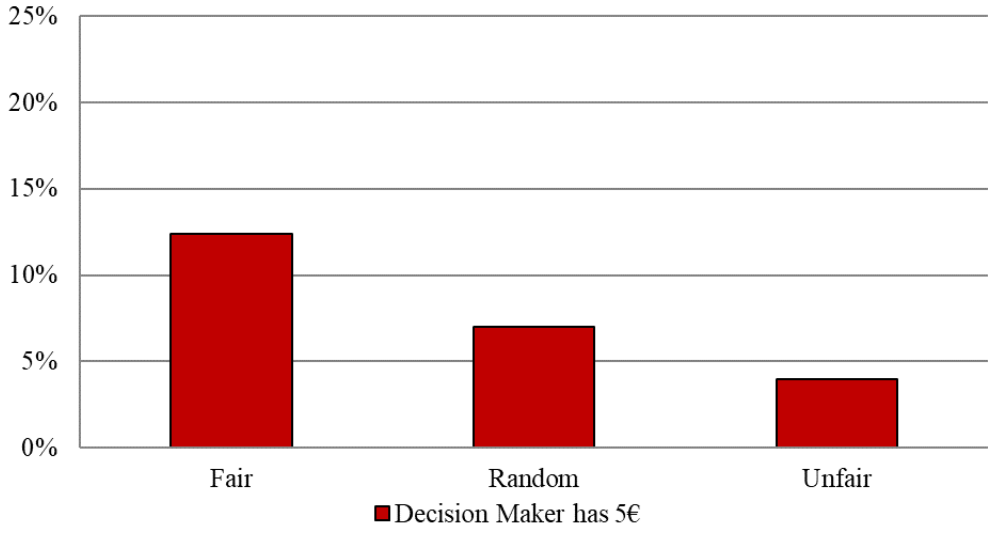
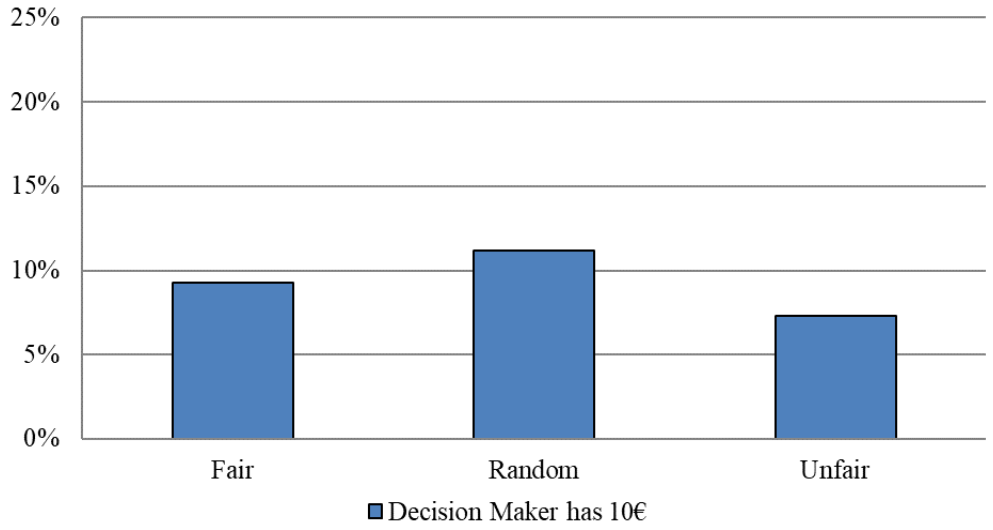


Figure 1.4: Average percentage destruction if target has €10



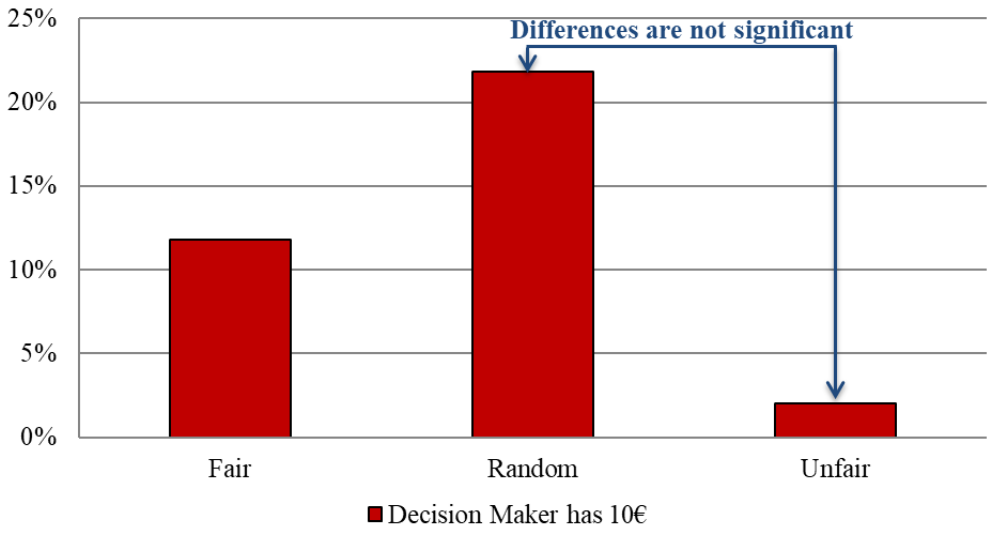
Therefore, hypothesis 4 can mostly be confirmed.

Result 4: *There are no significant differences across treatments concerning destruction rates for the combinations (low, low) and (high, high).*

Case 4: Combination (high, low)

In the last case, most models and theories predict low destruction rates overall, as the decision maker is in a privileged position anyway. Results are displayed in figure 1.5.

Figure 1.5: Average destruction if target has €5



In contrast to those predictions, differences across treatments are noticeable in this case, especially between the random and the unfair treatment. They are close to but still not significant (Wilcoxon rank sum test, $p = 0.14$).[10]

Result 5: *There are no significant differences across treatments concerning destruction rates for the combination (high, low).*

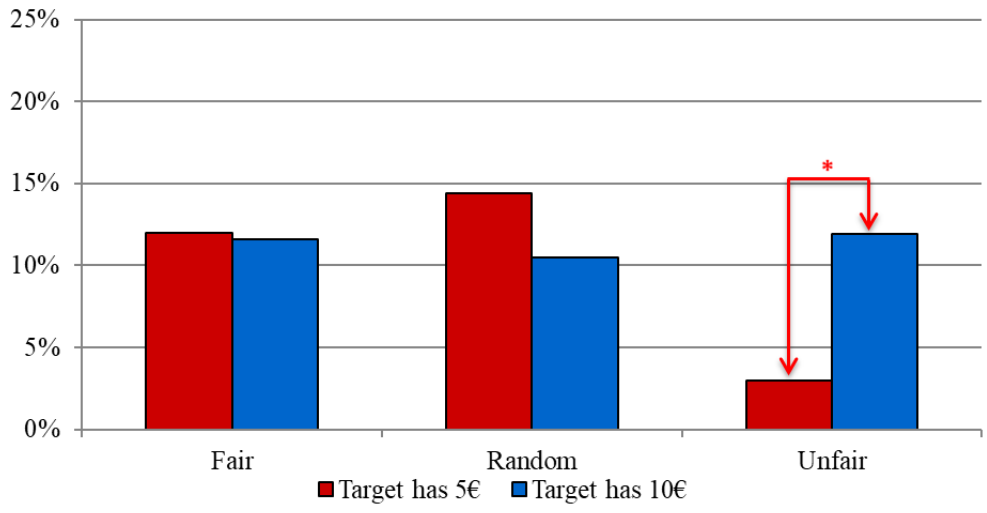
Looking at figure 1.5 this may seem surprising, but as mentioned before, the reason for most results being not significant are low destruction frequencies overall. To illustrate that point, we take a closer look at the data on an individual level for the last combination (high, low). In the random treatment in 5 out of 17 decisions, subjects decide to reduce somebody's payoff. Additionally, every time they chose a rather high amount and in 2 cases even the entire endowment of the group member is destroyed. In contrast, in the

[10] As we will see in the regression in section 1.4.5, the interaction term of the unfair treatment and the income of the decision maker is also close to being significant. So it seems there is the tendency that in the unfair treatment subjects with €10 destroy less money from those having €5 compared to the other treatments.

unfair treatment money is burned only in 3 out of 22 decisions and only to a very little extent in each case. That leads to high percentage differences on average, driven only by a few single decisions in the random treatment and explains why the difference between the random and the unfair treatment is not significant. Two of those subjects burning money in the random treatment state in the questionnaire as reason for their decision: They expected "the other participant would have done the same, if he or she had been given the opportunity to do so." Obviously the motive of preemptive retaliation still seems to be a relevant factor, even though the design allows only unilateral destruction decisions.

The only significant effect found is *within* the unfair treatment for pooled decisions, while examining which endowment class is more prone to be target of destruction. Pooling in this case means, we consider both the decisions made by subjects having an endowment of €5 or €10. The endowment of the target is held constant in this case. As one can see in figure 1.6, the percentage of endowment destroyed of those in the unfair treatment, who received an endowment of €10 is much higher than the amount reduced from subjects having €5.

Figure 1.6: Average destruction if target has €5



* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

These differences are clearly significant (Wilcoxon rank sum Test, $p = 0.016$; two tailed) and would even be larger, if one compared absolute values instead of percentage ones. As figure 1.6 suggests, this effect is mainly driven by the fact that subjects with an endowment of €5 (target has €5) is less money subtracted in the unfair treatment compared to the other treatments. In the other treatments, differences are insignificant.

Result 6: *Within the unfair treatment the percentage destruction targeting subjects having a high endowment is significantly higher than the percentage targeting those with a low endowment.*

Besides counterweighting effects of different income classes, another reason for the similar burning rates across treatments could be that participants burning decisions are not strongly affected by their fairness evaluations. In the next section we analyze the assessments of those subjects burning any positive amount of money and those who decide not to do so.

1.4.4 Correlation of Fairness Evaluations and Destruction Decisions

As shown before, subjects rate the unfair treatment as clearly less fair than the other two treatments. Nevertheless, their judgments do not seem to (strongly) affect their destruction decisions. Looking at the correlation of one subject's burning decision (yes=1 or no=0) and her fairness evaluation of the treatment, there is almost no correlation at all (Spearman's rho=0.11; Test of independence $p = 0.23$). The figures are similar, if you control for specific income combinations. Therefore it seems, those subjects who make use of the opportunity to destroy another player's endowment, do not mainly act in this way, because they perceive the endowment determining mechanism as unfair.

Result 7: *Negative fairness evaluations do not trigger subjects' decisions to reduce another player's payoff in this context.*

1.4.5 Regression Analysis

In this section we report the results of the regression analysis. This allows us to perform the analysis with demographics and other controls. Results are similar as before. Analogously to section 1.4.2, we first look at pooled values of destruction. We use the percentage of destruction as the dependent variable. An alternative specification would be a regression with the destruction decision as binary outcome. But as discussed, this analysis additionally incorporates the magnitude of destruction and provides us with more detailed results.

Results of the pooled regression are displayed in table 1.4. None of the treatments have a significant effect on destruction rates. Male subjects and economists show more antisocial behavior, while people, who donate to charity destroy significantly less money.

Table 1.4: Regression: Pooled destruction (in percent)

	Treatments only	Treatments with controls
T_fair	-0.028	-2.465
	(4.864)	(4.827)
T_unfair	-2.815	-6.329
	(4.788)	(4.973)
Male		6.170
		(4.097)
Economics		7.747^{+}
		(3.935)
Charity		-8.134*
		(4.050)
Constant	11.784**	12.818*
	(3.596)	(5.296)
Observations	119	119
R^2	0.004	0.083

Standard errors in parentheses.
Destruction is measured by adding up both intended values and dividing them by sum of endowments of the targets (=15).
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.5: Regression: Destruction behavior of specific income classes (in percent)

	Target has 5	Target has 10	Target has 5	Target has 10
T_fair=1	5.322	4.045	3.416	1.278
	(7.659)	(7.149)	(7.520)	(7.243)
Income_10=1	14.706^{+}	1.412	14.153^{+}	0.989
	(8.052)	(7.516)	(7.724)	(7.439)
T_fair=1 X Income_10=1	-15.287	-5.971	-16.701	-5.122
	(10.892)	(10.166)	(10.547)	(10.158)
T_unfair=1	-2.968	5.917	-6.732	2.107
	(7.581)	(7.076)	(7.452)	(7.178)
T_unfair=1 X Income_10=1	-16.797	-8.912	-15.451	-8.622
	(10.721)	(10.007)	(10.258)	(9.880)
Male			3.792	7.078
			(4.597)	(4.428)
Economics			9.554*	6.375
			(4.387)	(4.225)
Charity			-12.360**	-6.426
			(4.499)	(4.333)
Constant	7.059	9.765^{+}	11.626	9.975
	(5.694)	(5.314)	(7.224)	(6.958)
Observations	119	119	119	119
R^2	0.072	0.016	0.177	0.072

Standard errors in parentheses.
In this regression the income of the target is held constant and the income of the decision maker is used as dummy variable. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In table 1.5 we report the results by specific income combinations, both for the decision maker and the target. In each regression the income of the target is held constant. Both models are estimated with and without controls. The income of the decision maker is used as an explanatory dummy variable. We also look at interaction effects between the treatment variable and the income of the decision maker. The treatments have no significant effect on the destruction decision. When the target and the decision maker have an income of €5, a higher percentage of the target's income is deducted than in the case when the decision maker has an income of €10. This effect is weakly significant, but only in the random treatment. In the fair and unfair treatment this relationship is

exactly reversed. There, less money is subtracted from subjects with €5 by those with an income of €10. As before, the interaction effect is not statistically significant. This might be the case, because of rather low destruction rates in general.

1.5 Discussion and Conclusion

In this paper we analyze the extent of antisocial behavior in a money burning experiment, in which endowment is determined in different ways, varying in their degree of fairness. This is the first study, which focuses on the fairness of the money allocating process, while in most other comparable experiments, money is provided in the form of windfall gains. Altogether, we find moderate degrees of antisocial behavior in all treatments. However, there are no significant differences in destruction rates across the treatments, even if one controls for specific income classes. One reason might be that individual burning decisions in this context do not to depend much on the perceived fairness of the treatment. This is shown by the results of the questionnaire. There is no correlation between the individual fairness assessment of the procedure and the decision to burn money. Below we discuss several possible explanations for this result:

First, it could be that the fraction of people showing antisocial behavior is rather constant and independent of the context: Some evidence for this hypothesis comes from Sadrieh and Schröder (2012) and Zhang and Ortmann (2013). They both suggest, there is some relatively constant fraction of people exhibiting both pro- and antisocial behavior, depending on the exact choice set they are offered (e.g. Joy of Destruction Game or Dictator Game). Sadrieh and Schröder assume these so called "influencers" derive their utility mainly from having the power to change the payoff of others, no matter in which direction. This argument goes in a similar direction as experimenter demand or boredom effects. By these effects participants also gain utility from the process in which payoffs are generated and not only from the final distributional consequences (e.g. They want to be "active" or have "fun", while taking part in the experiment). This is especially true for low stakes scenarios, as is most often the case in standard economic experiments. But there is also abundant evidence against this argument: For example in Zizzo and Oswald

(2001) or in D. Fehr (2018) the amount of destruction varies a lot as treatment parameters change and has very low values in some cases.

An alternative explanation is uncovered by looking into responses to our questionnaire: Several participants state as their reason for not reducing somebody else's payoff, that they do not like the endowment allocation mechanism, but see it "not as the fault of the other subject". Two aspects are important here: First, strictly speaking, the procedure of endowment determination is ex ante also random across all treatments: Subjects are randomly assigned to one of these treatments, as well as to the early or late group. Presumably, some participants perceive the whole mechanism as a sort of (unfair) "lottery", in which some participants are lucky (the ones who do not have to do any work and receive the high endowment) and others are not. Adopting this view, one could argue that the criterion of process fairness is satisfied here, as all participants had ex ante the same chances being in each of the possible positions.

And secondly, subjects in our experiment have no opportunity to balance initial differences. Possibly, because of this they are not blamed or held responsible for these differences and therefore are not target of destruction more often. This would explain why destruction rates are at a low level in all treatments.

For both explanations we find clear evidence in other studies: In Bolton et al. (2005) "unfair" (in the sense of strongly unequal) outcomes in an Ultimatum Game are widely accepted, if they are determined by a fair (random) procedure, in which the players both had the same chance to receive the favorable outcome. Furthermore, they find that unfair offers are only frequently rejected, if the proposer had the chance to choose a more equal allocation, but not if he had no other choice. In a similar fashion, Blount (1995) finds that much lower offers than common are accepted in the Ultimatum Game, if these offers are randomly generated by a computer instead by the participant himself. E. Fehr and Fischbacher (2004) also provide additional evidence for the second point: They examine second and third order punishment in distributional choices. In their experiment subjects play a Dictator Game. In one treatment the recipient himself can, after seeing the decision, costly punish the dictator. In another treatment a neutral third party has the right to do

so. Punishment for low transfers is widespread and used both by the recipients and also neutral third parties. So, there is clear evidence, that if people violate common sharing norms, they get sanctioned. The main difference to our experiment is that people have no option to change initial endowments. Thus we can presume that destruction rates would have been much higher - especially in the unfair treatment - if subjects were offered an option to redistribute incomes.

Therefore we can conclude that antisocial behavior in our experiment is probably so low, even in the unfair treatment, because subjects cannot be held responsible for the resulting (unfair) distribution.

The latter explanation could be tested by performing a similar experiment and adding an additional stage, in which subjects with the high endowment can transfer some share to the ones with the low endowment before the destruction decision is made. Or instead the design could be changed such that the high endowment is not directly allocated, but subjects are assigned the right to choose one of these two values (that means in the unfair treatment participants from the late group would have the right to choose which endowment they would like to receive). Then we would expect a much stronger treatment effect on destruction rates targeting those subjects who did not redistribute.

Appendix

A.1 Instructions (in German)

Experiment Anleitung

Platz-Nr._____

Vielen Dank für Ihre Teilnahme und herzlich willkommen zu diesem Experiment!

Bitte sprechen Sie ab jetzt nicht mehr miteinander und schalten Sie ihre Mobiltelefone aus. Wenn Sie eine Frage haben, heben Sie Ihre Hand, wir kommen dann zu Ihnen an den Platz und beantworten Ihnen diese soweit möglich.

Bei diesem Experiment geht es darum, Entscheidungen zu treffen. Sie haben die Möglichkeit, durch diese Entscheidungen Geld zu verdienen, welches Ihnen am Ende des Experiments anonym und in bar ausbezahlt wird. Der genaue Geldbetrag hängt ab von den von Ihnen getroffenen Entscheidungen, sowie den Entscheidungen Ihrer Mitspieler.

Das Experiment hat zwei unterschiedliche Anfangszeiten für je die Hälfte der Teilnehmer. Wer zu welchem Zeitpunkt startet, wurde im Vorfeld per Zufall bestimmt. **Diejenigen Teilnehmer, die zum „offiziellen" Termin eingeladen wurden, bekommen für etwa 25-30min eine Aufgabe, die sie erledigen müssen.**

Dabei handelt es sich um die Korrektur von ausgefüllten Bögen eines IQ-Tests für ältere Personen, die aus einem kürzlich durchgeführten Experiment der Uni Heidelberg stammen und noch nicht ausgewertet wurden. Die Teilnehmer bekommen eine Lösungsschablone und haben die Aufgabe, zu ermitteln wie viele Punkte bei jeder Aufgabe erzielt wurden.

Die Teilnehmer, die zum späteren Zeitpunkt beginnen, müssen keine solche Aufgabe erledigen und starten gleichzeitig mit der ersten Hälfte der Personen zum zweiten Teil des Experiments, der wiederum für alle identisch ist.

Die Teilnehmer, die zuvor die Aufgabe erledigt haben erhalten ein Einkommen von 10€ [Treatment unfair: ***5€***]**, diejenigen, die später begonnen haben bekommen ein Einkommen von 5€ [**Treatment unfair: ***10€***]**.**

[Treatment Zufall: ***Zuerst wird die Höhe Ihres Anfangseinkommens ermittelt. Dieses wird zufällig bestimmt und nimmt entweder den Wert 5€ oder 10€ an.***

Dazu sind Briefumschläge mit je einem dieser zwei Geldbeträge gefüllt und jedem Gruppenmitglied wird durch die Experimentleitung zufällig einer dieser Umschläge zugeteilt. Die jeweilige Anzahl der beiden Beträge ist so gewählt, dass genau die Hälfte der Teilnehmer zu Beginn 5€ bekommen und die anderen 10€.]

In der zweiten Stufe haben Sie dann unter Umständen die Möglichkeit, das Einkommen von einem Ihrer Mitspieler zu reduzieren. Die genauen Details werden Ihnen nach der Ermittlung des Starteinkommens mitgeteilt.

All diese Informationen sind jedem bekannt, diese Anleitung ist für alle Teilnehmer identisch…

Abschließend folgt noch ein Fragebogen und danach wird Ihr Geld ausbezahlt.

Ihr Einkommen beträgt 10€ [5€] **Platz-Nr._____**

Wie bereits angekündigt, kann sich nun das Einkommen einzelner Spieler noch verändern. Dazu bilden Sie mit einem zufällig ausgewählten Teilnehmer eine 2er Gruppe. Sie können dabei entweder einem Spieler zugeordnet sein, der den gleichen Geldbetrag wie Sie erhalten hat, oder einem, der den anderen Betrag bekommen hat. Während des gesamten Experiments herrscht vollständige Anonymität über die Gruppenzusammensetzung.

Sie haben nun die **Möglichkeit, das Einkommen des anderen Spielers** um einen gewissen Betrag **zu reduzieren**. Die Auszahlung des betreffenden Spielers verringert sich dann um den angegebenen Wert. (Hinweis: Sie bekommen von diesem Geld nichts selbst hinzu)

Diese Möglichkeit ist allerdings mit eigenen Kosten in Höhe von 10% des gewählten Betrags verbunden. Konkret heißt das, für jeden Euro, den Sie dem anderen Spieler abziehen möchten, müssen Sie selbst 0,10€ = 10 Cent bezahlen. Die Höhe der Abzüge ist frei wählbar, die maximale Menge ist aber auf den Wert des Starteinkommens des anderen (entweder 5€ oder 10€) beschränkt. Sie müssen von dieser Möglichkeit keinen Gebrauch machen.

Nachdem jeder seine Wahl getroffen hat, wird pro Gruppe einer der beiden Teilnehmer ausgelost und dessen Entscheidung wird entsprechend durchgeführt. Das bedeutet, dass entweder nur Ihre Entscheidung oder nur die Ihres Spielpartners wirksam wird. Natürlich fallen bei Ihnen die Kosten auch nur dann an, wenn Ihre Entscheidung ausgewählt wurde. (Aus Gründen der Berechenbarkeit sollte der gewählte Betrag ein Vielfaches von 10 Cent betragen). Sie und ihr Spielpartner treffen Ihre Wahl gleichzeitig und völlig unabhängig voneinander.

Tragen Sie jetzt Ihre Entscheidung in den Bogen auf der nächsten Seite ein. Danach werden wie angekündigt zufällig die Gruppeneinteilung, sowie die wirksamen Entscheidungen ausgelost. Dazu zieht jeder Spieler ein Kärtchen mit einem Buchstaben darauf, der die

Gruppenzuordnung bestimmt, sowie welche Entscheidung durchgeführt wird. Tragen Sie bitte diesen in das dafür vorgesehene Kästchen auf dem Entscheidungsbogen ein. Stecken Sie den Bogen anschließend in den Briefumschlag, den Sie zuvor bekommen haben. Diese werden eingesammelt und ausgewertet. Daraus ergeben sich dann die endgültigen Auszahlungen für alle Teilnehmer.

Entscheidungsbogen

Platz-Nr._____

Ihr Einkommen beträgt 10€ [5€].

Betrag, den Sie dem anderen Spieler abziehen möchten, falls dieser **ein Einkommen von 5€** hat:

Betrag, den Sie dem anderen Spieler abziehen möchten, falls dieser **ein Einkommen von 10€** hat:

Gezogener Buchstabe für die Gruppeneinteilung:

Chapter 2

Gender Differences in Bargaining

Experimental evidence[†]

Abstract We study gender differences and gender pairing effects in a laboratory experiment with alternating-offers bargaining for a fixed pie, framed as an employer-employee interaction. We vary the degree of asymmetry in bargaining power between roles, as well as the disclosure of genders of the negotiating partners. This allows us to disentangle differences based on gender identities and differences due to inherent characteristics, such as for example risk preferences. With low asymmetry, we find no gender effects in behavior or outcomes. When there is high asymmetry and gender is known, men achieve more favorable deals than women in both roles, especially in mixed gender pairings. These differences are significant, despite gender information being transmitted only in a very subtle way. However, differences disappear, when no gender information about the bargaining partner is provided. Additionally, we examine the bargaining strategies of the players in detail. In presence of high asymmetry, men behave more aggressively in mixed gender parings, while the opposite effect is true for women. This explains higher earnings for men conditional on achieving a deal. Our results provide a potential explanation for the remaining part of the gender wage gap. When we include the cases in which negotiations fail, the picture is not as clear cut: Men and women achieve similar outcomes on average, as women are more likely to reach an agreement.

[†]Joint work with Graciela Kuechle

2.1 Introduction and Literature Review

The gender wage gap has been the focus of research in economics for several years. In the last decades the gap has narrowed considerably in many countries, mostly because women have acquired higher levels of human capital and spend more years being active in the labor market (see Blau & Kahn, 2017 for a survey). Nevertheless, the unadjusted gap still has a value of about 20-25% in Western European countries, depending on how it is measured (Blau & Kahn, 2017; Boll et al., 2017). The unadjusted gap can be split into an explained and an unexplained part: The explained part consists of factors, which are directly observable and measurable, such as e.g. education, job experience, work force interruptions, occupations and industries. The overall reduction of the gap can mostly be attributed to a reduction of the differences belonging to the explained part. Still, there remains a small but persistent part of the gap of about 6-12%, which the traditional factors cannot account for (Boll et al., 2017).

Among others, a prominent explanation for the remaining part are gender differences in negotiating or bargaining behavior and in the resulting outcomes (Blau & Kahn, 2017). These differences can stem either from inherent characteristics, such as preferences or from more external forces based on gender identity, as for example social norms and stereotypes (Marianne, 2011). Concerning inherent characteristics, there is a lot of evidence in the literature: Women are found to be more risk averse than men (Eckel & Grossman, 2008b), women are more likely to avoid competition (Niederle & Vesterlund, 2007), they are more concerned with the preferences of other people (Eckel & Grossman, 1996; Selten & Ockenfels, 1998) and behave more cooperatively in Dictator and Ultimatum Games (Eckel & Grossman, 2008a).[10] These traits are likely to lead to lower expected payoff in negotiations. For differences due to norms and stereotypes, evidence is provided for example by women receiving worse wages offers than men (Säve-Söderbergh, 2007) or being less rewarded for their work, even when delivering the same performance (Heinz et al., 2016).

[10]For an additional overview of gender differences in preferences see also Croson and Gneezy (2009) and for an overview about gender differences in labor markets using evidence from lab and field experiments see Azmat and Petrongolo (2014).

As preferences and norms are not directly observable, the literature has shifted the focus to examining differences in negotiation behavior and outcomes. Since initial wages, pay raises and promotions are often subject to individual bargaining, negotiation behavior can possibly explain a substantial part of the remaining gap. In this regard, women have been found to lag behind men in several ways: Women are less likely to start negotiations both in the field (Babcock & Laschever, 2009) and in lab experiments (Bowles et al., 2007; Exley et al., 2016), especially, if wages are not explicitly described as negotiable (Leibbrandt & List, 2014). Furthermore, women, who start negotiations, are more likely to be penalized for that decision by male evaluators (Bowles et al., 2007). This expectation may explain why, according to some studies, women ask for less and are offered less than men in bargaining settings (Azmat & Petrongolo, 2014; Säve-Söderbergh, 2007).

Some of the previous findings point at the constraining effects of gender roles in the context of negotiation practices, as predicted by social role theories. According to congruity theory (Eagly & Karau, 2002), individuals behave in ways that are consistent with culturally accepted roles. In Western cultures, women are expected to behave more accommodating and less aggressive than men, traits that are detrimental to successful bargaining (Amanatullah & Tinsley, 2013; Stuhlmacher & Linnabery, 2013). In a meta-analysis of gender differences in bargaining, Mazei et al. (2015) conclude that although men obtain slightly higher earnings than women, the differences are moderated by various contextual factors shaped by socially entrenched gender roles (as seen for example in Eagly & Karau, 2002).

Insights on this topic are relevant for real-world applications and institutional designs in the context of labor markets. There might be efficiency losses, if mutual beneficial deals are not achieved due to biases in behavior resulting from gender effects. These losses can go in different directions. Possibly, some profitable contracts are not materialized, because the offered wage is below the threshold of the negotiating partner. Conversely, perhaps partners demand substantially higher wages than their reservation values and end up without an agreement. If differences in outcomes are in a large part the result of stereotypes and discrimination, one could think of making the application process more

anonymous (by for example not providing social and demographic information) or, if possible, base wages to a greater extent on objective criteria, such as individual performance, instead of bargaining. Another possibility would be to allow for more transparency, by showing how the amount of wage is determined or making earning ranges for certain positions publicly known.

Gender differences in experiments where partners fight for a share of financial resources have been investigated by means of many different designs. The existing literature suggests that gender differences are especially likely to occur, if there is both **an asymmetric situation of power** and **disclosure of gender information**. If one or both of these factors are missing, then one usually does not observe significant differences.

In symmetric situations, both Lutzker (1961) and Conrath (1972) find no gender differences in behavior in a Chicken Game independently of whether genders are known to the partners or not. Also in the studies of Hernandez-Arenaz and Iriberri (2018) and Dittrich et al. (2014), who both employ a bargaining task similar to ours, no differences are observed when genders are revealed but there is no asymmetry in bargaining power. These two experiments will be discussed later in more detail.

Studies based on asymmetric situations like the Ultimatum Games provide evidence that the average offers and acceptance rates of men and women are not significantly different, if genders are not salient (Eckel & Grossman, 2001; Exley et al., 2016 ; Solnick, 2001). An exception is Rigdon (2012). In this study, the author finds that women demand significantly less than men, but when information about other players' demands and offers is provided, gender differences vanish completely.

In asymmetric environments where gender is mutual knowledge, results from Ultimatum Games show that the gender of the proposer affects the acceptance rate, but evidence is very mixed. While in Eckel and Grossman (2001) women are more likely to accept offers from other women than from men, in Solnick (2001) this effect goes in the opposite direction. However, in a recent study of Li et al. (2018), Solnick's key findings could not be replicated.

Concerning experimental evidence from bargaining games, there are two studies based on Rubinstein (1982) model that bear similarities with ours: Dittrich et al. (2014) conduct a lab experiment, in which an employer representing a firm and a prospective employee repeatedly bargain over a wage by means of alternating offers. Each treatment involves a different minimum wage. The authors find that wages negotiated by women are lower than those negotiated by men regardless of the employer's gender. They also provide evidence of gender pairing effects, in a sense that gender differences in a given role may depend on the counterpart's gender. In a recent working paper of Hernandez-Arenaz and Iriberri (2018), the authors examine gender differences in the presence of different kinds of asymmetry. They observe significant role differences: Compared to female responders, male responders are less likely to reach an agreement, spend more time bargaining and obtain a larger share of the pie conditional on reaching an agreement. But the overall unconditional earnings of men and women are not significantly different, a result that holds across treatments and is similar to our findings. In contrast to Dittrich et al. (2014), Hernandez-Arenaz and Iriberri (2018) find no interaction effects, in a sense that the gender of the bargaining counterpart does not affect the outcomes of the different genders.

However, none of these studies examine the effect of both influencing factors on gender pairings systematically, so it is hard to assess their impact in isolation. In this respect, our paper contributes to the existing literature by varying both factors independently in a controlled environment. This allows us to examine their effect in isolation and disentangle possible explanations for differences in bargaining behavior. Furthermore, we consider it a particular important feature of our design to have a very subtle manipulation for gender revealing. This information is embedded under the heading of other general demographic information, which ensures that we do not observe differences that are induced merely by demand effects.[11]

Our design is similar to Dittrich et al. (2014), except for the way in which gender information is transmitted and in the absence of minimum wages. Instead of face to face

[11] See for example Zizzo (2010) for a survey about this topic.

interactions, we provide subjects with demographic information of their counterparts that includes gender and four other items, which are expected to be non-informative[12] (age, place of residence, occupation, and semester). Our design also bears similarities with Hernandez-Arenaz and Iriberri (2018), except for the implementation of different forms of asymmetry and the disclosure of gender information. In their computer aided experiment, subjects see on the monitor an avatar, representing their own gender and that of their bargaining partner. This makes gender very salient in their context and could potentially create experimenter-demand effects, as it is the only sort of demographic information, which is provided. A similar argument can be made about the study of Dittrich et al. (2014), but in face-to-face interactions gender does not necessarily seem to be an explicitly pronounced factor.

Our results suggest that gender and gender paring effects play an important role in the bargaining process. We analyze behavior always role-dependently, comparing outcomes and strategies of female vs. male employers and employees respectively. Overall, we find no differences in behavior and outcomes in the environment with low asymmetry. We conjecture that this environment is very close to a symmetric situation, where the 50:50 sharing norm is very salient. Under high asymmetry and if a deal is reached, men are able to achieve more favorable outcomes than women in both roles. This effect is strongest in the pairing of a male employer with a female employee, and if genders are known. Under high asymmetry, if genders are not revealed, men achieve only slightly better outcomes than women in both roles. The reason behind these differences is that men bargain more aggressively in mixed gender pairings when genders are revealed, while the opposite is true for women. If genders are not known, this tendency still remains, but is much less pronounced. However, if one additionally accounts for the cases, in which the game ends before a deal has been reached, results are mixed: Men and women perform equally well on average, as women are more likely to reach a deal, which offsets the differences from

[12]We asked participants at the end of the experiment, if they took any of the demographic information into account and if yes, which of those exactly. The item mentioned most frequently was “gender”, followed by “age”. The other items were not mentioned very often. As age cohorts differed not much across subjects, one might assume this factor did not play a big role. Further evidence for this claim is provided in the regression analysis of the results section.

beforehand. Overall, these results provide further evidence that gender differences mainly arise, when there is both an asymmetric environment and knowledge about the genders of the interacting partners. So it seems a certain degree of asymmetry is a necessary, but not a sufficient condition for gender differences to occur. Probably, in symmetric situations the fair split is very salient, because of common sharing norms. But given there is some room left for bargaining, our results suggest that gender differences in behavior depend both on gender identities and on differences in preferences.

An additional factor, which also might be relevant, is the domain of the negotiation. Bear and Babcock (2012) show that gender differences disappear, if the very same strategic situation is framed in a "female context" instead of a "male context". This is also pointed out in the meta-analysis of Mazei et al. (2015), in which the authors conclude that "multiple influences may affect role congruity for women in negotiations, so that gender differences in economic outcomes should depend on the specific context".[13] Due to practical limitations such as sample sizes, we do not examine this factor further in our experiment and choose a setting, which is comparable to most preexisting studies.

2.2 Experimental Design

2.2.1 Bargaining Setup

Subjects are matched into groups-of-two and bargain over the division of a pie of 100 experimental units by making alternating offers . Simultaneously to the proposer making her decision, the responder is asked to state her minimum share, for which she would just accept the offer. In this respect our design is slightly different to the classical Rubinstein bargaining game. But this way allows us to elicit more detailed information about the strategies of the players, compared to just observing, if an offer is accepted or rejected. To be in line with the strategic framework of the original setup, subjects learn about the proposal of the previous round, but the minimum is always private information of the

[13] In most studies as well as in most real world situations, negotiations take place in financial or business domains. It might well be possible, that women behave much more aggressively, if for example they bargain about benefits for their children than if they sell their car.

players. If the share offered to the responder is higher than her stated minimum, the proposal is automatically implemented and the game ends. As the actual distribution of shares is not influenced by the reported minimum, this guarantees that there are no strategic considerations involved and that it is weakly dominant to truthfully state one's reservation value. If no agreement is reached, the other subject makes a counteroffer in the next round. Bargaining continues in this way, until either the parties achieve a deal or the game ends automatically. From round 3 onwards there is a probability of 20% that the negotiation will break down, if partners have failed to reach an agreement. If this happens, a given outside option is implemented. The exact values of the outside option vary between treatments and reflect the degree of bargaining power of the partners. In the beginning, subjects are randomly assigned to different roles, named "employer" and "employee". The employer is always in a more privileged position, as his outside option is higher than the one of the employee. Furthermore the employer makes the proposal in the first and all odd-numbered rounds, which gives him an additional strategic advantage. in this setting. Our aim was to model an asymmetric environment between negotiation parties. Firstly, this specification reflects real world (wage) negotiations in a more realistic way and secondly, as evidence suggests, asymmetry is a necessary condition for gender differences to occur since there is no clear sharing norm of the pie.

In contrast to similar experiments, the game is played only once, in order to avoid learning effects or the outcomes of past periods influencing future behavior . Additionally, income effects or risk preferences might play a role in repeated interactions, depending on the exact payoff scheme.[14] For these reasons the observations of a participant, playing the same bargaining game several times, cannot be seen as independent. A drawback of our design is that it yields fewer observations. This reduces the power of the statistical analyses to some extent, especially if we analyze effects of specific gender pairings.

The whole scenario is framed as a business context. Throughout the description of the situation, we use terms, which are linked to an employer-employee interaction, in which partners negotiate about the distribution of the surplus of a (potential) collaboration.

[14] If for example earnings of different rounds are added up, subjects might become more or less aggressive in bargaining, depending on whether their earnings are currently above or below their expectations.

We use this kind of language, as this might make potential effects more pronounced. Studies have suggested that gender differences in bargaining behavior might be domain-dependent. Our design allows us to additionally implement a neutral framing[15] in order to test if differences are reduced.

2.2.2 Treatments

The experiment consists of three treatments:

1. **Treatment Info & low asymmetry ("Info_low")**
2. **Treatment Info & high asymmetry ("Info_high")**
3. **Treatment No info & high asymmetry ("No_high")**

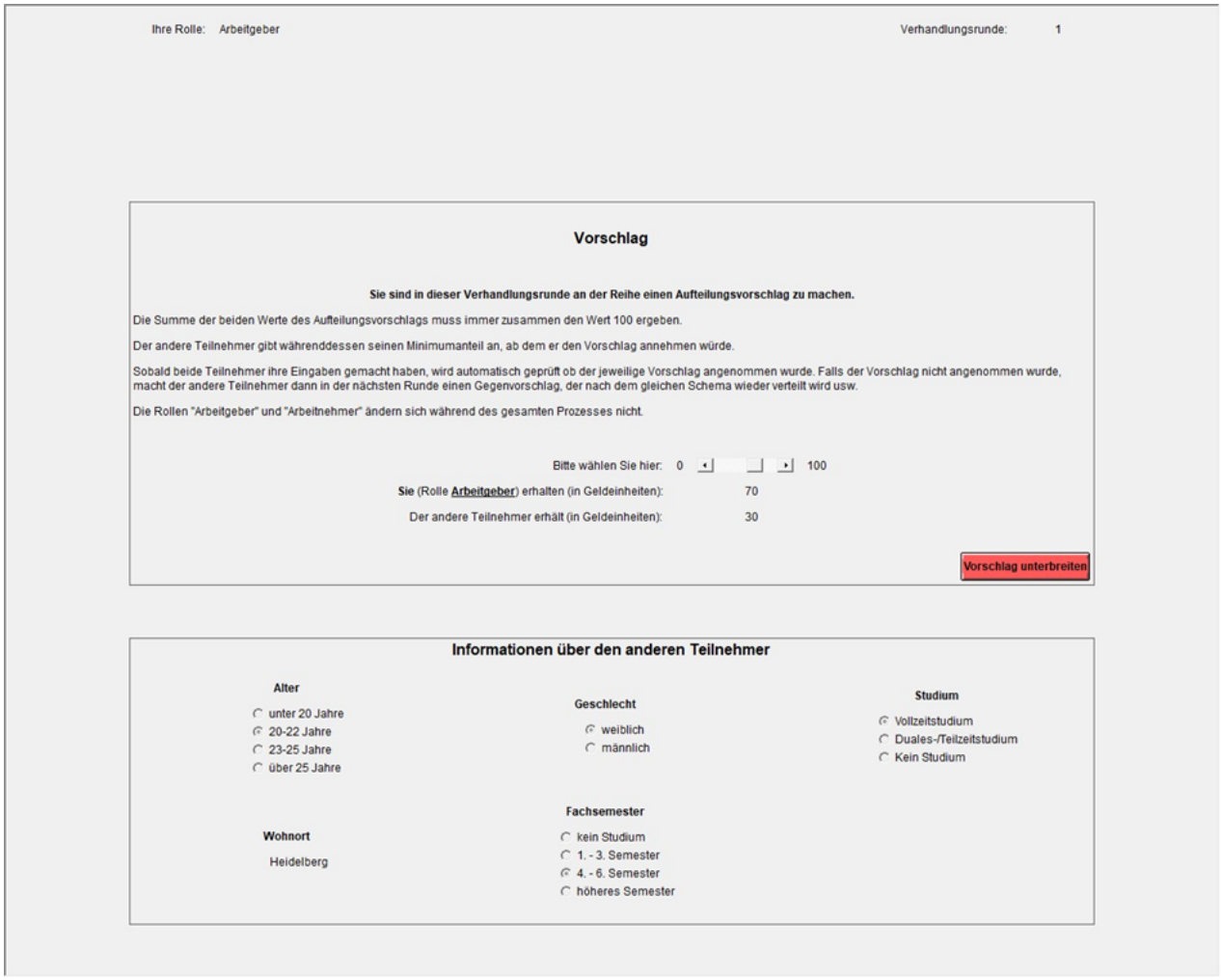
All sessions of all treatments follow the same timeline:

1. Demographic questionnaire
2. Instructions for the bargaining task
3. Comprehension quiz (incentivized)
4. Eliciting of expectations
5. Bargaining task (main part)
6. Debriefing questionnaire

In all treatments subjects respond to a demographic questionnaire (age, gender, place of residence, if they study and if yes, which semester they are enrolled in). In the info treatments, this information is revealed to the other player for the whole duration of the negotiation (see figure 2.1).

[15] One could for example use terms as "Player A" and "Player B" for the roles and avoid calling the game a "negotiation".

Figure 2.1: Screen gender revealing in Info treatments



We chose on purpose a very subtle way of revealing gender in order to avoid potential experimenter-demand effects, which could arise when providing gender as the only piece of information. Subjects are informed about this procedure in the instructions for the bargaining task. In the no info condition, the procedure is exactly the same, except for the revealing of demographic information. Across all treatments, when the game ends without reaching an agreement, an asymmetric outside option is implemented. In the treatment with low asymmetry, the outside option has values of (20, 0), while in the high asymmetry treatment it is (40, 0). The first number corresponds to the payoff of the employer and the second to the payoff of the employee respectively. The design allows us to vary these parameters easily, so one could in future treatments implement a completely

symmetric situation or an even more asymmetric environment. Before subjects start the bargaining task, they have to answer some incentivized control questions with immediate feedback. This was meant to increase the likelihood that subjects understand the rules of the game.[16] We also ask to give their expectations about the outcome of the negotiation. After the task, subjects complete a debriefing questionnaire, in which, they are asked to describe their bargaining strategy and offer a self-report on their risk preferences.

2.2.3 Implementation

The experiment was conducted between 08/2017-11/2018 in the experimental lab of the University of Heidelberg. The bargaining process was implemented using z-Tree experimental software (Fischbacher, 2007). A total of 471 subjects, mostly students, took part.[17] The average duration of each session was about 35-40min and average earnings were 8.52€per subject, with payoffs ranging from 4.00€-14.00€. Further details are summarized in table 2.1 below:

Table 2.1: Sessions overview

Treatment	Number of Sessions	Number of Subjects
Info_low	12	160
Info_high	12	159
No_high	10	152

The original instructions for each treatment can be found in the Appendix.

2.2.4 Game Theoretic Predictions

In order to be able to assess and compare the degree of asymmetry in bargaining power across treatments, the payoffs in the resulting sub-game-perfect equilibrium of these games are derived. Our setup is similar to the classical Rubinstein bargaining model with in-

[16]The indented effect worked well, as more than 90% of subjects managed to answer at least 6 out of 7 questions correctly. Additionally, all subjects received detailed feedback about the correct answers before the experiment continued.

[17]One observation had to be removed, as one participant suddenly became sick and abandoned the experiment.

finite time horizon with players having an equal constant discount factor δ. Under the assumption of risk neutrality, the equilibrium prediction with a shrinking pie according to a discount factor of δ or a constant continuation probability of $p = \delta$ is identical.[18] As the game can end automatically only after finishing the third round, the first two rounds are irrelevant for the equilibrium prediction. The reason is that in these two initial rounds there is no discounting of the payoffs and hence neither player would accept any share below her equilibrium share she would receive in round 3. The bargaining game from round 3 onwards can be transformed into a game, in which players receive their outside options.[19] as a fixed payoff and bargain about the remaining pie.[20] Applied to our context, the game can be transformed to the standard case, in which players bargain over the remaining pie of $(100 - x)$ units and player 1 receives a fixed payoff of x units in addition. The outcome of the subgame perfect equilibrium of the remaining pie and the equilibrium proposal in round 1 is $(\frac{1}{(1-\delta}, \frac{\delta}{(1-\delta)}) * Piesize$. For the calculation of the final payoffs, simply the values of the outside option are added. With the parameters implemented in the experiment this corresponds to the following equilibrium payoffs and equilibrium proposals of the employer (all treatments have the same continuation probability/discount factor of $\delta = 0.8$):

Table 2.2: Outside options and equilibrium payoffs

Treatment	**Outside Option**	**Equilibrium Payoffs**
Info_low	$(20, 0)$	$(64\frac{4}{9}, 35\frac{5}{9})$
Info_high	$(40, 0)$	$(73\frac{1}{3}, 26\frac{2}{3})$
No_high	$(40, 0)$	$(73\frac{1}{3}, 26\frac{2}{3})$

[18] If players are risk-averse, then the equilibrium predictions might be slightly different, as a risk-averse player would prefer a sure payoff of 0.8 times the original pie compared to a lottery, in which he receives the whole pie with probably 0.8 and a payoff of zero with probability of 0.2.

[19] This transformation is only correct, if the values of the outside options are not discounted, which is true in our scenario.

[20] For further details and a proof for this claim see the working paper of Miller, Montero, and Vanberg (2015) Their model includes our case as a special case, with the only difference that the role of the proposer is randomly determined before each round, which does not influence the transformation of the game.

2.3 Results

2.3.1 Outcome Analysis

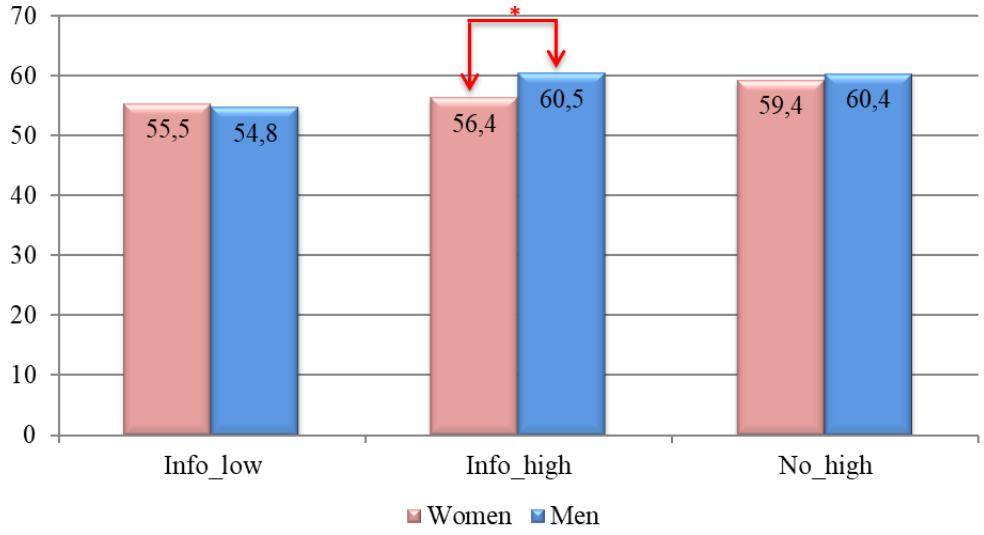
To analyze the results we will focus on the cases, in which a deal was reached. Firstly, it makes results comparable to preexisting work, where the main focus is also on gender differences. Secondly, there is no clean comparison for the other cases, because the outcomes depend additionally on random draws that terminate the game. This could confound the true effects.[21] The last and main reason for this selection is our focus on explaining potential gender differences in real-world wage negotiations. In this situation one also observes only the outcomes of the contracts which materialize.

Nevertheless, in order to be able to evaluate the performance of a negotiating strategy, it is important to also take into account the cases where negotiations failed. We report and briefly discuss these results in the appendix performing the same analysis (see figures 2.14 and 2.14).

Compared to the game theoretic predictions, subjects are never able to fully exploit their bargaining power from the employer's perspective. Maybe, social norms, inequity-aversion and fairness considerations attenuate given imbalances of bargaining power.
In figures 2.2 and 2.3 average payoffs for women and men are displayed in each treatment. All significant differences are indicated with arrows and stars.[22] We analyze the data in terms of roles, which means we only compare payoffs of the same role gender-wise (for example payoffs of female employers with outcomes of male employers). First, we look at differences within treatments and then we examine differences between treatments Info_high and No Info_high to assess, if there is a treatment effect in revealing genders. Treatment Info_low can be seen as a robustness check to confirm, that gender differences do not arise, when the asymmetry is not high.

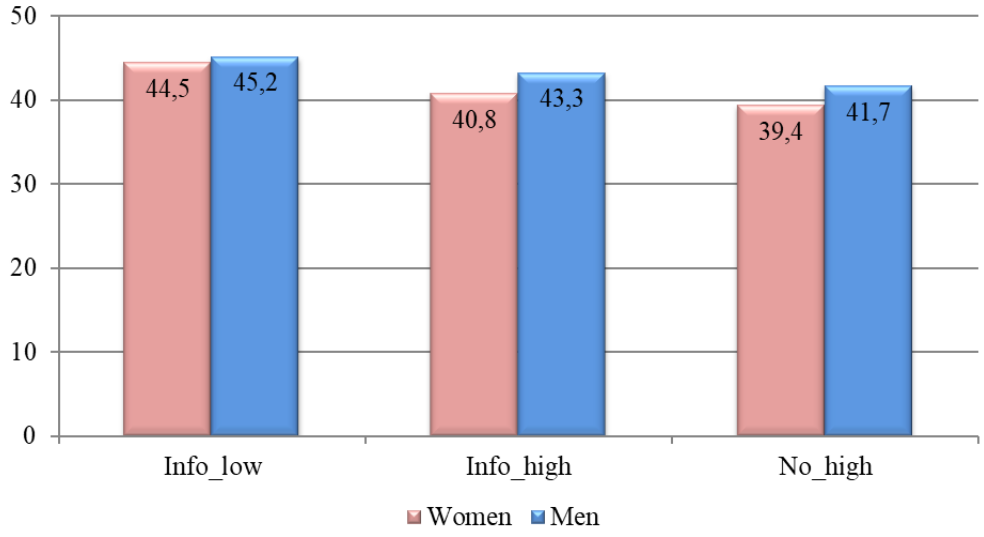
[21] If genders are not equally affected from the random draws terminating the game, payoffs may differ substantially, despite behavior being identical.
[22] * indicates significance at the 5 % level and ** at the 1% level.

Figure 2.2: Mean payoffs of employers



* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: Considering only cases, when a deal was reached.

Figure 2.3: Mean payoffs of employees



Note: Considering only cases, when a deal was reached.

A Shapiro-Wilk test shows that the data of the payoffs is clearly not normally distributed ($p = 0.000$). Therefore we use a Wilcoxon rank-sum test for the analysis. In treatment Info_low there are no significant gender differences in either role. A plausible explanation is that subjects often agreed to an approx. 50:50 split. In treatments with a more asymmetric environment, within treatments men achieve more favorable outcomes than women across all cases. These differences are only significant within treatment Info_high

for employers (rank-sum test, $p = 0.02$, see figure 2.2). When comparing average payoffs between treatment Info_high and treatment No_high (figure 2.2), women earn less as employer, but differences are not significant (rank-sum test, $p = 0.18$). Looking at the employee-side, we do not find significant differences in outcomes, neither within nor between treatments. (see figure 2.3).

The data so far does not yet provide a full picture of the results, as it does not take into account the gender of the bargaining partner. In the next section we analyze in more detail gender pairing effects to gain a more in depth understanding of the mechanisms at play.

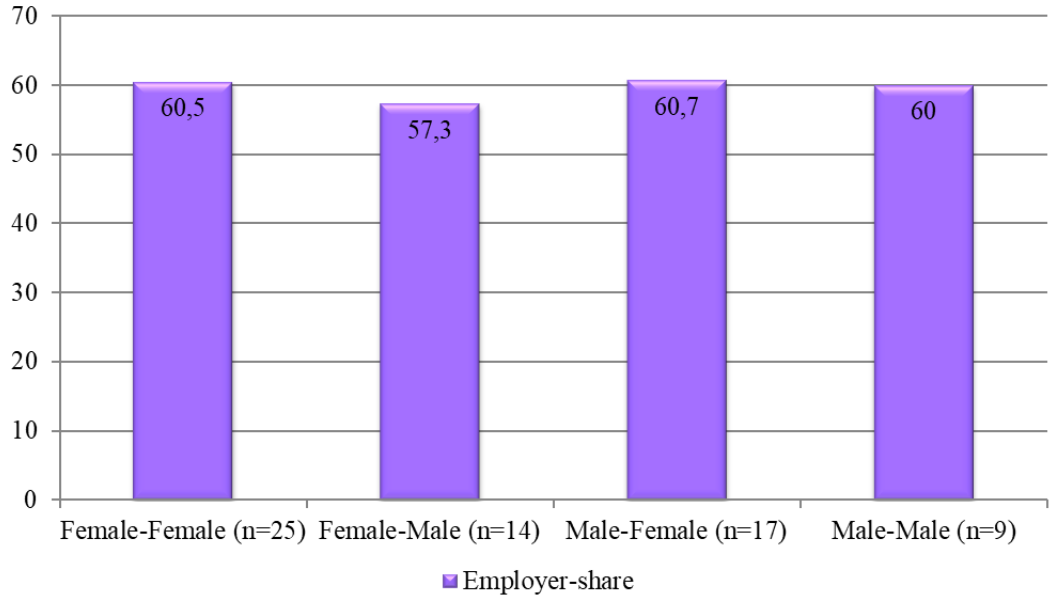
2.3.2 Gender Pairing Effects

In figures 2.4 and 2.5 average shares of employers in treatments Info_high and No info_high are displayed for each of the four gender-pairings.

Figure 2.4: Mean payoffs in treatment Info_high

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: Considering only cases, when a deal was reached.

Figure 2.5: Mean payoffs in treatment No_high



Note: Considering only cases, when a deal was reached.

The first attribute refers to the gender of the employer, e.g. in the paring "Female-Male", a female employer is interacting with a male employee. We only report employers'-shares, as when looking at specific gender pairings the shares of the employees' are exactly determined and correspond to the residual.[23] We perform the analysis from both the employers' and the employees' perspective by comparing a specific gender pairing (for example "Female-Male") to the gender parings where either the gender of the employer ("Male-Male") or that of the employee ("Female-Female") is varied.

The results for treatment Info_low can be found in the appendix (figure 2.13). Similar to the previous subsection, in treatment Info_low there are no significant differences in outcomes for any gender pairing. In the second treatment, however, there are clear differences in gender pairings (see figure 2.4): Male employers receive a substantially higher share, when bargaining with a woman, instead of bargaining with a man (rank-sum test, $p = 0.05$). Similarly, women earn less than men in the role as employer, when paired with a female employee. In the latter case, effects are strongest (rank-sum test, $p = 0.01$), but in these categories there also exist more observations. The effect seize is about the same ($Cohen's\ d = 0.81$ and $d = 0.84$).

[23] $(employee's\ share = 100 - employer's\ share)$

The previous differences for the pairings "Male-Female" and "Male-Male" are not present in treatment No_high (see figure 2.5). Female employers achieve worse outcomes, when bargaining with male employees. However, the differences remain insignificant (rank-sum test, $p = 0.70$). From the analysis of bargaining behavior in section 2.3.4, it will be clear that the results are driven by male employees, who do not make as many concessions as female employees do.

2.3.3 Regression Analysis

In this subsection, we report regression analyses that examine gender interaction effects. This allows us to take several control variables into account.[24] Again, we analyze data role-dependently.

Gender pairing effects are only found in treatment Info_high. From the employer's perspective(see table 2.3), males achieve significantly better payoffs than female employers when facing a female employee ($p = 0.01$). As the interaction term suggests, this difference almost completely disappears, when bargaining with a male employee. The interaction term is not found to be statistically significant, but it is nevertheless in the direction we expect, given the earlier analysis. Analogously, from the perspective of employees (see table 2.4), facing a male employer leads females employees to receive significantly worse payoffs ($p < 0.01$). This effect disappears if the employee is also male.

In treatments Info_low and No_high we find no gender or gender pairing effects for either role. Instead the behavior of subjects is significantly influenced by their expectations about the outcome of the negotiation and their risk preferences.[25] Higher expectations of employers have a positive influence on payoffs both in treatment Info_low and in treatment No_high. Players with higher expectations demand higher shares which pays off on average. More risk-loving employees in treatment No_high achieve significantly better deals. But it is important to note that we focus on the cases, where a deal was reached.

[24] As control variables we include expectations, risk-preferences, age and the age of the opponent. More detailed information about how expectations and risk preferences are measured can be found in section 2.3.5 Both age and age of opponent are measured in a scale consisting of five categories. The other variables, place of residence and occupation, do not have any significant impact; especially as in our sample the vast majority of subjects have exactly the same characteristics.

[25] In treatment No_high also the age of the opponent has a significant effect, despite being not revealed.

Table 2.3: Regression: Employer's share

	Info_low	Info_high	No_high
male=1	-1.443	5.863*	-1.980
	(2.186)	(2.281)	(2.645)
male_opp=1	-1.146	0.480	-2.516
	(2.039)	(2.716)	(2.722)
male=1 X male_opp=1	0.683	-5.857	7.286
	(3.313)	(3.851)	(4.406)
expected_share	0.339**	-0.127	0.560***
	(0.116)	(0.109)	(0.125)
risk_loving	-0.173	1.071*	0.369
	(0.379)	(0.483)	(0.563)
age	-1.104	-2.084	-0.240
	(0.907)	(1.321)	(1.195)
age_opp	1.081	-1.169	-3.402*
	(0.987)	(1.129)	(1.404)
Constant	38.33***	67.02***	35.08***
	(7.247)	(8.313)	(8.113)
Observations	72	71	65

Standard errors in parentheses

Considering only the cases, in which a deal has been reached

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.4: Regression: Employee's share

	Info_low	Info_high	No_high
male=1	0.725	-1.249	0.605
	(2.195)	(2.700)	(3.024)
male_opp=1	0.351	-7.022**	-1.614
	(2.278)	(2.218)	(2.791)
male=1 X male_opp=1	-0.162	6.897+	-0.386
	(3.502)	(3.811)	(4.896)
expected_share	-0.0978	0.203*	0.180
	(0.086)	(0.092)	(0.122)
risk_loving	-0.00628	0.374	1.755**
	(0.415)	(0.391)	(0.608)
age	-1.807+	0.568	1.038
	(1.010)	(1.137)	(1.581)
age_opp	1.176	1.919	0.792
	(0.955)	(1.299)	(1.279)
Constant	50.69***	25.96***	19.68**
	(5.894)	(6.514)	(7.234)
Observations	72	71	65

Standard errors in parentheses

Considering only the cases, in which a deal has been reached

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This makes these effects potentially biased, as higher expectations and more risk-loving behavior are positively correlated with the probability of not reaching an agreement, which counterweights the previously described effects.

We additionally analyze gender revealing effects using regression analysis contrasting treatment Info_high with treatment No_high. This makes it possible to evaluate the effect of gender revealing on behavior. The regression includes gender pairings and interaction effects between gender and treatment. The results are presented in table 2.5. Treatment Info_high corresponds to the baseline. There male employers receive higher average payoffs than female employers but these differences are not statistically significant. Facing a male employer significantly decreases a female employee's share in treatment Info_high ($p = 0.01$). Both interaction terms "treatment x opponent's gender" and "gender x opponent's gender" in the employee's share regression show that the decrease in payoffs is reduced in treatment No_high and if the employee is male. Again, the interaction terms are not significant, but they are in the expected direction.

Table 2.5: Regression: Effects of gender revealing

	Employer's share	Employee's share
male=1	3.623	0.278
	(2.306)	(2.576)
treatment No_high=1	1.276	-1.795
	(2.241)	(2.175)
male=1 X treatment No_high=1	-2.597	-0.289
	(2.936)	(2.984)
male_opp=1	-1.965	-5.483*
	(2.623)	(2.221)
male_opp=1 X treatment No_high=1	1.599	3.003
	(3.064)	(2.879)
male=1 X male_opp=1	-1.698	3.204
	(3.073)	(3.035)
expected_share	0.168+	0.174*
	(0.087)	(0.073)
risk_loving	0.719+	0.935**
	(0.392)	(0.333)
age	-0.958	0.787
	(0.935)	(0.937)
age_opp	-1.937*	1.116
	(0.940)	(0.901)
Constant	52.18***	25.63***
	(6.182)	(4.958)
Observations	136	136

Standard errors in parentheses

Comparing only treatments Info_high and No_high with treatment Info_high as baseline.

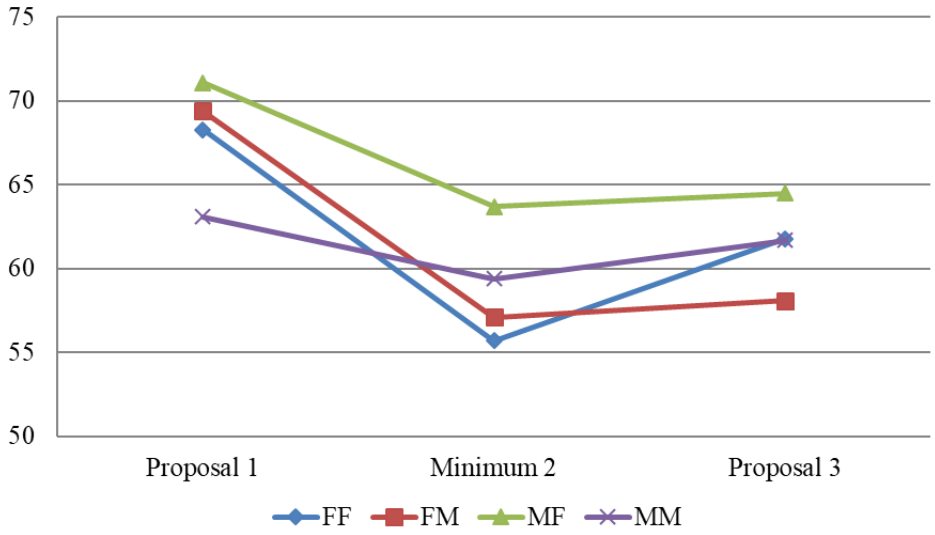
Considering only the cases, in which a deal has been reached.

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2.3.4 Analysis of Bargaining Behavior

In this subsection we analyze bargaining strategies employed by the players. Our design allows us to see how shares demanded by each party evolve per round. In figures 2.6 and 2.7 average demands of employers in the first 3 rounds for different gender pairings are shown. The abbreviations "F" and "M" stand for "female" and "male". The same information is displayed in figures 2.8 and 2.9, for the employee's perspective. Note, that employers always make the proposal in the first round ("proposal 1") and then state their minimum in the second round ("minimum 2") etc. For employees this applies analogously in reversed order. As we did not find significant effects in treatment "Info_low", we only report bargaining behavior of treatments Info_high and No_high[26].

Figure 2.6: Treatment Info_high: Employers' behavior



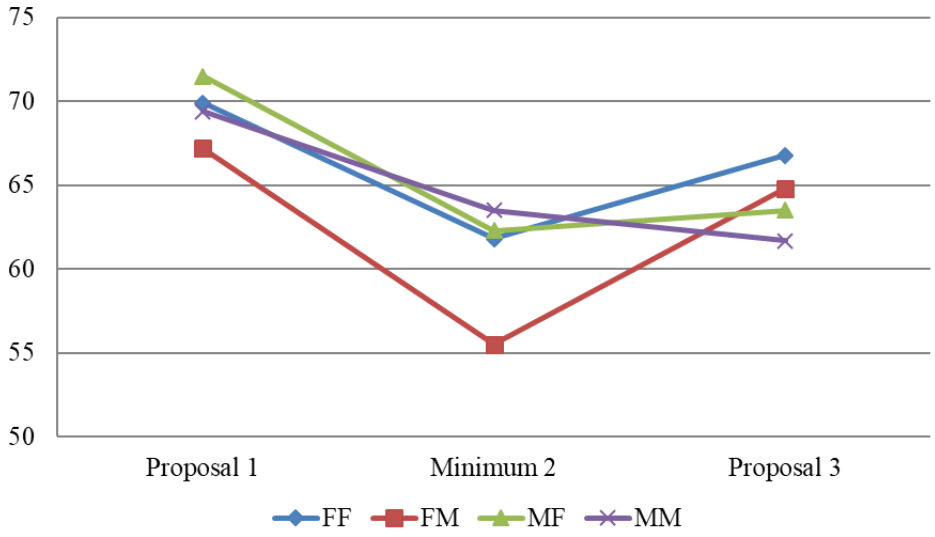
In general there is a moderately high path-dependency between proposals per round.[27] It is lower in treatment Info_high than in treatment No_high. Additionally, we find substantial differences between employers and employees: While the correlation between the first and second round proposal is very similar for both sides[28], the correlation between

[26] In treatment Info_low, average demands for each role are quite close to each other throughout all rounds. We do not find significant differences in bargaining strategies concerning gender pairings.

[27] The Pearson correlation coefficient ("PCC") between proposals of consecutive rounds mostly has values between $0.50 - 0.70$.

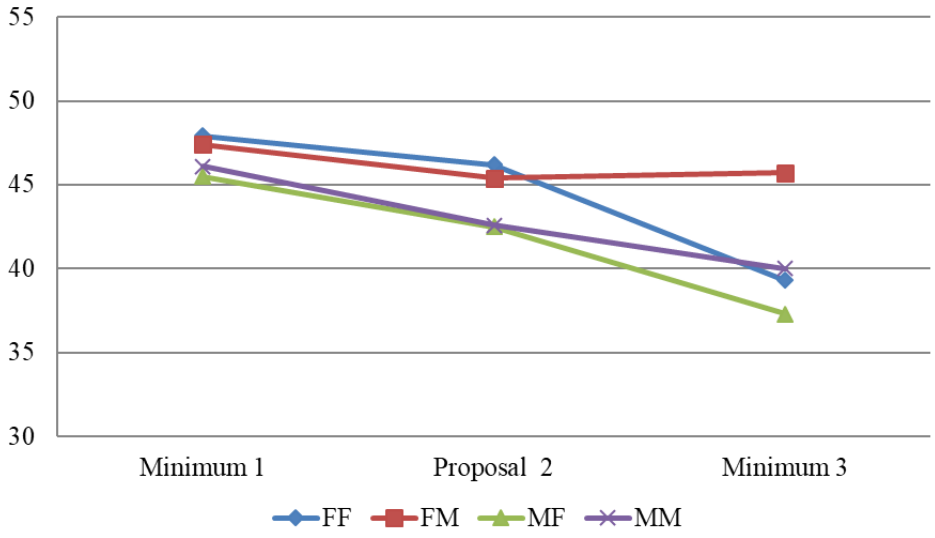
[28] The PCC is around 0.60 in each case.

Figure 2.7: Treatment No_high: Employers' behavior



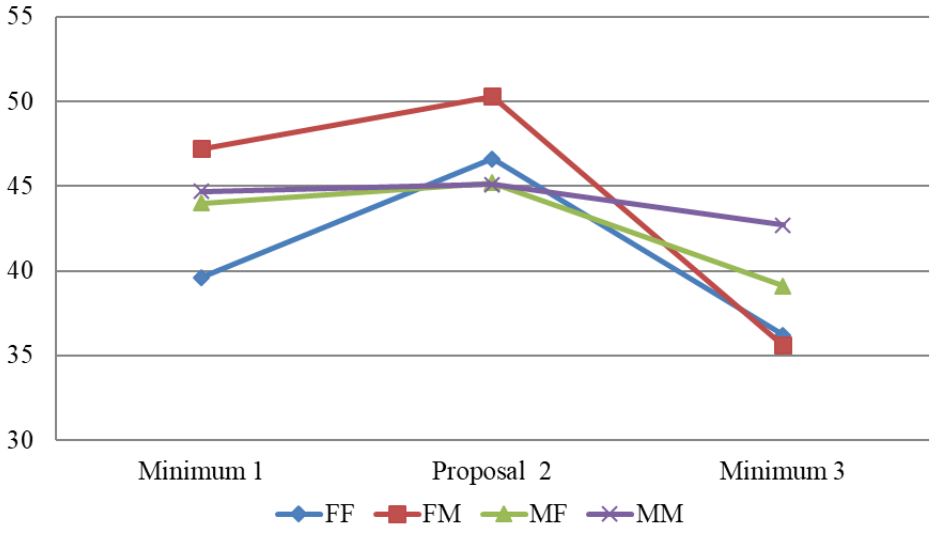
round 2 and 3 increases for employers and decreases strongly for employees.[29] As results below suggest, this effect is mainly driven by females making bigger concessions between rounds 2 and 3 when being an employee (see the green and blue lines in figures 2.8 and 2.9).

Figure 2.8: Treatment Info_high: Employees' behavior



[29]PCC= 0.76 for employers and PCC= 0.41 for employees.

Figure 2.9: Treatment No_high: Employees' behavior



With respect to gender pairing effects, in treatment Info_high we find that as employers men behave more aggressively, when facing a woman (see the green line in figure 2.6). Men submit significantly higher proposals in the first round, when interacting with a female employee, compared to interacting with a male employee (rank-sum test of proposal 1, $p = 0.04$). Also in round 2, male employers lower their demands significantly less, when facing a woman, than female employers do. (rank-sum test of minimum 2, $p = 0.05$). In treatment No_high these differences disappear (see figure 2.7). Looking at the red line in figure 2.7 it appears as if in treatment No_high female employers ask for smaller shares in round 2 when facing a male employee, but these differences are not significant.

Focusing on employee behavior, one can see that in treatment Info_high (figure 2.8) all gender pairings start with similar minimums, but especially female employees make bigger concessions than male employees in subsequent rounds. This leads to significant differences between the pairing FF and FM round 3 (rank-sum test, $p = 0.08$). As a substantial amount of deals already have taken place in the first two rounds, we lack power to adequately analyze if the differences are significant.

In treatment No_high, men as employees start with higher demands than women (rank-sum test of minimum 1, $p = 0.07$) against female employers (see figure 2.9). As genders are not revealed in this treatment, one might pool the values for both opponents' genders. When doing so, women demand on average slightly less than men do, but significance

does not change (rank-sum test of minimum 1, $p = 0.07$).

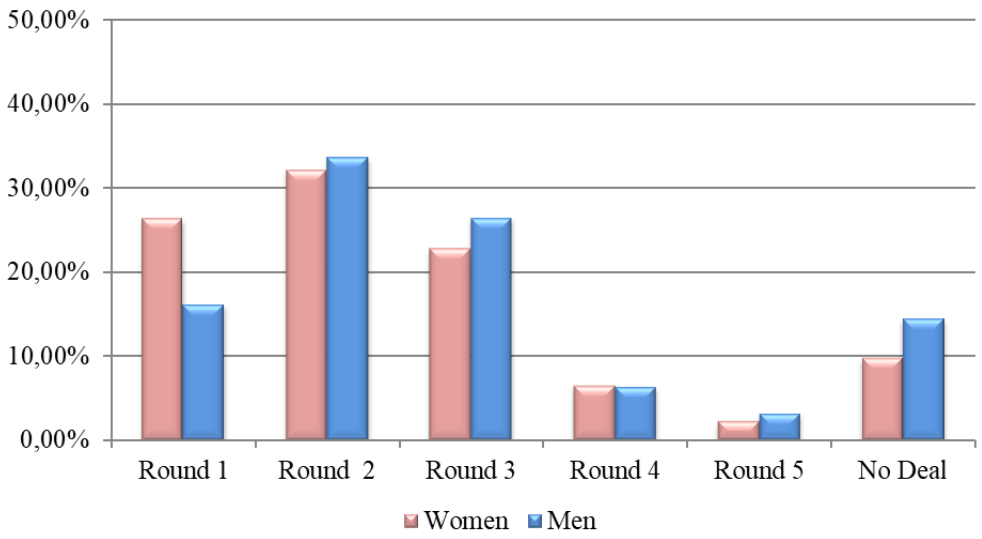
Overall, when genders are revealed, men employ a more aggressive bargaining strategy. They demand higher shares in both roles and make smaller concessions in subsequent rounds. The opposite effect is true for women. This is a likely mechanism to explain, why men are able to achieve better outcomes in treatment Info_high in mixed gender pairings.

2.3.5 Further Analysis

I. When is a deal achieved?

In figure 2.10, the numbers of deals reached within each round are depicted by gender.

Figure 2.10: Fraction of deals reached (all treatments)



Women appear to behave more deal-oriented in both roles and as shown in section 2.3.4 for this purpose they are willing to accept lower shares. As a result, women reach a deal earlier than men. Frequencies for achieving an agreement in the first round are significantly different between genders as well as for the cases, when no deal is reached.

II. Outcome Expectations

In order to better understand bargaining behavior, we asked subjects beforehand, which share they are expecting to receive.[30] In figure 2.11 and 2.12 average expected shares for both roles are displayed treatment- and gender-wise.

Figure 2.11: Mean expected shares of employers

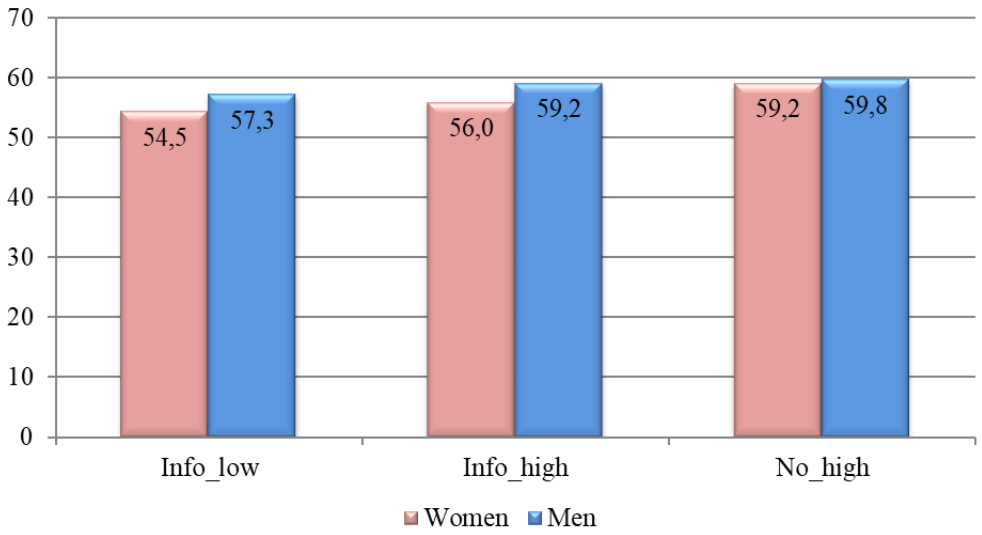
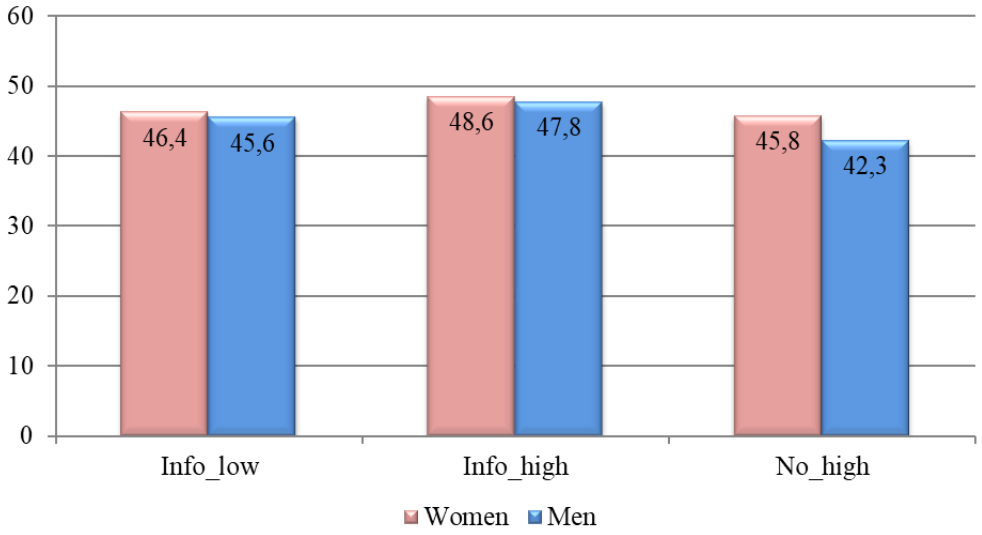


Figure 2.12: Mean expected shares of employees



[30] On a scale ranging from 0-100.

An interesting pattern emerges: Male employers expect higher payoffs than women across all treatments.[31] This exactly reverses for the employee-side. In this role women expect higher shares across all treatments.[32] One possible reason could be that men accept to a greater extent the strategic aspect of the situation and adjust their expectations accordingly. Hence, their beliefs are closer to payoffs of the game-theoretic equilibrium, which are still far more unbalanced. In contrast, women's expected shares are always closer to an equal split.

Subjects' expectations play an important role, as they are strongly correlated to their first round offers.[33] Thus, gender differences in bargaining behavior can to some extent be explained by gender differences in expectations. Note that subjects at this point in time are only informed, that demographic information will later be revealed during the negotiation, but they do not know the exact gender of their opponent yet.

III. Risk-Preferences

A subject's bargaining strategy can be expected to depend on her risk attitude. Subjects always face a tradeoff between aiming for a higher share, while accepting a lower probability of success or the other way around. More risk-averse subjects would potentially post lower demands for both roles. The literature suggests that women are more risk-averse than men, which is confirmed by our data: On a scale ranging from 0-10, women assess their risk tolerance on average with 4.7 points, while men have an average value of 5.6. These differences are strongly significant (rank-sum test, $p < 0.001$). Lower demands by women can to some extent be explained by differences in risk-attitudes. There exists a positive correlation between risk-preferences and both the first proposal and the first minimum[34]. As this factor clearly plays an important role, we have taken it into account as a control variable, when analyzing gender pairing effects in outcomes. Still, as we have seen in treatment Info_high the differences in gender parings remain significant, even when controlling for risk-preferences.

[31] Differences are significant at the 10%-level in treatments Info_low and Info_high (two-sided t-test).
[32] But differences are not significant at treatment level.
[33] Pearson's correlation coefficient is 0.36 for the correlation between expected share and proposal 1, being highly significant, $p < 0.001$.
[34] Spearman's rho=0.18 and Spearman's rho=0.22 respectively, $p < 0.01$ in both cases.

2.4 Conclusion

This paper examines gender differences and gender paring effects in an experimental bargaining setup. In line with previous findings, we provide further evidence, that a certain degree of asymmetry concerning bargaining power is a necessary condition for gender differences to occur. In the symmetric treatment subjects tend to choose an equal split of the pie. This may be due to participants underestimating the amount of bargaining power the employer has, as seem from their expectations for this role being clearly lower than the game theoretic predictions. Alternatively, perhaps they might be influenced by fairness considerations or social norms and do not try to exploit their bargaining power to the full extent.[35]

Under higher asymmetry we find gender differences in behavior and outcomes. Differences are twofold in nature: On the one hand, we observe gender differences in preferences, which are of an inherent type. Women are more risk averse and their expectations are closer to the equal split for both roles. This seems to make them more deal-oriented, as women have lower demands on average for both roles, even when genders are not revealed.[36] This allows female participants to reach more agreements in total and in earlier rounds. On the other hand, we provide evidence for gender differences and gender pairings effects due to gender identity effects. When genders are known, men choose, independently from their role, a more aggressive bargaining strategy, when facing a female counterpart. This effect is in line with findings from Dittrich et al. (2014) and Hernandez-Arenaz and Iriberri (2018). In contrast, women make more concessions in mixed gender pairings. Taking these two behavioral patterns together, women achieve worse outcomes, when interacting with a male participants, when information about genders is provided. It is important to note, that when examining average payoffs of all pairings while including the cases where no deal was reached, the male advantage is reduced almost completely. These findings are also in line with the above mentioned literature.

[35] Similar to how in Ultimatum games subjects regularly offer an equal split, despite their first mover advantage.

[36] See results in treatment No_high.

The results can help to explain part of the gender wage gap that remains after controlling for known factors such as education, job experience, work force interruptions, occupations and industries. In real-world wage negotiations, genders are usually known to the interacting partners. As our results suggest, this could have an important effect on earnings, as in these kind of situations, men achieve significantly better outcomes than women. On the other hand, if field data is in line with the experimental findings, this would also mean, that men more often fail to reach an agreement. A phenomenon that is not reflected in the gender wage gap, as this is only based on observed wages.

We implemented a rather subtle manipulation to inform subjects about the gender of their bargaining partner. As a next step one could test the effects of revealing gender in a more salient way, for example, by providing it as the only piece of demographic information.

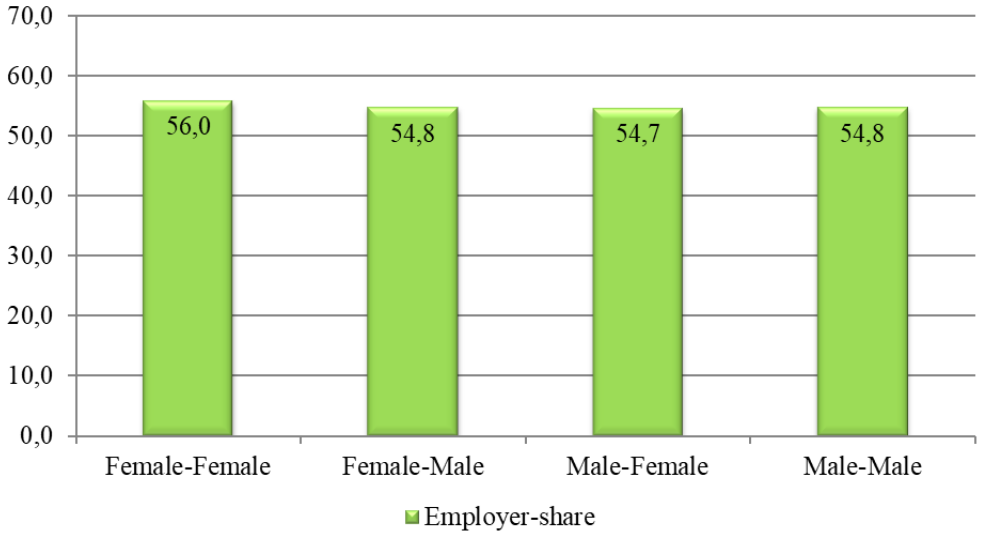
Another idea for future research would be to further investigate the motivational forces underlying gender differences in behavior, given they exist. One explanation could be that differences stem from differences in beliefs about the behavior of the other person. For example, men might think that women are more easily willing to make concessions than men and for this reason will demand larger shares, when bargaining with a woman. Alternatively, differences might result from changes in own behavior by acting according to socially accepted role models (see e.g. congruity theory by Eagly & Karau, 2002). This could for example lead women to behave less demanding when they are aware that their gender is known to the bargaining partner. A possible design for testing both hypotheses would be to unilaterally reveal gender. This would allow one to distinguish between those two explanations mentioned before and isolate individual effects.[37]

[37] Consider for example the case of a male person A bargaining with a female person B and person A is unilaterally informed about the gender of person B. If only the first explanation is true, person A will change his behavior, while B will not. If only the second explanation is true, person B will change her behavior, while A will not. Accordingly, if both persons change their behavior both factors would seem to play a role.

Appendix

A.1 Mean Payoffs Treatment Info_low

Figure 2.13: Mean payoffs in treatment Info_low

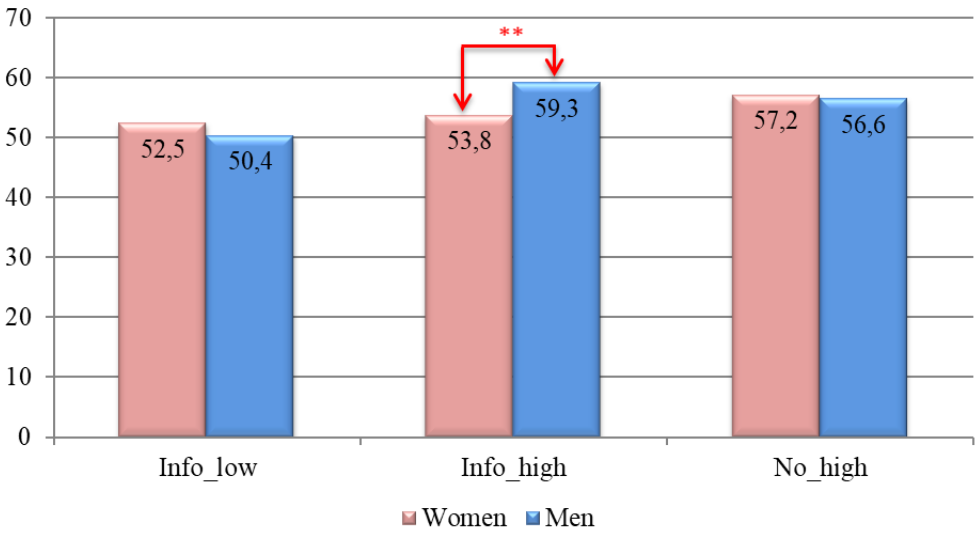


Note: Considering only cases, when a deal was reached

A.2 Mean Payoffs (including cases, when no deal was reached)

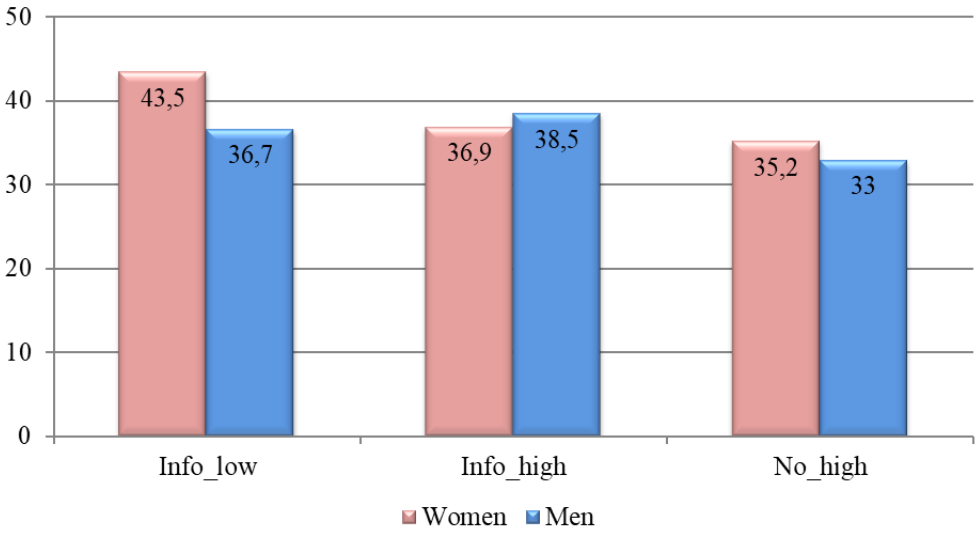
In figures 2.14 and 2.15 average outcomes of women and men are displayed for each role and treatment. In contrast to before, these values also include the cases, when no deal has been reached and the game was terminated automatically. As before, the only significant effect is found in treatment Info_high. Again, male employers achieve significantly better outcomes than female employers (rank-sum test, $p < 0.01$). The remainder provides a rather mixed picture: The worst outcomes belong to male employees in treatment No_high. They are paying the price for having high demands, that lead to several cases, in which no deal was achieved.

Figure 2.14: Mean payoffs of employers



* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: Including cases, when no deal was reached

Figure 2.15: Mean payoffs of employees



Note: Including cases, when no deal was reached

A.3 Instructions (in German)

Experiment - Allgemeine Informationen

Herzlich willkommen zu diesen Experiment und vielen Dank für Ihre Teilnahme!

Bitte schalten Sie nun Ihr Handy aus und kommunizieren Sie ab jetzt nicht mehr mit den anderen Teilnehmern. Wenn Sie eine Frage haben, heben Sie bitte die Hand, wir kommen dann zu Ihnen an den Platz und beantworten diese persönlich.

In diesem Experiment können Sie einen gewissen Geldbetrag verdienen, der Ihnen am Ende in bar ausbezahlt wird. Die genaue Höhe dieses Betrags ist abhängig von den von Ihnen getroffenen Entscheidungen sowie denen Ihrer Mitspieler. Für Ihre ordnungsgemäße Teilnahme erhalten Sie einen Grundbetrag in Höhe von 4€. Zudem bekommen Sie den erzielten Betrag aus der nachfolgenden Entscheidungssituation. Und zusätzlich können Sie 1€ verdienen, falls Sie mindestens sechs der sieben Kontrollfragen richtig beantworten.

Somit ergibt sich als Gesamtauszahlung für Sie:

$$\boldsymbol{Gesamtauszahlung} = \mathbf{4}€ + (\mathbf{1}€) + \boldsymbol{Betrag\ aus\ Entscheidungssituation}$$

Ihre Angaben und Entscheidungen während des Experiments werden komplett anonym behandelt und können nicht mit Ihrer Person/Identität in Verbindung gebracht werden.

Der generelle Ablauf des Experiments ist für alle Teilnehmer identisch und wird im Folgenden näher beschrieben:

1) Zuerst füllen Sie einen kurzen Fragebogen mit allgemeinen demografischen Angaben aus.
2) Danach erhalten Sie detaillierte Instruktionen zu der darauffolgenden Entscheidungs-situation.
3) Bevor diese durchgeführt wird, beantwortet jeder Teilnehmer ein paar kurze Kontrollfragen.
4) Nun treffen Sie Ihre Entscheidungen.
5) Wenn der Prozess beendet ist, erhalten Sie einen weiteren kurzen Fragebogen.
6) Danach findet die Auszahlung statt.

Anleitung Entscheidungssituation

Im Folgenden bilden Sie und ein anderer Teilnehmer für den Rest des Experiments eine **2er Gruppe**. Ein Teilnehmer der Gruppe erhält dabei die Rolle **„Arbeitgeber" (AG),** der andere die Rolle **„Arbeitnehmer" (AN)**. Die **Einteilung** der Gruppen und Rollen **erfolgte zufällig.**

Ihre zugeteilte Rolle wird Ihnen am Computer angezeigt.

Beide Partien befinden sich in einer Verhandlungssituation über ein mögliches Beschäftigungsverhältnis, welches einen **Gewinn von 100 Geldeinheiten erzielt, über dessen**

Aufteilung Sie innerhalb Ihrer Gruppe verhandeln können. Dazu werden Sie und Ihr Verhandlungspartner pro Runde abwechselnd einen Vorschlag machen, wie die Aufteilung der 100 Einheiten erfolgen soll. Der Arbeitgeber ist in der Rolle als "Startspieler" und macht den ersten Vorschlag. Sie bzw. der andere Teilnehmer geben währenddessen an, ab welchem Minimalanteil Sie den Vorschlag gerade noch akzeptieren würden. Ist der angebotene Anteil größer oder gleich diesem Wert, kommt der Vorschlag automatisch zustande und die Aufteilung der 100 Einheiten erfolgt gemäß diesem Vorschlag. Falls der angebotene Anteil kleiner als Ihr geforderter Minimalanteil ist, wird der Vorschlag abgelehnt und es beginnt eine neue Verhandlungsrunde, in der der jeweils andere Teilnehmer an die Reihe kommt und einen neuen Aufteilungsvorschlag macht.

Beispiel 1: Spieler A bietet als Aufteilungsvorschlag an, 60 Geldeinheiten für sich und 40 Geldeinheiten für Spieler B. Das akzeptierte Minimum von B für diese Runde beträgt 30 Geldeinheiten. Dieser Vorschlag kommt zustande und die daraus resultierende Aufteilung ist 60 Geldeinheiten für Spieler A und 40 Geldeinheiten für Spieler B.

Beispiel 2: Falls das akzeptierte Minimum von B 50 Geldeinheiten (statt wie zuvor 30) beträgt, kommt der Vorschlag nicht zustande und es beginnt eine neue Runde, in der Spieler B einen Gegenvorschlag macht.

Dieser Verhandlungsprozess geht dann immer so weiter, bis entweder ein Vorschlag angenommen wird oder das Verhandlungsende eintritt. Das Verhandlungsende kann frühestens nach Runde 3 eintreten. Falls in Runde 3 oder einer der nachfolgenden Runden keine Aufteilung zustande kam, wird jeweils per Zufallsgenerator ermittelt, ob eine weitere Verhandlungsrunde durchgeführt wird oder nicht. Die Wahrscheinlichkeit, dass es eine weitere Runde gibt, beträgt konstant 80%.

Der Teilnehmer mit der Rolle **„Arbeitgeber“ erhält einen Anteil von 40 [*20*] Geldeinheiten, falls das Verhandlungsende eintritt** und zuvor kein Aufteilungsvorschlag angenommen wurde, während der Teilnehmer mit der Rolle **„Arbeitnehmer“ 0 Geldeinheiten erhält.**

Am Ende des Experiments werden die Geldeinheiten zu einem Kurs von 10:1 in Euro umgerechnet, d.h. für je 10 Einheiten erhalten Sie eine Auszahlung von 1€ für diesen Teil.

Zusätzlich erfahren Sie und Ihr Mitspieler zuvor gegenseitig über sich die zuvor ausgefüllten demografischen Angaben wie beispielsweise Alter, Geschlecht, Beschäftigungsstatus…

[*Die zuvor ausgefüllten demografischen Angaben wie beispielsweise Alter, Geschlecht, Beschäftigungsstatus…werden den Verhandlungspartnern nicht mitgeteilt.*]

Chapter 3

Mutual Knowledge of Preferences

Does mutual knowledge of preferences lead to more Nash equilibrium play? Experimental evidence[†]

Abstract

In some experiments, Nash equilibrium fails to accurately predict behavior. Usually, it is assumed that the monetary payoffs in the game represent subjects' utilities. However, subjects may actually play a very different game. In this case, mutual knowledge of preferences may not be satisfied. We run an experiment where we first elicit subjects' preferences over the monetary payoffs for all players. This allows us to identify equilibria in the games that subjects actually are playing. We then examine whether revealing other subjects' preferences leads to more equilibrium play and find that this information indeed has a significant effect. Furthermore, it turns out that subjects are more likely to play maxmin and maxmax strategies than Nash equilibrium strategies. This indicates that subjects strongly rely on heuristics when selecting a strategy.

[†]Joint work with Christoph Brunner and T. Florian Kauffeldt

3.1 Introduction

Applied game theory usually relies on the standard Nash equilibrium (Nash et al., 1950, Nash, 1951). At the same time, it remains largely unclear whether the underlying (implicit) assumptions are met. For instance, it is often not clear whether there is common knowledge about payoffs and rationality. In particular, common (or, at least mutual) knowledge about preferences is a core assumption in game theory. In the words of Polak (1999):"In games of complete information, common knowledge of payoffs is usually taken to be implicit. Indeed, this is often taken to be the definition of complete information."[36]

Mutual or even common knowledge about preferences is not only assumed in traditional game theory, but also often in behavioral game theory. In fact, most level-k models assume that payoffs are mutually known and that agents form beliefs about other agents' play based on this information (see, e.g., Costa-Gomes et al., 2001). Other approaches such as Quantal Response Equilibrium[37] introduced by McKelvey and Palfrey (1995) incorporate a stochastic element that can be interpreted as uncertainty about other players' preferences.

Despite the ubiquity of the (implicit) assumption of mutually or commonly known preferences, there is little empirical evidence about the degree to which it affects the reliability of the Nash prediction. However, previous experimental research suggests that it should not be taken for granted. For example, Healy (2011) finds that subjects fail to accurately predict other subjects' preferences over possible outcomes in normal-form 2×2 games. The purpose of the experiment reported in this paper is to test whether mutual knowledge of preferences is important for the Nash prediction in dominance-solvable 2×2 games.

Our results can be summarized as follows: (1) subjects are indeed significantly more likely to play a Nash equilibrium strategy when they are informed about their opponents' preferences over the possible outcomes of the game. When preferences are not mutually

[36]This means that "complete information" cannot be part of the rules of the game (*the game-form*) because it involves assumptions about knowledge of individual preferences. In general, a game with complete information can be seen as an interactive situation in which both the game-form and players' preferences are commonly known.

[37]In Quantal Response Equilibrium, there is an error term in players' payoff functions whose distribution is assumed to be known.

known, the frequency of equilibrium play is rather low.

(2) A strategy is more likely to be played when it cannot lead to the lowest payoffs (*maxmin strategy*) or when it can lead to the highest one (*maxmax strategy*). Furthermore, maxmin and maxmax strategies predict behavior better than Nash equilibrium strategies, especially when preferences are not mutually known.

Result (1) shows that subjects not only fail to accurately predict other players' preferences as previous evidence already suggests, the lack of such information also significantly affects their behavior. Whenever it is unlikely that players know each other's preferences and some players have no strictly dominant strategy, it might therefore be advisable to use a more general equilibrium concept. Following Polak (1999), we may view a situation where preferences are not mutually known as a game with incomplete information. Such a situation can then be modeled as a Bayesian game (Harsanyi, 1967).[38] Result (2) suggests that subjects largely rely on heuristics rather than on strategic considerations. The reason may be that subjects do not believe that the other player is rational and/or are uncertain about his payoff function.

Theoretically, in the tested dominance-solvable 2×2 games, mutual knowledge of payoff functions along with mutual knowledge of rationality[39] suffices to ensure that agents will play a Nash equilibrium.[40] To see this, suppose one player (called "D") has a strictly dominant strategy. Given that D is assumed to know his own payoff function and is rational, D will play his dominant strategy. The other player (called "ND") believes that D is rational and that he has a strictly dominant strategy. Therefore, ND believes that D will play this strategy. Since ND is himself assumed to be rational and to know his own payoff function, ND will play a best response to D's dominant strategy.

[38] Players with different preferences can be thought of as different types and it is then assumed that the prior distribution of types is commonly known. This approach has been used in various fields. In auction theory, for example, the assumption that all bidders are risk neutral and that this is commonly known has been relaxed. Instead, the prior distribution of risk preferences rather than other bidders' actual risk preferences are assumed to be commonly known (see, e.g., Hu & Zou, 2015).

[39] Note that there is a difference between "knowledge" and "(probability one) belief". Roughly, "knowledge" refers to true belief justified by either direct observation or logical deduction, whereas "belief" may be false. Therefore, it would be more accurate to assume that players believe that others are rational with probability one.

[40] See Aumann and Brandenburger (1995) for sufficient conditions that ensure Nash equilibrium in general normal-form 2×2 games.

Uncertainty about opponent's rationality and/or payoff function can lead to uncertainty about the other agents' strategy choices with unknown probabilities. Uncertainty about probabilities (ambiguity) can affect peoples' behavior, as Ellsberg (1961) showed.[41] The strategic ambiguity model of Eichberger and Kelsey (2014) shows that maxmin or maxmax strategies can be a best response to strategic ambiguity.[42]

3.1.1 The experiment

In stage 1 of the experiment, we elicit subjects' preferences over monetary payoff pairs (they will be referred to as "payment pairs"). The same payment pairs are then used to construct four different 2×2 games (or more precisely four different game-forms). In stage 2, each subject plays each of these games exactly once. This design allows us to avoid the assumption that subjects only care about their *own* monetary payments. Instead, we can use the preferences elicited in stage 1 to describe the game that our subjects play.[43] This will be illustrated with the help of Example 3.1 below, which corresponds to one of the games played in the experiment.

Example 3.1. *Consider the prisoner's-dilemma-type game-form in Figure 3.1. The numbers in the matrix correspond to the amount of money paid to the players, where the first number is the row player's payment and the second number is the column player's payment.*

Let r be the row and c be the column player of the game in Example 3.1 and denote the payment pairs by $(x_r, x_c) \in \mathbb{R}^2$. Suppose that both players

[41] When people face ambiguity, they frequently do not behave as if they were governed by subjective probabilities.

[42] This model allows for optimistic responses to strategic ambiguity. Most other strategic ambiguity models such as those of Lo (1996), Eichberger and Kelsey (2000), and Lehrer (2012) assume ambiguity-averse behavior. While these models can explain maxmin behavior, they cannot rationalize maxmax behavior.

[43] We maintain the assumption that preferences depend only on players' monetary payments. That is, the specific game-form, other subjects' preferences, or any other factors have no effect on subjects' ordinal ranking of payment pairs. Of course, this is to some degree a consequentialist approach and consequentialism has been criticized in the literature repeatedly. In Section 3.3.4, we will discuss evidence suggesting that such considerations do not play an important role in the games used in this study. However, we cannot completely exclude that violations of consequentialism might have caused some noise and the "true" treatment effect is even higher.

Figure 3.1: Prisoner's-dilemma-type game-form

	L	R
U	$4, 4$	$8, 3$
D	$3, 8$	$7, 7$

(a) are selfish. That is, each player's preferences over payment pairs can be represented by a strictly monotone increasing utility function $v_i(x_i)$ $(i \in \{r, c\})$ that depends only on his own payment or

(b) have other-regarding preferences represented by a function $\tilde{v}_i : \mathbb{R}^2 \to \mathbb{R}$,

then the games that result in cases (a) and (b) are depicted in Figure 3.2.

Figure 3.2: Induced games in Example 3.1

	L	R
U	$v_r(4), v_c(4)$	$v_r(8), v_c(3)$
D	$v_r(3), v_c(8)$	$v_r(7), v_c(7)$

(a) Players with selfish preferences

	L	R
U	$\tilde{v}_r(4,4), \tilde{v}_c(4,4)$	$\tilde{v}_r(8,3), \tilde{v}_c(8,3)$
D	$\tilde{v}_r(3,8), \tilde{v}_c(3,8)$	$\tilde{v}_r(7,7), \tilde{v}_c(7,7)$

(b) Players with social preferences

The game that results if players are selfish (a) is a prisoner's-dilemma-type game. For all strictly monotone increasing utility functions, v_i, the game has only one Nash equilibrium (U, L), i.e., everyone defects. That is not necessarily true for the induced game (b), where players have social preferences. For example, if $\tilde{v}_r(7,7) > \tilde{v}_r(8,3)$ and $\tilde{v}_c(7,7) > \tilde{v}_c(3,8)$, then mutual cooperation, (D, R), is a Nash equilibrium in (b).

In this paper, whenever we refer to a "Nash equilibrium", we refer to the Nash equilibrium of the induced game using the preferences elicited in stage 1 of the experiment. We focus on those situations, in which a pure unique Nash equilibrium exists (according to the reported preferences): That corresponds to those cases, in which one player has a strictly dominant strategy and the other player has a non-dominant unique pure Nash equilibrium strategy in the induced game.[44]

[44]In our experiment, we only ask subjects to rank payment pairs ordinally. Eliciting a cardinal ranking of payment pairs would require a more complicated procedure that some subjects might fail to understand. It is not obvious that subjects can reliably assign a cardinal utility to each payment pair. As a result, we cannot compute Nash equilibria in mixed strategies for the induced games. Moreover, we will exclude the decisions of subjects who have a strictly or weakly dominant strategy in the induced game.

Consequently, we consider situations where subjects' opponents have a strictly dominant strategy. In the baseline treatment the reported preferences are not revealed. Hence, subjects cannot be certain that their opponents have a dominant strategy.

For example, suppose the row player in the induced game above (b) is selfish. His pure strategy U is then strictly dominant. A column player who prefers $(4, 4)$ to $(8, 3)$ and $(7, 7)$ to $(3, 8)$ then has a unique equilibrium strategy that is not dominant: L. In treatment baseline, such a column player may not be sure whether row is selfish or not and might therefore occasionally play R rather than L.

In our second treatment (called "info"), the column player can see that row has a strictly dominant strategy and might therefore play the unique equilibrium strategy L more often. Intuitively, this logic can explain our first result that subjects are more likely to play a Nash equilibrium strategy in treatment info compared to treatment baseline. Furthermore, if a subject is uncertain about the strategy choice of his opponent, then, depending on his attitude towards uncertainty, he will try to avoid the lowest ranked payment pair (maxmin), or, to reach the highest ranked one (maxmax). Intuitively, this explains our second result.

3.1.2 Related literature

The papers closest to ours are Healy (2011) and recent working papers by Wolff (2014) and Attanasi et al. (2016).

Healy examines whether the sufficient conditions for Nash equilibrium identified by Aumann and Brandenburger (1995) are satisfied when subjects play normal-form 2×2 games in the laboratory. For that purpose, subjects first chose a strategy and then state their beliefs about behavior and preferences of their opponent. Subjects' own preferences and rationality are also measured. Healy finds that there are only very few instances where all conditions are satisfied. Focusing on mutual knowledge of preferences, he finds that both players correctly predict how their opponent ordinally ranked the payment pairs in

Information about their opponent's preferences is not necessary for those subjects to compute a best response and as a result, information about the other player's preferences should not be expected to have an effect on behavior.

only 64% of games played. Healy concludes that "The failure of Nash equilibrium stems in a large part from the failure of subjects to agree on the game they are playing."

Since mutual knowledge of preferences is one of three conditions that are together sufficient for Nash equilibrium in 2×2 games (see Aumann and Brandenburger, (1995)) and since the other two are also not fully satisfied in Healy's experiment, it is difficult to assess the impact of the failure of mutual knowledge of preferences on equilibrium play in isolation. By introducing a treatment in which information about the opponent's preferences is directly revealed, we can identify the impact of mutual knowledge on equilibrium play by holding all other factors constant.

Wolff (2014) studies behavior in three-person sequential public good games. In contrast to our experiment, he does not reveal subjects' preferences over the material outcomes. Instead, he elicits subjects' best-response correspondences to the contributions of the other players. In one of his treatments, these are then revealed to all group members. This information has a much smaller effect on the frequency of equilibrium play compared to the treatment effect in our experiment.

Revealing best-response correspondences is obviously not sufficient for subjects to be able to predict how much their opponents will contribute: Wolff measures beliefs about others' contributions to the public good and finds that subjects tend to overestimate these. As a result, they often fail to play an equilibrium strategy even though their contributions tend to be consistent with their beliefs and their own reported best-responses. As opposed to the dominance-solvable 2×2 games that we study, several iterations of alternating best responses are required in Wolff's experiment to compute the Nash equilibrium. Some subjects might not be able to do so.

Attanasi et al. (2016) also argue that when subjects have belief-dependent or other-regarding preferences, they are actually playing a game of incomplete information. In their experiment, subjects form beliefs about their opponent's type (e.g., selfish or prosocial) and choose their strategy based on these belief. Attanasi et al. then test whether revealing information about opponent's preferences and beliefs changes behavior in a Mini Trust Game. They find that first movers are more likely to transfer the money when they face

a non-selfish trustee ("guilt-averse" trustee) and vice versa. The Mini Trust Game can be considered as a 2×2 coordination game with two pure equilibria (trust, share) and (not trust, not share). Subjects clearly coordinate better on one of these two equilibria when belief-dependent preferences are disclosed. While this result points in a similar direction as our results, Attanasi et al. do not systematically test the impact of mutual knowledge of preferences on the Nash prediction. In particular, the second movers can observe the decisions of the first movers. Therefore, the preferences of the first movers are not relevant for their strategy choices.

This paper is organized as follows. The next section describes the experimental design. We then present our results, and conclude in Section 3.4. The appendix provides additional information about the experiment.

3.2 Experimental design

Our experiment consists of two treatments (called "baseline" and "info") with two stages each. In the first stage of both treatments, we elicit subjects' preferences over eight different payment pairs. These payment pairs are then used to construct four different 2×2 games. In stage 2 of each treatment, subjects play each one of these games exactly once. In treatment "info", subjects can see their opponent's ordinal ranking of the four payment pairs used in the current game, whereas in treatment "baseline", this information is not disclosed.

3.2.1 Stage 1 of the experiment

Stage 1 is identical in both treatments. Subjects are asked to create an ordinal ranking over the following set X_{row} of eight payment pairs (x_r, x_c):

$$X_{row} = \{(8,3),\ (7,7),\ (5,8),\ (4,4),\ (6,2),\ (3,8),\ (3,3),\ (2,2)\} \subset \mathbb{R}^2 \tag{3.1}$$

The first number, x_r, corresponds to the amount of money (in Euros) paid to the decision-maker in the role of a row player. The second number, x_c, is paid to some other subject

in the role of a column player (the "recipient").[45] Subjects are informed that they will not interact with the recipient in any other way in either stage of the experiment.

The order in which the payment pairs appear on the screen was randomly determined beforehand and remains constant in all sessions. Subjects rank the payment pairs by assigning a number between one and eight to each pair, where lower numbers indicate a higher preference. The same number can be assigned to multiple payment pairs, thus allowing for indifference.

In treatment info, subjects are told that their rankings would be disclosed to other participants at a later stage of the experiment.[46] In treatment baseline, we made it clear that this information would not be revealed. We will explain at the end of this section how the elicitation of preferences was incentivized. After subjects confirm their ranking, they proceed to stage 2, in which they play four one-shot 2×2 games. We ran two waves of experiments. In the first wave, subjects played the games in Figure 3.3 (all numbers are payments in Euro). In the second wave, they played the games in Figure 3.4.

3.2.2 Stage 2 of the experiment

Figure 3.3: Games in wave 1 of the experiment

Game 1

	L	R
U	$4,4$	$8,3$
D	$3,8$	$7,7$

Game 3

	L	R
U	$4,4$	$8,3$
D	$3,3$	$7,7$

Game 2

	L	R
U	$5,8$	$7,7$
D	$6,2$	$3,3$

Game 4

	L	R
U	$8,3$	$2,2$
D	$7,7$	$3,8$

[45] Subjects who were assigned the role of a column player ranked the same payment pairs but the first number corresponds to the other player's payoff. Rewriting X_{row} for column players such that the first number corresponds to the column player's payment and the second to the row player's, we obtain $X_{column} = \{(8,3),\ (7,7),\ (8,5),\ (4,4),\ (2,6),\ (3,8),\ (3,3),\ (2,2)\} \subset \mathbb{R}^2$.

[46] We will discuss the possibility that subjects might strategically misrepresent their preferences in the results section.

Figure 3.4: Games in wave 2 of the experiment

Game 5

	L	R
U	$3,8$	$8,3$
D	$3,3$	$7,7$

Game 7

	L	R
U	$8,3$	$6,2$
D	$7,7$	$5,8$

Game 6

	L	R
U	$8,3$	$2,2$
D	$2,2$	$3,8$

Game 8

	L	R
U	$3,3$	$8,3$
D	$2,2$	$7,7$

As described in the introduction, only one type of strategic situation is of interest (regardless of the monetary payoffs): a unique pure Nash equilibrium in the induced game where one player has a unique equilibrium strategy that is neither weakly nor strictly dominant. This is only possible when the other player has a strictly dominant strategy. Therefore, games in which one and only one player has a strictly dominant strategy are most useful to testing whether mutual knowledge of preferences increases equilibrium play and some of the games were selected accordingly. All games were constructed using the same eight payment pairs. Furthermore, as described below, we made sure that the games exhibit some diversity with respect to the number of pure strategy Nash equilibria under the assumption that subjects are selfish payment maximizers.

The 8 games were selected on the basis of **2 key criteria** that seem to play an important role in the context of our study:

(i) # players, who have a strictly dominant strategy (0, 1 or 2) and

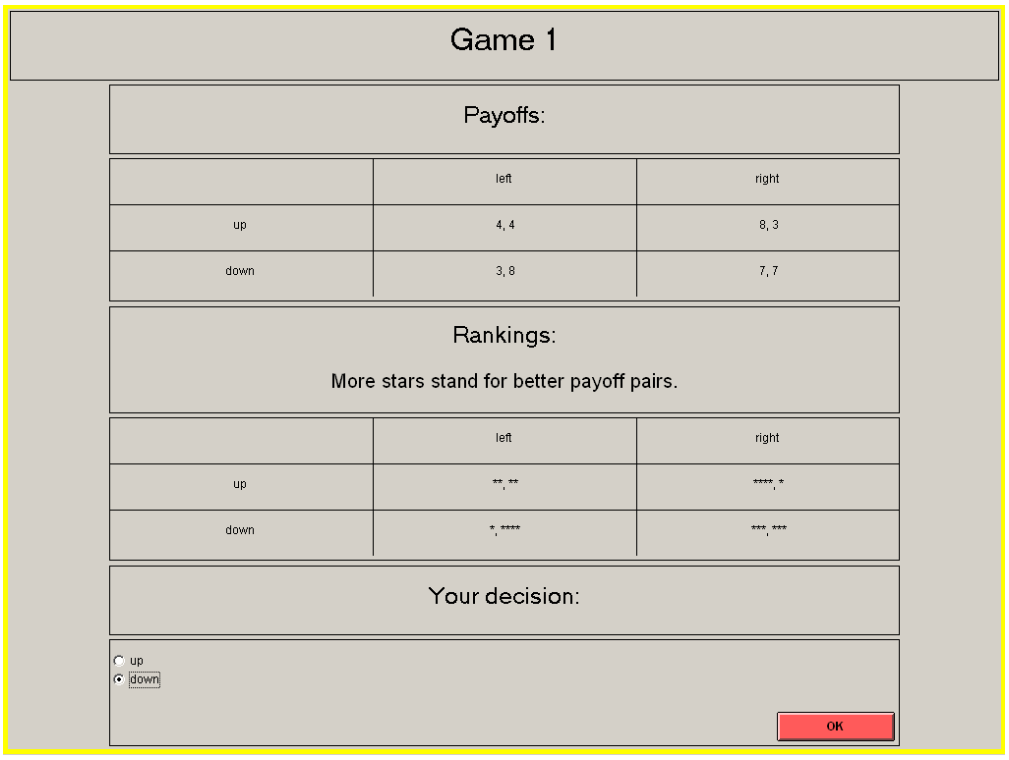
(ii) # pure Nash equilibria (0, 1 or 2).

Both criteria were determined for the case where preferences correspond to monetary payoffs (i.e., on the basis of the game-forms). 2×2 games can be grouped into 6 categories based on these two criteria (some combinations are not possible, e.g., 2 players with strictly dominant strategies and 2 Nash equilibria). Games with more than 2 pure equilibria are unlikely to offer valuable insights for our analysis because they are not expected to generate many relevant observations. We first run wave 1 of the experiment, then we selected the games of wave 2 so that we have at least one game of each of the 6

categories. Furthermore, we wanted to cover most of the 2×2 games that are frequently used in experimental economics (e.g., Prisoners' Dilemma, Matching Pennies, and Battle of Sexes).

In both treatments, subjects can see how they ranked the four payment pairs of the currently played game. This information is displayed by assigning 1-4 stars to each outcome, where more stars indicate a better outcome. In treatment info, subjects are shown both their own *and* their opponent's ranking in matrix-form (see Figure 3.5). Just like in the payment matrix, the first entry corresponds to the subject's own ranking while the second entry reveals the opponent's ranking. In treatment baseline, subjects are shown the same rankings matrix but this matrix only contains their own rankings.

Figure 3.5: Information screen



All subjects play each of the four games of their wave exactly once, each time against a different anonymous opponent. Games are played one after another and feedback about the outcome is only provided at the end of the experiment when subjects are paid, but not while subjects still make decisions.

In both treatments, each subject is paid for exactly one of his decisions, which is randomly selected at the end of the experiment. If a decision from stage 1 is chosen, two of the eight payment pairs from the set X_{row} are randomly selected. The row subject is then paid the first number, x_r, of the payment pair that he ranked more highly in stage 1. The second number, x_c, is paid to some other column subject. In order to avoid reciprocity considerations, we made it clear that the second number is paid to a subject with whom subjects will not interact in the second stage of the experiment. Column subjects are paid in a similar manner.

The probability that stage 1 is paid is $\frac{7}{8}$ while stage 2 is paid with a probability of $\frac{1}{8}$. These probabilities are consistent with selecting each of the $\binom{8}{2}$ possible pairs of payment pairs and each of the four decisions made in stage 2 with equal probability. Paying stage 1 with a substantially higher probability also reduces the odds that subjects might misrepresent their preferences. This issue will be discussed in more detail in Section 3.3.4.

Subjects were given printed instructions and they could only participate after successfully answering several test questions. Test questions as well as the rest of the experiment were programmed using Z-Tree (Fischbacher, 2007). All sessions of the experiment were conducted at the AWI-Lab of the University of Heidelberg. Subjects from all fields of study were recruited using Orsee (Greiner, 2015). Fewer than half of the subjects were economics students. Sessions lasted about 40-50 minutes on average. The following table summarizes the number of participants per session as well as average payments:

Table 3.1: Summary of treatment information

Treatment	Wave	Sessions	Subjects	Average payment
baseline	1	9	97	€ 12.02
baseline	2	7	91	€ 10.54
info	1	8	95	€ 11.78
info	2	7	85	€ 11.41

Decisions made by subjects who made more than 10 mistakes when answering test questions are excluded from the data (including Table 3.1).[47]

[47]The main treatment effect (Table 3.6) is still significant when these 10 subjects are included. In treatment baseline, 2 subjects made more than 10 mistakes, in treatment info, there were 8 such

3.3 Results

In this section, we first characterize subjects' preferences as measured in stage 1 of the experiment. We then present the main treatment effect: **subjects are significantly more likely to play their unique equilibrium strategy in treatment info than in treatment baseline.** This effect can be observed in 6 of the 8 games. **Subsequently, we show that maxmin and maxmax strategies are more likely to be played in both treatments.** We argue that it is unlikely that subjects misrepresent their true preferences or that many preferences changed when subjects are shown their opponents' preferences.

3.3.1 Characterization of measured preferences

In stage 1 of the experiment, we elicit subjects' preferences over the payment pairs $(x_r, x_c) \in X_{row}$ defined in equation (3.1). Tables 3.2 and 3.3 show the ordinal rankings reported by at least two subjects who were assigned the role of a row and column player respectively. Payment pairs that are assigned a lower number are preferred to payment pairs with a higher number.

subjects. It is not plausible that the decisions of the excluded subjects affected other subjects' decisions since all of our games are simultaneous games and subjects were not informed about the decisions of their opponents during the experiment.

Table 3.2: Preferences reported by at least two subjects who were assigned the role of a row player, both treatments. Smaller numbers are assigned to better ranked payment pairs.

(8,3)	(7,7)	(5,8)	(4,4)	(6,2)	(3,8)	(3,3)	(2,2)	n
1	2	4	5	3	6	7	8	63
1	2	4	5	3	7	6	8	15
2	1	4	5	3	6	7	8	15
1	2	4	5	3	6	6	8	12
2	1	3	5	4	6	7	8	10
2	1	3	6	4	5	7	8	7
1	2	3	5	4	6	7	8	5
2	1	3	4	5	6	7	8	5
3	1	2	5	6	4	7	8	3
3	1	2	5	5	3	7	8	3
1	2	3	5	3	6	7	8	2
1	2	3	4	5	6	7	8	2
3	1	2	6	5	4	7	8	2
3	1	2	4	5	6	7	8	2
1	1	4	5	3	6	7	8	2
1	2	5	4	3	7	6	8	2
1	2	4	6	3	5	7	8	2

Table 3.3: Preferences reported by at least two subjects who were assigned the role of a column player, both treatments. Smaller numbers are assigned to better ranked payment pairs.

(8,3)	(7,7)	(8,5)	(4,4)	(2,6)	(3,8)	(3,3)	(2,2)	n
2	3	1	4	7	5	6	8	68
3	2	1	4	7	5	6	8	14
3	1	2	4	7	5	6	8	14
3	1	2	5	6	4	7	8	8
1	3	1	4	7	5	5	7	5
3	1	2	4	8	6	5	7	5
1	3	1	4	8	6	5	7	4
3	2	1	5	7	4	6	8	4
2	3	1	4	7	5	5	7	4
3	2	1	5	6	4	7	8	3
1	3	1	4	7	5	6	8	3
4	1	2	3	6	5	7	8	2
1	3	2	4	8	6	5	7	2
3	1	2	4	6	5	7	8	2
2	3	1	4	6	5	7	8	2
3	2	1	4	8	6	5	7	2
3	1	2	4	8	7	5	6	2

To characterize subjects' preferences, we introduce four properties: pareto-efficiency, strict pareto efficiency, maximization of own payoff, and maximization of total payoff. These properties are defined as follows:

Definition 1 (Pareto efficiency)**.** A subject's preferences $\succsim$ on X are said to satisfy *pareto-efficiency* if, for all $x, y \in X_{row}$, $x \succ y$ whenever $x_r \geq y_r$ and $x_c \geq y_c$ with at least one inequality strict.

Definition 2 (Strict pareto efficiency)**.** A subject's preferences $\succsim$ on X are said to satisfy *strict pareto-efficiency* if, for all $x, y \in X_{row}$, $x \succ y$ whenever $x_r > y_r$ and $x_c > y_c$.

Definition 3 (Own payoff maximization)**.** A row (column) subject is said to *maximize his own payoff* if, for all $x, y \in X_{row}$ $(x, y \in X_{column})$, $x \succ y$ whenever $x_r > y_r$ $(x_c > y_c)$.

Definition 4 (Total payoff maximization)**.** A subject is said to *maximize total payoff* if, for all $x, y \in X_{row}$, $x \succ y$ whenever $x_r + x_c > y_r + y_c$.

Table 3.4 shows the fraction of subjects whose preferences are consistent with the properties defined above.

Table 3.4: Measured preferences

Treatment	Pareto efficiency	Strict pareto efficiency	Own payoff max.	Total payoff max.	n
Pooled	70.9%	90.2%	48.6%	4.6%	368
Baseline	71.8%	90.4%	46.8%	4.3%	188
Info	70.0%	90.0%	50.6%	5.0%	180

Preferences that satisfy pareto efficiency or own payoff maximization must also satisfy strict pareto efficiency. The vast majority of subjects report preferences that are consistent with strict pareto efficiency. Table 3.5 further classifies those preferences. Clearly, most preferences that satisfy strict pareto efficiency also satisfy either pareto efficiency or own payoff maximization or both. Only 6.9% of the preferences that satisfy strict pareto

efficiency are not consistent with either pareto efficiency or own payoff maximization (pooled data). Notice also that preferences that satisfy total payoff maximization must simultaneously satisfy pareto efficiency.

Table 3.5: Preferences that satisfy strict pareto efficiency

Treatment	Pareto	Own payoff max.	Pareto and Own payoff max	Only strict pareto	n
Pooled	78.6%	53.9%	39.5%	6.9%	332
Baseline	79.4%	51.8%	37.6%	6.5%	170
Info	77.8%	56.2%	41.4%	7.4%	162

3.3.2 Nash equilibrium play

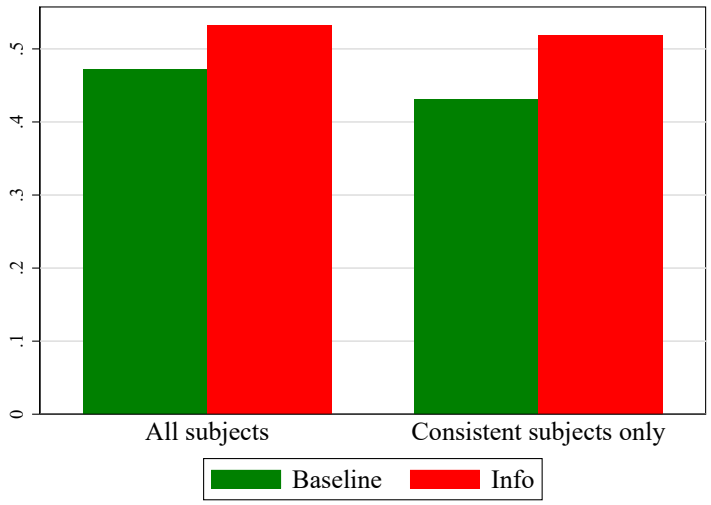
Our first hypothesis is that subject behavior is more consistent with the Nash equilibrium when preferences are mutually known. As outlined in the introduction, playing a maxmin or a maxmax strategy can be a response to strategic uncertainty. Therefore, our second hypothesis is that a strategy is more likely to be played when it is a maxmin and/or a maxmax strategy.

We test the first hypothesis by using two different subsets of our data. Recall that each subject played four games. Since there are a total of 368 subjects who participated in the experiment, we have data on 1472 individual decisions, 752 in treatment baseline and 720 in treatment info. As described earlier, we focus on situations in which one player has a unique pure non-dominant equilibrium strategy. Therefore, we exclude those decisions where both strategies are played with strictly positive probability in some Nash equilibrium, which leaves us with 862 decisions (424 in treatment baseline and 438 in treatment info). We also exclude those decisions where the equilibrium strategy is weakly or strictly dominant. In such a situation, the best response does not depend on the other player's strategy and therefore, it should not matter whether or not the other players'

preferences are known. This leaves us with 279 individual decisions, 140 in treatment baseline and 139 in treatment info.

In all of these 279 games, the subject whose decision we study has a unique pure equilibrium strategy and that subject's opponent has a strictly dominant strategy. We test our main hypothesis using these 279 observations and will refer to the corresponding subset of our data as "all subjects". Figure 3.6 shows that subjects play an equilibrium strategy more often in treatment info than in treatment baseline.

Figure 3.6: Frequencies of played unique equilibrium strategies



To test whether these differences are significant, we run a logit regression. The dependent variable "equilibrium strategy played" assumes a value of 1 if a subject plays the unique equilibrium strategy and 0 otherwise. We include an intercept as well as a dummy variable, which assumes a value of 1 if the observation is generated in treatment info and 0 otherwise. These results are shown in Table 3.6. The treatment effect is significant indicating that informing subjects about their opponents' preferences leads to a higher frequency of equilibrium play.

We run the same test a second time with a smaller subset of our data which no longer includes the decisions made by subjects who played a strictly dominated strategy in at least one of the four games. Either the preferences that these subjects reported in stage 1 do not reflect their true preferences or they are not rational in the sense that their

choice in stage 2 is inconsistent with their reported preferences. Table 3.7 shows that approximately one fourth of our subjects violate strict dominance at least once.

Table 3.6: Logit regression "equilibrium strategy played", robust standard errors clustered by subject

Dependent variable: equilibrium strategy played	All Subjects	Consistent subjects only
info	0.54**	0.60**
	(0.26)	(0.29)
constant	-0.41**	-0.41**
	(0.18)	(0.20)
n	279	226
Clusters	212	166
Pseudo R^2	0.013	0.016

** significant at 5% level

Table 3.7: Violations of strict dominance

Treatment	Subjects	Games played	Games with dominant strategy	Dominated strategy played	Subjects who played dominated strategy at least once
Baseline	188	752	280	23.2%	26.1%
Info	180	720	295	24.4%	29.4%

Similar to the subset "all subjects" we also only use games where the subject has a unique equilibrium strategy that is not dominant. Removing the choices made by inconsistent subjects therefore further reduces the number of observations to 226 individual decisions, 115 in treatment baseline and 111 in treatment info. We will refer to this subset of our data as "consistent subjects only". The treatment effect is comparable when we only use the decisions made by these consistent subjects, even though the number of observations is reduced by approximately 20%.

We also test whether there is a significant treatment effect using a two-tailed two-sample Wilcoxon rank-sum test. The dependent variable is the frequency with which a subject

played an equilibrium strategy. Each subject who plays at least one game where the subject has a unique equilibrium strategy that is not weakly or strictly dominant counts as one observation. We run the same test for all subjects and for consistent subjects only. When using all (only consistent) subjects, we have 107 (87) observations in treatment baseline and 105 (79) in treatment info. The null hypothesis that the distribution of the frequency of equilibrium play is the same in both treatments can be rejected regardless of which data set we use.[48]

Result 1: *Subjects are more likely to play their unique Nash equilibrium strategy when preferences are mutually known.*

As a robustness check, we also compute the frequency of equilibrium play for each game separately. These results are shown in Figure 3.7 for all subjects and in Figure 3.8 for consistent subjects only. Regardless of which subset of our data we use, the frequency of equilibrium play is higher in treatment info than in treatment baseline for every game[49] except for games 5 and 6.

At first glance, subject behavior in Game 5 appears to be surprising: there is less equilibrium play in treatment info than in treatment baseline. A detailed check shows that all subjects, who did not take the equilibrium stategy in treatment info, were column players who played strategy R. The row players had the strictly dominant strategy U. Our second main result in Section 3.3.3 shows that many subjects followed a heuristic approach by selecting maxmax and/or maxmin strategies. Game 5 exhibits a special feature: the equilibrium and maxmax/maxmin strategy especially often fall apart. In several cases the (non-equlibrium) strategy R is both the maxmax and the maxmin strategy. These cases occur considerably more often in treatment info than in treatment baseline: in treatment info, 6 out of 10 subjects who violated the equilibrium prediction faced such a situation, while this is only the case for 1 out of 3 subjects in treatment baseline.

[48] $p = 0.083$ using all subjects, $p = 0.086$ using consistent subjects only.

[49] Using a Fisher exact test, this difference is significant at the 5% level for Game 3, when we use all subjects. We have more observations for Game 3 than for any other game. In Game 3, it occurred particularly often that one subject had a strictly dominant equilibrium strategy while the other subject did not have a strictly or weakly dominant strategy. Details of these tests can be found in the appendix (tables 3.10 and 3.11).

In Game 6 the frequency of equilibrium play is zero in both treatments as Figure 3.7 shows. This is in line with what we have expected: recall that Game 6 is a "Battle of Sexes"-type game-form and we expected that the game that subjects actually play (the induced game) is in most cases a "Battle of Sexes"-type game. Consequently, the situation that one subject has a unique non-dominant equilibrium strategy occurs very rarely here (for consistent subjects, we only have 5 relevant cases in both treatments together). Nonetheless, it makes sense to incorporate this situation in our analysis. First, we intend to provide a comprehensive analysis of the most popular games used in the experimental literature as described in the introduction. Second, we wanted to test whether the outcome meets our expectations.

Figure 3.7: Frequency of equilibrium play by game, all subjects

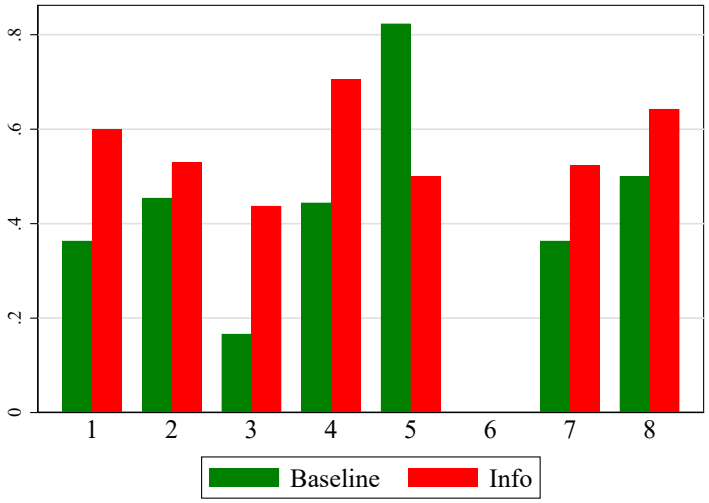
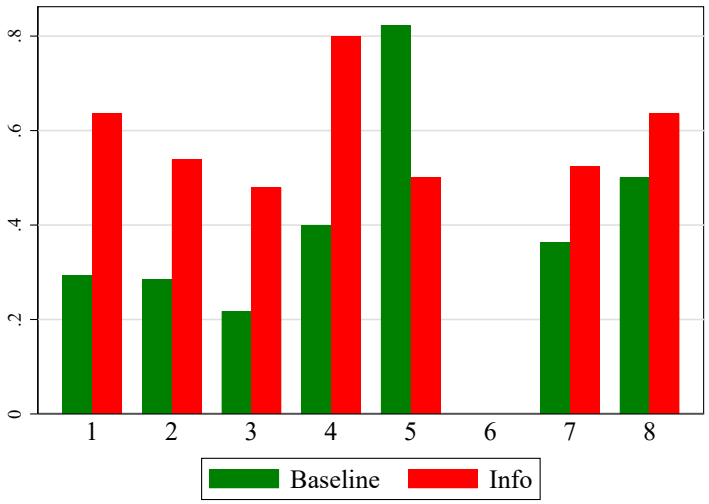


Figure 3.8: Frequency of equilibrium play by game, consistent subjects only



3.3.3 Maxmin and maxmax strategy play

In the following, we discuss our second hypothesis that maxmin and maxmax strategies are more likely to be played. Subjects may use these strategies when they are uncertain about other players' payoff functions and/or other players' rationality. In treatment baseline, subjects face both types of uncertainty, whereas the uncertainty about other players' payoffs is removed in treatment info. Since there is some uncertainty in both treatments, we would expect a strategy to be played more often if it is a maxmin or a maxmax strategy in both treatments. Both effects are expected to be stronger in treatment baseline compared to treatment info.

We test these conjectures by running a conditional logit regression. An observation corresponds to a pure strategy. The dependent variable ("played") assumes a value of 1 if a strategy is played and 0 otherwise. Three independent variables are used to characterize each strategy: "equilibrium" indicates whether a strategy is a Nash equilibrium strategy. "maxmax" assumes a value of 1 if a strategy can lead to a most highly ranked payment pair. "maxmin" indicates whether a strategy can result in the realization of a lowest ranked payment pair (maxmin = 0 if that is the case, maxmin = 1 otherwise).

We only use decisions made by subjects who never played a strictly dominated strategy. Table 3.8 shows that whether or not a strategy is a Nash equilibrium strategy

only matters in treatment info when predicting which strategies subjects will play. In contrast, the coefficients of maxmax and maxmin are highly significant in both treatments. While the three independent variables ("equilibrium", "maxmax" and "maxmin") are correlated, all pairwise correlation coefficients are lower than 0.5. Further details on the relationship of the three independent variables can be found in the appendix.

Table 3.8: Conditional logit regression "played", robust standard errors clustered by subject

Dependent variable: played	Baseline	Info
equilibrium	0.09	0.89****
	(0.21)	(0.22)
maxmax	1.61****	1.17****
	(0.18)	(0.15)
maxmin	1.39****	1.29 ****
	(0.17)	(0.15)
n	1112	1016
Clusters	139	127
Pseudo R^2	0.40	0.41

**** significant at 0.1% level

Result 2: *In both treatments, a strategy is more likely to be played when it cannot lead to the lowest ranked payment pair and when it can lead to the highest ranked payment pair.*

In line with Result 1, the coefficient estimate for the variable "equilibrium" differs significantly among the two treatments and is only useful to predict play in treatment info but not in treatment baseline. In contrast, the highest and lowest ranked payment pair seems to attract our subjects' attention in both treatments. As expected, the according coefficient estimates are higher in treatment baseline than in treatment info. However, the difference is not signficant.[50]

[50]The coefficient estimate of an interaction term of maxmin and the treatment dummy (maxmax and the treatment dummy) is not significant.

3.3.4 Did we manage to elicit subjects' true preferences?

When preferences are elicited in stage 1 of the experiment, subjects in treatment info are aware that these preferences will be revealed to other subjects. However, they are not informed about the specific games that are played in stage 2. Hence, subjects did not have the information necessary to figure out what kind of misrepresentation might be most advantageous: in some 2×2 games, it could be beneficial to be perceived as having social preferences whereas in other games, the contrary is more likely (e.g., in the chicken game). Moreover, recall that a decision made in stage 2 affects a subject's payment with a probability of only 1/8. All in all, in treatment baseline, subjects had clear and strong incentives to truthfully report their preferences. It was clear to subjects that their reports would not be revealed to anyone but the experimenter. Therefore, it is not plausible that a rational subject would misrepresent his preferences.

We test the claim that subjects truthfully reported their preferences in stage 1 of treatment info by using the frequency with which subjects play strictly dominated strategies in stage 2 of the experiment. To identify strategies that are strictly dominated, we use the preferences elicited in stage 1. If these reflect a subject's true preferences, a rational subject should never play such a strictly dominated strategy. In contrast, if subjects strategically misrepresent their preferences in stage 1, a strategy that we classify as strictly dominated may in fact not be dominated according to the subjects' true preferences.

Since preferences in treatment baseline are not revealed to other subjects, it is clear that subjects in treatment baseline have no reason to misrepresent their preferences. Therefore, we can compare the frequency with which subjects play a strictly dominated strategy in the two treatments to test the claim that preferences are truthfully revealed in stage 1 of treatment info. If that claim is true, no difference should be observed. Otherwise, subjects should be more likely to play a strictly dominated strategy in treatment info than in baseline.

Table 3.7 shows how often subjects play a strictly dominated strategy using the preferences reported in stage 1 to define the according games. Each subject played 4 games, thus resulting in 752 games played in treatment baseline and 720 in info. In 280 of these games

in treatment baseline and 295 in info, one of the strategies was strictly dominated. In roughly a quarter of these cases, the strictly dominated strategy was played.

In order to check the assumption that subjects do not misrepresent their preferences in both treatments, we run a logit regression using the 280 games in treatment baseline as well as the 295 games in treatment info as observations. The dependent variable "dominated strategy played" is a dummy variable that assumes a value of 1 if the strictly dominated strategy was played. The only explanatory variable other than the intercept is a treatment dummy ("info") (see Table 3.9).

Table 3.9: Logit regression "dominated strategy played", robust standard errors clustered by subject

Dependent variable: dominated strategy played	
Info	0.07
	(0.23)
Constant	-1.20***
	(0.16)
n	575
Clusters	333
Pseudo R^2	0.0002

***significant at 1% level

The coefficient estimate for the treatment dummy is not significantly different from 0. Hence, the null hypothesis cannot be rejected. We also test the same assumption using a two-tailed two-sample Wilcoxon rank-sum test. The dependent variable is then the frequency with which a subject plays a dominated strategy. Each subject who had a strictly dominant strategy in at least one of the four games corresponds to an observation. There are 165 such observations in treatment baseline and 168 in treatment info. We cannot reject the null hypothesis that the frequency with which strictly dominated strategies are played follows the same distribution in the two treatments ($p = 0.81$).

Result 3: *Subjects are equally likely to play a strictly dominated strategy in both treatments.*

Moreover, the fraction of subjects whose reported preferences are consistent with own payoff maximization is even slightly larger in treatment info compared to treatment baseline, though the difference is not significant ($p = 0.53$ using a Fisher exact test). All other properties of measured preferences that we discussed in section 3.3.1 are also satisfied equally frequently in both treatments (see tables 3.4 and 3.5). We therefore maintain the assumption that subjects truthfully report their preferences in stage 1 of the experiment in both treatments.

In psychological game theory, Rabin (1993) and Dufwenberg and Kirchsteiger (2004) introduced models of reciprocity in which players reward kind actions and punish unkind ones. Reciprocity could lead to a problem equivalent to the misrepresentation of preferences discussed in this section. For instance, consider Game 1 in stage 2 of treatment info. Suppose an own-payoff maximizer (row) is matched with a total-payoff maximizer (column). The row player might then believe that column will cooperate (play R), even though column expects row to defect (play U). This expected kindness on the part of column might then induce row to also cooperate, thus violating our assumption that only outcomes matter. In other words, subjects' preferences might change once they are shown their opponents' ranking of payment pairs in stage 2 of treatment info. Another potential violation of our assumption might arise if subjects' preferences over payment pairs changed once they are shown the specific game-form.

If there were many preference reversals of the reciprocity type, we would expect to observe some differences between treatment info and baseline since subjects' preferences are only revealed in treatment info. In fact, we found no evidence that would support this claim. For instance, Result 3 indicates that there is no significant difference concerning the play of strictly dominated strategies. This indicates that reciprocity effects probably do not matter much in our experiment. If they did, one would expect to observe that subjects play strictly dominated strategies more often in treatment info compared to treatment baseline (e.g., because some subjects, who reported selfish preferences, would reward opponents with social preferences). We cannot exclude that there were some preference reversals arising from the strategic situation in stage 2 (as compared to the ranking of

payment pairs in stage 1). However, since there is a treatment effect, this would be only a problem if such preference reversals led to a systematic upward bias (indicating a false-positive result). We checked our data and found no evidence indicating such a bias. This is plausible since subjects face the same game-forms in both treatments. Hence, it is unlikely that there were very different preference reversals in the treatments that caused an upward bias. After all, we found a significant treatment effect despite many factors that potentially caused noise in the experiment. In our view, this makes our result more robust and indicates that the "true" treatment effect might be even stronger.

3.4 Conclusion

The assumption that players' preferences are mutually known is often not satisfied in the laboratory. It seems plausible that similar difficulties exist in many real-world situations as well. Our experiment shows that it is a relevant assumption: making sure that preferences are mutually known leads to significantly more equilibrium play.

When deciding what model to apply to a specific situation, whether or not agents can reasonably be expected to know other agents' payoff functions should therefore play an important role. At least in the games we analyzed, subjects are unlikely to play a Nash equilibrium strategy when payoff functions are not mutually known. Many other models that are used in behavioral game theory (e.g., level-k models) also rely on the assumption of mutual knowledge of preferences. These models might also fail to accurately predict behavior whenever preferences are not mutually known.

Our results also show that subjects are more likely to play maxmin or maxmax strategies rather than the Nash equilibrium strategy. Hence, many subjects seem to rely on heuristics rather than on strategic considerations when selecting a strategy.

Acknowledgements. We thank Gary Charness, Jürgen Eichberger, Jörg Oechssler, Christoph Vanberg, the seminar participants at University of Exeter and University of Heidelberg, and the conference participants at ESA European Meeting 2014 and ESA World Meeting 2017 for very helpful comments.

Appendix

A.1 Details of the robustness check tests for the main result

Tables 3.10 and 3.11 report the results of a two-tailed Fisher exact test of the null hypothesis that the probability that a subject plays the equilibrium strategy is the same in both treatments. These tests were run separately for each of the 4 games. n_base is the number of observations in treatment baseline and n_info the number of observations in treatment info. The tests reported in Table 3.10 include all subjects while those reported in Table 3.11 include consistent subjects only.

Table 3.10: Fisher exact test (two-tailed), all subjects.

Game	n_base	n_info	p-value
1	22	15	0.193
2	11	17	1.000
3	30	32	0.028
4	18	17	0.176
5	17	20	0.082
6	4	3	n.A.
7	22	21	0.364
8	16	14	0.484

Table 3.11: Fisher exact test (two-tailed), consistent subjects only.

Game	n_base	n_info	p-value
1	17	11	0.121
2	7	13	0.374
3	23	25	0.075
4	10	10	0.170
5	17	18	0.075
6	3	2	n.A.
7	22	21	0.364
8	16	11	0.696

A.2 Details on the conditional logit regression (table 8)

Table 3.12 provides additional information for both regressions that are displayed in table 3.8: We check for each pure strategy available to consistent subjects whether it is a Nash equlibrium strategy ("equilibrium" = 1), whether it is the maxmax strategy ("maxmax" = 1) and whether it is the maxmin strategy ("maxmin" = 1). "n_baseline" indicates the number of pure strategies in treatment baseline, "n_info" the number of strategies in treatment info.

Table 3.12: Properties of strategies available to consistent subjects, by treatment

equilibrium	maxmax	maxmin	n_baseline	n_info
0	0	0	332	322
0	1	0	236	199
1	1	1	187	189
0	0	1	182	143
0	1	1	61	56
1	1	0	56	44
1	0	1	34	37
1	0	0	24	26

A.3 Instructions

Treatment Baseline: Instructions Part 1

1 General Information

Welcome to this experiment and thank you very much for your participation! Please switch off your mobile phone now and do not communicate with each other any more. If you have a question, raise your hand, we will come over to your seat and answer it individually. In this experiment, you can earn a substantial amount of money. The amount you earn depends on your own decisions, the decisions of the other participants and on chance. The amount of money earned will be paid out to all participants individually in cash at the end of the experiment. During the experiment, everyone makes his decisions anonymously on his own. At no point in time will your decisions be linked to your identity.

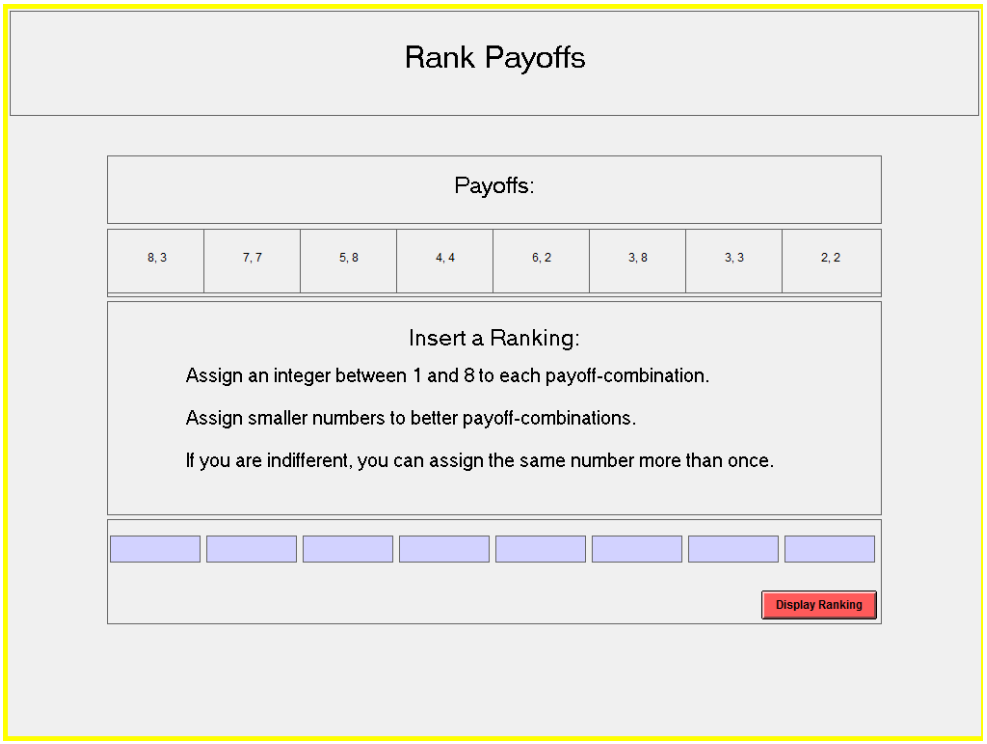
This experiment consists of two parts, which are identical for all participants: In the first part you are shown eight different payoff-combinations, which you are supposed to evaluate. Each of these combinations consists of two numbers (x, y). The first number x corresponds to the amount of Euro that you receive yourself in this situation. The second number y corresponds to the amount that another participant receives. You are supposed to establish a ranking (a so called "preference relationship") over all these payoff-combinations (x, y). That means, you indicate which of these combinations you like best, which one second-best, and so on. The exact procedure will be explained again step by step later on.

The ranking created in this way, as well your decisions in part two of the experiment, will not be revealed to any other participant. After each participant has created such a ranking over the payoff-combinations, part two of the experiment will begin. Both parts of the experiment are run at the computer. Before they start, you are asked several control questions, which shall help you in your understanding of the experiment. For the second part, you will receive separate instructions. At the end of the experiment, there will be a short questionnaire and then you will be paid in cash.

Your total payoff consists of two payments. In order to determine these payments, one of the decisions made in either part 1 or part 2 of the experiment will be randomly selected. Further details will be provided later on.

2 Evaluation of Payoff-combinations

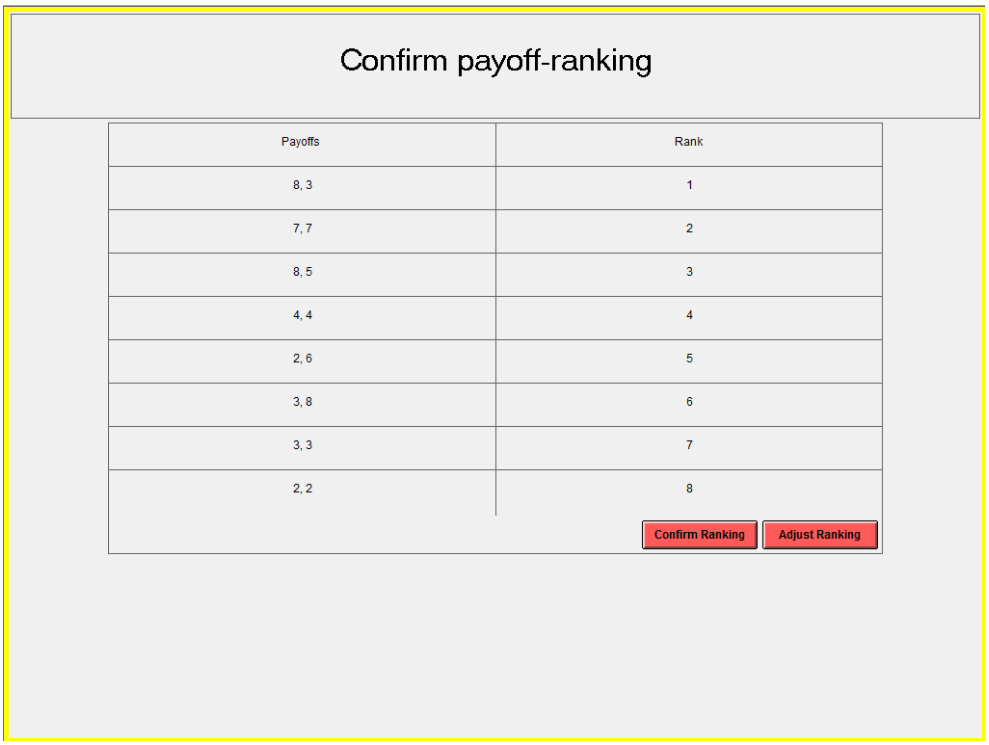
We will now explain the first part of the experiment, the evaluation of payoff-combinations. You will perform this task immediately afterwards at the computer. You will first be shown the following screen:



In the row below "Payoffs" you see the eight different payoff-combinations (x, y), which you are supposed to rank (all amounts are in Euro). The payoff combinations are currently ordered randomly. *(Remember: The left value x is the amount you receive yourself and the right value y is given to a randomly selected other participant.)*

You will now assign a number between 1 and 8 to each of these payoff-combinations. The number 1 corresponds to the first rank, which you shall assign to the combination you like best. Analogously the second rank shall be assigned to your second-best combination and so forth until rank 8, which corresponds to your least preferred combination. If you consider two or more combinations as equally good, you are allowed to assign the same rank/number to them.

After you created your ranking, you will see the following screen:



Here you see the payoff-combinations, ordered according to your previously stated preferences. If you like, you can still make modifications. After all participants confirmed their ranking, the second part of the experiment will begin.

3 Calculation of your Final Payoff

The one and only payoff-relevant decision will be randomly selected at the end of the experiment. Your total payoff depends on whether a decision from the first or the second part of the experiment is selected.

With a probability of $\frac{7}{8}$ a decision of part one will be chosen. In this case, two of the eight payoff-combinations will be randomly selected. The payoff combination that you ranked more highly will then be paid out. (If both combinations have the same rank, one of these two will be randomly selected.). You will receive the first amount, the value x. In addition, every participant receives exactly one additional payment y that corresponds to the second amount y of a payment-combination selected for some other participant. *(The assignment is carried out in such a way that the second amount y from your decision is not distributed to a participant you are interacting with during the experiment or from whom you receive the second amount yourself.)*

Payoff, if selected decision is from part one:

Total payoff = Amount x from own decision + Amount y from decision of some other participant

The probability that a decision from part two is chosen for payment is $\frac{1}{8}$. In that case, payments depend on the actions chosen by the participants in part two. The calculation of the final payoff for this case will be explained in the instructions for this part. *(The random draw will be performed by a participant at the end of the experiment. For that purpose he draws a card from a deck containing 32 cards numbered 1 to 32. The numbers 1-28 correspond to all possible combinations of two out of the eight payoff-pairs (x, y)) from the first part. If a number between 29-32 is drawn, a decision from the second part will be paid out.)*

Treatment Baseline: Instructions Part 2

The second part of the experiment is run at the computer as well. This part consists of four strategic decision situations, in the following referred to as "games". In each of these situations, you will be matched with a different participant as game partner, that means you never interact with the same person twice. You and the other player simultaneously select one of two possible actions. The row player always chooses between one of the two actions "up" and "down" and the column player always decides between the actions "left" and "right". *(For the sake of simplicity, the game will be displayed for every participant in such a way, that he always acts in the role as row player and the game partner in the role as column player.)*

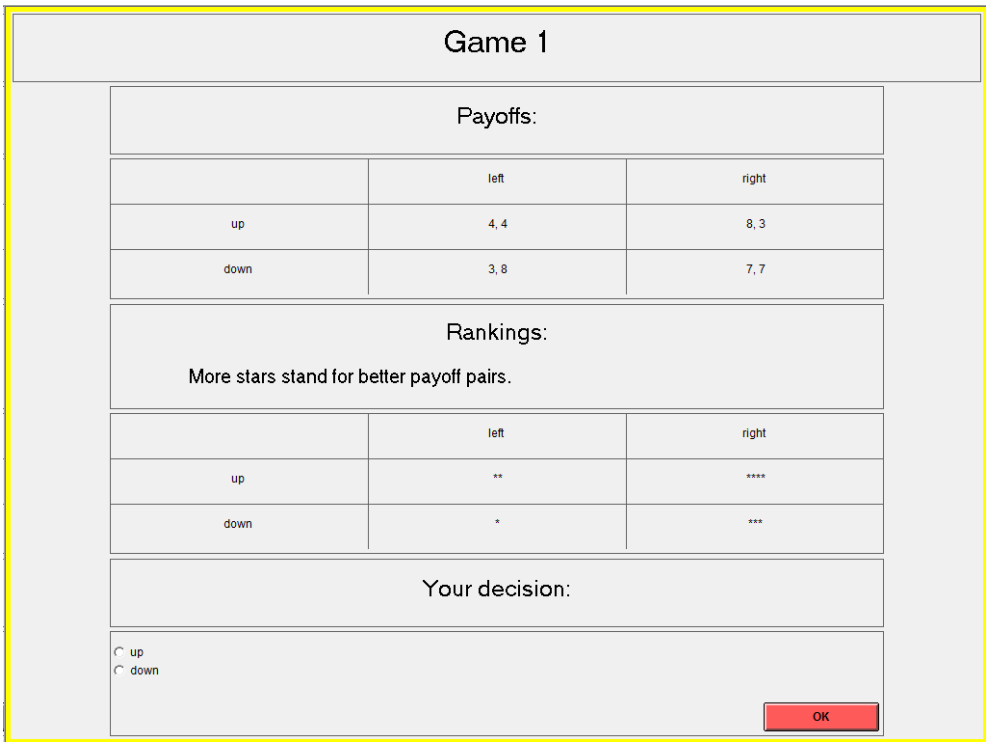
In every game, there are four possible outcomes. Which one of these outcomes is selected depends on the action you chose as well as on the action the other player chooses. The four outcomes are are displayed in the form of a payoff matrix. The combination (x, y) in one cell of the matrix corresponds to the amounts of money the two players receive, if the corresponding actions have been chosen. Analogously to the first part, the left value x indicates the amount of money in Euro that you receive and the right value y corresponds to the payoff of the other player. **The combinations (x, y) are chosen in such a way, that they assume the exact same values as those from the first part of the experiment.** Thus in every game there appear four out of the eight payoff pairs evaluated in part one.

If a situation from the second part is chosen for payment, the involved players receive the payoffs that correspond to the outcome of the game. In contrast to the first part, each player only receives one amount of money from the payoff-relevant decision. In addition, each player is given a fixed payment of 5 Euro.

Total payoff = 5 Euro + Payment x obtained in the selected game

In addition to the monetary payments, you are also shown the ranking of the payoff-pairs used in the current game that you submitted in the first part of the experiment.

In the computer program, you will see the following screen:



For the sake of clarity, not the exact numbers of the ranking will be shown there, but instead 1-4 stars. A value of four stars (****) means that the corresponding payoff-combination was ranked by you as the best combination (among those appearing in the game). Accordingly, the worst combination is marked by one star (*)

Example:
Let us consider the game shown on the screen "Game 1". If, for example, you decide to play "up" and the other player chooses "right", then you receive a payoff of 8 Euros and your game partner a payoff of 3 Euros. Additionally, you can see in the matrix below, that this is your most preferred outcome.

Are there any questions?

If this is not the case, the second part of the experiment will start shortly...

Treatment Info: Instructions Part 1

1. General Information

Welcome to this experiment and thank you very much for your participation! Please switch off your mobile phone now and do not communicate with each other any more. If you have a question, raise your hand, we will come over to your seat and answer it individually. In this experiment, you can earn a substantial amount of money. The amount you earn depends on your own decisions, the decisions of the other participants and on chance. The amount of money earned will be paid out to all participants individually in cash at the end of the experiment.

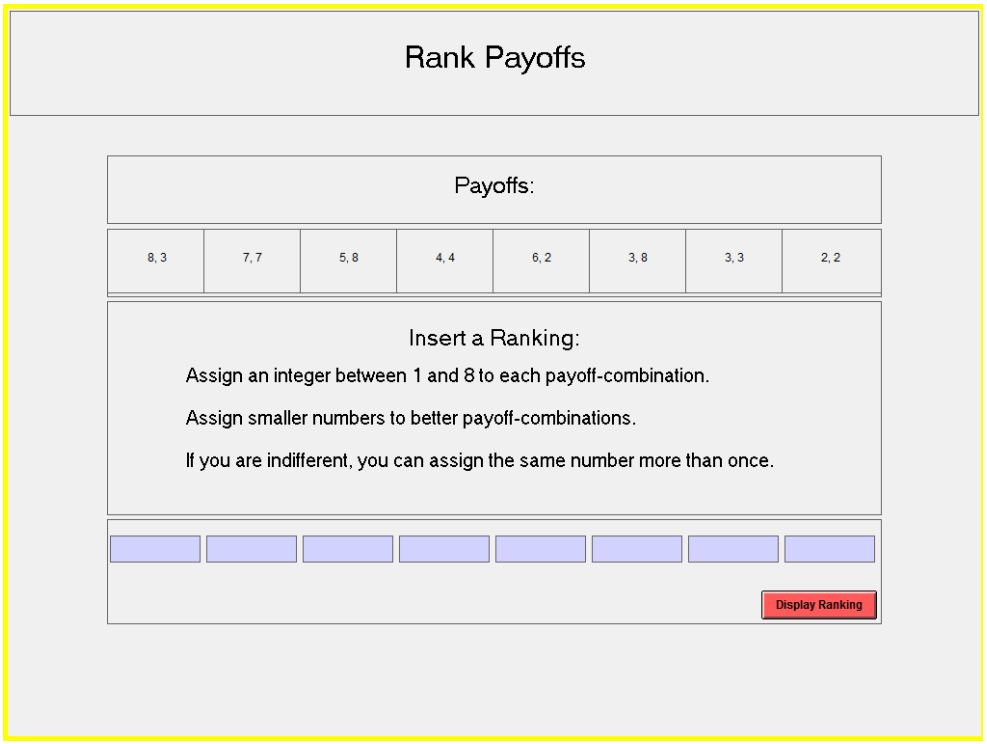
This experiment consists of two parts, which are identical for all participants: In the first part you are shown eight different payoff-combinations, which you are supposed to evaluate. Each of these combinations consists of two numbers (x, y). The first number x corresponds to the amount of Euro that you receive yourself in this situation. The second number y corresponds to the amount that another participant receives. You are supposed to establish a ranking (a so called "preference relationship") over all these payoff-combinations (x, y). That means, you indicate which of these combinations you like best, which one second-best, and so on. The exact procedure will be explained again step by step later on.

After each participant has created such a ranking over the payoff-combinations, part two of the experiment will begin. In this part, the information provided in the first part of the experiment will be used. Two participants at a time will be shown each others' ranking of the payoff-pairs provided in part one of the experiment. In both parts of the experiment, you will interact with other participants using a computer. Before we start, you will be asked several control questions, which shall help you in your understanding of the experiment. For the second part, you will receive separate instructions. At the end of the experiment, there will be a short questionnaire and then you will be paid in cash.

Your total payoff consists of two payments. In order to determine these payments, one of the decisions made in either part 1 or part 2 of the experiment will be randomly selected. Further details will be provided later on.

2. Evaluation of Payoff-combinations

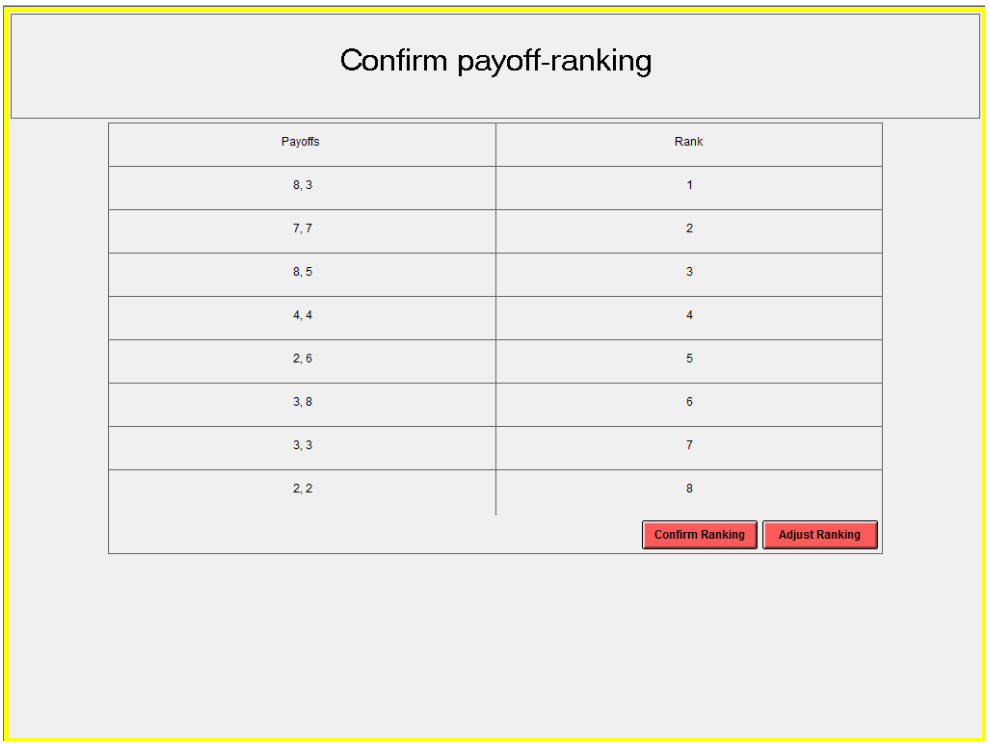
We will now explain the first part of the experiment, the evaluation of payoff-combinations. You will perform this task immediately afterwards at the computer. You will first be shown the following screen:



In the row below "Payoffs" you see the eight different payoff-combinations (x, y), which you are supposed to rank (all amounts are in Euro). The payoff combinations are currently ordered randomly. *(Remember: The left value x is the amount you receive yourself and the right value y is given to a randomly selected other participant.)*

You will now assign a number between 1 and 8 to each of these payoff-combinations. The number 1 corresponds to the first rank, which you shall assign to the combination you like best. Analogously the second rank shall be assigned to your second-best combination and so forth until rank 8, which corresponds to your least preferred combination. If you consider two or more combinations as equally good, you are allowed to assign the same rank/number to them.

After you created your ranking, you will see the following screen:



Here you see the payoff-combinations, ordered according to your previously stated preferences. If you like, you can still make modifications. After all participants confirmed their ranking, the second part of the experiment will begin.

3. Calculation of your Final Payoff

The one and only payoff-relevant decision will be randomly selected at the end of the experiment. Your total payoff depends on whether a decision from the first or the second part of the experiment is selected.

With a probability of $\frac{7}{8}$ a decision of part one will be chosen. In this case, two of the eight payoff-combinations will be randomly selected. The payoff combination that you ranked more highly will then be paid out. (If both combinations have the same rank, one of these two will be randomly selected.). You will receive the first amount, the value x. In addition, every participant receives exactly one additional payment y that corresponds to the second amount y of a payment-combination selected for some other participant. *(The assignment is carried out in such a way that the second amount y from your decision is not distributed to a participant you are interacting with during the experiment or from whom you receive the second amount yourself.)*

Payoff, if selected decision is from part one:

Total payoff = Amount x from own decision + Amount y from decision of some other participant

The probability that a decision from part two is chosen for payment is $\frac{1}{8}$. In that case, payments depend on the actions chosen by the participants in part two. The calculation of the final payoff for this case will be explained in the instructions for this part. *(The random draw will be performed by a participant at the end of the experiment. For that purpose he draws a card from a deck containing 32 cards numbered 1 to 32. The numbers 1-28 correspond to all possible combinations of two out of the eight payoff-pairs (x, y)) from the first part. If a number between 29-32 is drawn, a decision from the second part will be paid out.)*

Treatment Info: Instructions Part 2

The second part of the experiment is run at the computer as well. This part consists of four strategic decision situations, in the following referred to as "games". In each of these situations, you will be matched with a different participant as game partner, that means you never interact with the same person twice. You and the other player simultaneously select one of two possible actions. The row player always chooses between one of the two actions "up" and "down" and the column player always decides between the actions "left" and "right". *(For the sake of simplicity, the game will be displayed for every participant in such a way, that he always acts in the role as row player and the game partner in the role as column player.)*

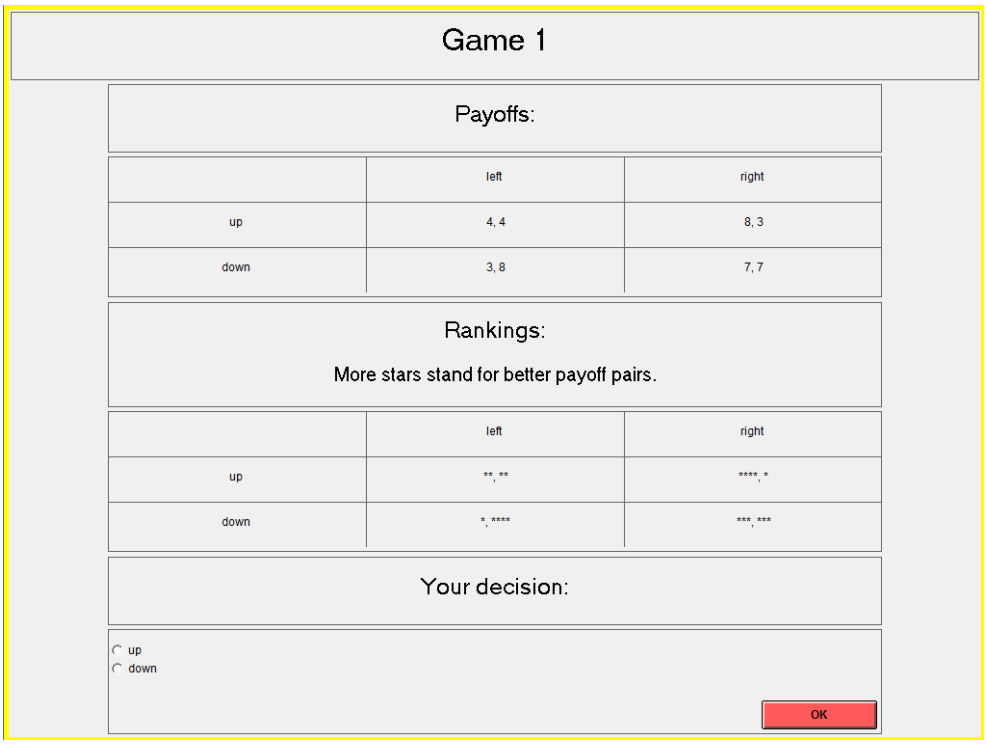
In every game, there are four possible outcomes. Which one of these outcomes is selected depends on the action you chose as well as on the action the other player chooses. The four outcomes are are displayed in the form of a payoff matrix. The combination (x, y) in one cell of the matrix corresponds to the amounts of money the two players receive, if the corresponding actions have been chosen. Analogously to the first part, the left value x indicates the amount of money in Euro that you receive and the right value y corresponds to the payoff of the other player. **The combinations (x, y) are chosen in such a way, that they assume the exact same values as those from the first part of the experiment.** Thus in every game there appear four out of the eight payoff pairs evaluated in part one.

If a situation from the second part is chosen for payment, the involved players receive the payoffs that correspond to the outcome of the game. In contrast to the first part, each player only receives one amount of money from the payoff-relevant decision. In addition, each player is given a fixed payment of 5 Euro.

Total payoff = 5 Euro + Payment x obtained in the selected game

As announced before, you will now receive information about each others' preferences. This means that in addition to the payoff matrix you are shown another matrix below, in which you can see how you and the other player ranked the payoff combinations used in the current game in the first part of the experiment. *Note: you interact with a different partner in every game and therefore the ranking of your opponent may change from one game to another.*

In the computer program, you will see the following screen:



For the sake of clarity, not the exact numbers of the ranking will be shown there, but instead 1-4 stars. A value of four stars (****) means that the corresponding payoff-combination was ranked by you as the best combination (among those appearing in the game). Accordingly, the worst combination is marked by one star (*)

Example:
Let us consider the game shown on the screen "Game 1". If, for example, you decide to play "up" and the other player chooses "right", then you receive a payoff of 8 Euros and your game partner a payoff of 3 Euros. Additionally, you can see in the matrix below, that this is your most preferred outcome, but the least preferred outcome of the other player.

Are there any questions?

If this is not the case, the second part of the experiment will start shortly...

Chapter 4

Social Capital

Inequality, Fairness and Social Capital[†]

Abstract We study the impact of unjust inequality on social trust and trustworthiness, and how it interacts with economic status in a large-scale controlled experiment. We document that unfair economic inequality is detrimental for social interactions, resulting in a significant decline in trust and trustworthiness. Probing the boundaries of this effect, we demonstrate that this erosion of social capital critically depends on the context: if an economically successful person is not directly responsible for the outcome of the unsuccessful per-son, we observe no negative effects on trust and trustworthiness in the aggregate. Finally, our data do not support the view that higher status or wealth leads to an erosion of pro-social attitudes: the successful are always more generous, whereas unsuccessful persons display the least efficient and generous behavior.

[†]Joint work with Dietmar Fehr, Stefan T. Trautmann and Yilong Xu

4.1 Introduction

The recent surge of income and wealth inequality in many developed countries is a widely discussed topic in the media and academic research. Much of these discussions revolve around the gains of the top-income decile and the stagnation of income for the bottom half of the distribution and its implications for society (e.g.,Piketty & Saez, 2003; Autor et al., 2008; Piketty & Saez, 2014; Piketty et al., 2017, Alvaredo et al., 2017). Indeed, inequality deriving from competitive economic environments is often associated with negative societal consequences (Stiglitz, 2012; Verhaeghe, 2014). In particular, it is sometimes conjectured that inequality may harm the social fabric, destroying social capital (trust, honesty, cooperation) and subsequently affecting economic outcomes (Wilkinson et al., 2009). Two hypotheses can be derived from the literature in economics and the social sciences. The first hypothesis states that higher inequality, especially if perceived as unjust and caused by competition, hampers economic interaction (Alesina & Perotti, 1996; Benabou, 1996; Camera et al., 2016). The second hypothesis states that those who are in an advantageous position (of higher status or wealth) in an unequal society, become self-focused and greedy (Piff et al., 2012; Piff, 2013; Fisman et al., 2015; Guinote et al., 2015; Nishi et al., 2015). That is, negative social consequences are caused by the behavior of the successful.

Both of these hypotheses are contested in the literature. However, empirical assessments of the effects of inequality and the role of the successful often suffer from an absence of counterfactuals and the endogeneity of status. Experimental methods offer an alternative approach for assessing the consequences of inequality as they make exogenous variation of inequality, the underlying causes of inequality, institutions and available information possible (e.g., Falk & Heckman, 2009; Charness & Fehr, 2015). While potentially having lower external validity, experiments thus provide a clear identification of causal effects and underlying processes.

We use experimental methods to study the impact of unjust inequality on subsequent social interactions, differentiating between the behavior of the economically successful and the unsuccessful. Our design thus aims to test both hypotheses within the same

setting. We create income inequality in dyads, using a real-effort procedure with varying payment schemes. Subsequently, we let these dyads interact in a modified trust game allowing us to measure both players' social trust and trustworthiness. Social trust has been interpreted as an important component of social capital in the literature (Glaeser et al., 2000; Bellemare & Kröger, 2007; Bjoernskov, 2018; Langer et al., 2017). As higher social capital is typically associated with better-functioning institutions and society in general (Putman, 2000) , social trust is a center piece in the debate on whether inequality erodes the social fabric.[51] In addition, our experimental measure for trustworthiness allows us to quantify subjects' greed or altruism absent strategic motives. It directly tests the hypothesis that higher inequality has a negative impact on social interactions because successful people become less generous, in particular less generous than the unsuccessful.

We create exogenous variation in income inequality in the real-effort task by randomly assigning subjects to two different payment schemes. In our baseline condition subjects receive a piece-rate payment. This results in relatively low inequality and is typically not perceived as unjust. We compare the trust-game outcomes in this setting with an environment that features unjust and high inequality. To generate high inequality, we implement a relative-payment scheme that gives an undue advantage to one participant in the dyad.[52] This undermines equality of opportunity and the payment scheme can thus be seen as unfair from a normative perspective (e.g., Roemer, 2008). In a third condition, we employ the same relative-payment scheme to generate unjust inequality as before, but randomly rematch participants in the trust-game stage (keeping earnings information constant across conditions). This eliminates the direct responsibility for each other's outcomes in the dyads and has the advantage of observing matches with equal

[51] More precisely, social capital can be defined as values and shared beliefs that help groups to cooperate in situations where contracts are difficult or impossible to enforce (cp., Guiso et al., 2004). According to this definition it is possible to measure social capital by eliciting values and beliefs with experimental tools such as the trust game (see e.g., E. Fehr, 2009 for an extensive account of the measurement of trust and trust beliefs). In the economic literature social capital has been positively associated with a plethora of economic outcomes, such as economic growth (e.g., Knack & Keefer, 1997), the size of firms (e.g., Bloom et al., 2012) or financial development (e.g., Guiso et al., 2004).

[52] There is evidence documenting that (high) inequality is not per se seen as unfair (e.g., Breza et al., 2017; D. Fehr, 2018). For example, D. Fehr (2018) illustrates that an increase in inequality leads to more antisocial behavior but only if higher inequality cannot be clearly attributed to work effort and is possibly the result of immoral behavior.

and unequal outcomes.

Our results support the view that unjust inequality can negatively affect social interactions. That is, we document a significant decline in trust and trustworthiness when income inequality is the result of an income-generating process that is eminently perceived as unfair. However, we also find that this observed decline depends on a direct interaction in the first stage, i.e., when the well-off ("successful") player causing the poor out-come of the worse-off ("unsuccessful") player. If we take away the direct interaction by re-matching participants in the trust game, we find that especially the successful players maintain a high level of trust and trustworthiness, in particular when interacting among themselves. That is, the detrimental impact of inequality on social interactions critically depends on contextual factors.

We do not find evidence that the advantageous social position makes people more selfish: successful players are consistently more generous than the unsuccessful in absolute terms. However, holding the successful accountable to higher normative standards (such as sharing the trust game pie equally), or evaluating generosity in terms of giving relative to someone's wealth position, we may well argue that they fall short on these standards.

In the next section, we introduce the experimental paradigm and design of our study followed by a description of how we induce unjust inequality. Section 3 shows that our experimental paradigm successfully induces inequality differences and a polarization of fairness perceptions. Clearly, neither inequality nor competitiveness have to be perceived negatively per se (e.g. Cappelen et al., 2013; Cappelen et al., 2014; Bartling et al., 2017). Rather, it is the combination of inequality and unequal opportunity within a competitive environment that triggers strong feelings of injustice in our experimental setup. While pooling these features hides their marginal contribution to the perception of unfairness, it guarantees a powerful prime to reliably quantify the effects of unjust inequality on social interactions and, arguably, mirrors many settings outside the laboratory.[53] Competition, unequal opportunities, and inequality are, for example, inherent features of school edu-

[53]Note that our paradigm can be extended to identify the marginal impact of the different features of the environment. However, the effects may not be additive making it impossible to disentangle them. See section 4.2.1 for a more thorough discussion of this issue.

cation,universities, workplaces or labor markets more generally.[54] Section 4 discusses the effects of unjust inequality in fixed dyads and Section 5 discusses the effects when direct attributions of responsibility for others' outcomes cannot be made. We evaluate these results in the context of the related literature in section 6.

4.2 Experimental Paradigm and Design

The current study employs an experimental paradigm in which dyads of participants interact in two stages. In the first stage, a repeated real-effort task involves either an individual piece-rate payment, or a competitive tournament with a favorable condition for the initial tournament winner (in a between-subjects design). While the piece-rate condition leads to modest inequality depending on individual performance, the tournaments amplify income differences in a way that is difficult to justify by the observed performance differences. In the second stage these same dyads then interact in a trust game. Consequently, we observe trust and trustworthiness depending on stage-1 conditions, and depending on stage-1 income. In a third treatment, the tournament-based real-effort stage is followed by a trust-game stage involving new matches of dyads, which have, however, exactly the same degree of information on each other's earnings as dyads in the fixed-pair tournament condition.

In the following, we first describe the stage-1 income manipulation, and the elicitation of fairness judgments. We then provide details on the trust game stage with fixed dyads, and new dyads. Our three treatments are called *Piece Rate* (first-stage piece rate - fixed dyads); *Tournament* (first-stage tournament - fixed dyads); and *Tournament-New* (first-stage tournament - new partner in stage 2).

[54]For example Lemieux et al. (2009) document an economy-wide increase of performance-pay jobs in the U.S. labor market, along with a substantial increase in wage inequality. Features of competitive environments are innately linked to relative status concerns or relative-income comparisons, and it is long known that individuals care about their standing relative to others (e.g., Veblen, 2017). Several recent experimental studies suggest that such comparisons have, for example, detrimental effects on well-being (Card et al., 2012) or ethical behavior (e.g., Gill et al., 2013; John et al., 2014).

4.2.1 Stage 1: Inequality Manipulation

We implement a repeated real-effort slider task (Gill & Prowse, 2012) and vary the payment scheme to manipulate inequality, i.e., low inequality versus high and potentially unjust inequality. In the slider task, participants see a number of sliders on their computer screen and have to adjust each slider to exactly the middle position within a certain time limit (see Figure A.1 in the Appendix). The goal in this task is to maximize the number of correctly positioned sliders before the allotted time runs out. Participants are only allowed to use their mouse to drag the sliders into the correct position.[55] The task requires little apriori knowledge and skills such that outcomes mainly depend on the expended effort of subjects. Unfairness or concerns about unequal opportunities arise only through institutional features, i.e., the details of the implemented payment scheme.

In the low-inequality condition (*Piece Rate*), participants complete four rounds of this task, each lasting for 120 seconds. In each round, they receive a flat payment of €0.50 plus €0.05 per correctly placed slider. Total earnings are calculated by summing up the earnings in the four rounds. Note that each subject in a dyad individually determines her own earnings, i.e., there is no interaction. However, at the end of each round both subjects in the dyad are informed about the correctly positioned sliders and the resulting earnings of each other. Thus, social comparison is also salient in this setting.

In the high-inequality conditions (*Tournament* and *Tournament-New*), participants in a dyad also complete the slider task four times. In contrast to the *Piece Rate* condition, participants' payoffs in each round are determined through a relative performance scheme. That is, the subject with the higher number of correctly placed sliders in a round receives €3.00, while the subject with the lower number of correctly placed sliders receives €0.30. In the case of equal performance, the two payments are randomly allocated. As in *Piece Rate*, participants receive information on the performance of each subject and the resulting payoffs after each round.

In addition to the high payoff, the subject with the higher performance receives a time bonus. More specifically, after an initial time budget of 120 second for both subjects, the

[55] To avoid cheating, we used a keyboard locker to prevent students from using the arrow keys or the mouse wheel.

winner of the first round obtains a time bonus of 8 seconds, and the winners of the second and third round get a time bonus of 6 and 4 seconds, respectively. The time bonus is subtracted from the time budget of the tournament loser in the respective round.

Tournament incentives are ubiquitous in economic life, and typically lead to a more spread pay distribution (and thus more inequality) than the underlying effort and ability justifies (Frank & Cook, 2010). We mimic this observation in our setup with a large difference in tournament prizes for winners and losers that hardly warrants the observed effort differences within dyads in a given round. This income difference magnifies over the rounds because of the substantial time gap (16 seconds) that arises after the first round and that makes it nearly impossible for the first-round loser to catch up in the subsequent rounds.[56] The condition thus induces inequality, caused by a competitive procedure that is difficult to justify on fairness grounds. In addition, this feature allows subjects to grow into their favorable or unfavorable economic positions over the course of the three remaining real-effort task rounds. This seems important in view of the conjecture that the successful are responsible for the erosion of the societal cooperation (e.g., Piff et al., 2012; Piff, 2013). For example, Piff (2013) observes that rich players in a rigged Monopoly Game experiment favoring their own economic status become increasingly imperious as inequality gets larger.

Note that our *Tournament* design includes two components - competition and unjust procedure - that are absent in the *Piece Rate* condition and additionally results in higher income inequality than the *Piece Rate* condition. These three aspects arguably go often hand in hand in real-world settings, where initial advantages are amplified in competitive contexts, leading to enhanced inequality (e.g., Frank & Cook, 2010; Stiglitz, 2012). For example, if performance in or quality of primary school determines access to better secondary schools and subsequently to college, students end up with better jobs and higher earnings (see e.g., Chetty et al., 2011). At the same time, combining these three aspects provides a powerful instrument to probe the effects of (unjust) income inequality on social interactions. This is important as previous evidence suggests that inequality

[56]Note that winning the first round depends on exerted effort and to a large degree on matching luck (i.e., the random assignment of the interaction partner).

effects are subtle (see discussion in Section 4.6). As such, our focus is on maximizing the impact of inequality in the *Tournament* conditions in comparison to the inequality in the *Piece Rate* condition, and not on fully differentiating the marginal effects of the three ingredients (higher inequality, competition, unjust procedure).

4.2.2 Stage 1: Measurement of Fairness Perception

We measure subjects' fairness evaluations of the payment schemes to assess whether the piece rate versus tournament manipulation was successful in creating perceptions of unfair inequality. To gauge the impact of the procedures on participants, we measure fairness perceptions both before and after the stage-1 game. At the beginning of the experiment, participants receive the detailed instructions about the stage-1 real-effort task and the payment procedures of their condition. They then answer three control questions about the procedure. Next, they are asked to indicate on a scale from 0 (very unfair) to 10 (very fair) how fair they consider the payment procedures in stage 1. They also indicate their gender, age, and field of study. After that they start with the real-effort task.

The first assessment provides a fairness judgment based on a verbal description of the mechanism, absent any experience of the task and the outcomes. Our second measurement takes place immediately after the end of stage 1. Subjects have then completed four rounds of the real-effort task and received feedback on the number of correctly placed sliders and the corresponding payoffs of both subjects in the dyad. Thus, we can observe whether and how experiencing the task and the resulting feedback affects subjects' fairness evaluations.

4.2.3 Stage 2: Measurement of Social-Interaction Effects

In the second stage, we use a trust game to measure the effects of the exogenous income variation on social interactions. In this game there are two player roles, the first mover (trustor) and the second mover (trustee). The first mover has an endowment of €6.00 while the second mover has an endowment of €0.00. The first mover decides whether or not to transfer her endowment to the second mover. If she does not transfer, the game ends and the earnings will be €6.00 for the first mover and €0.00 for the second mover.

In contrast, if she transfers her endowment, the experimenter triples the endowment such that the second mover receives €18.00 (and first mover has €0.00 now). The second mover then decides how much of the €18.00 to send back to the first mover (by the cent). Payoffs follow directly from the second mover's decision.

To obtain information on both decisions and the underlying processes, we use the strategy method. More precisely, we first elicit from each player in the dyad their decision as a first mover, and then their decision as a second mover conditional on having received a transfer (because otherwise there is no decision to be made). The player roles in the game are randomly determined after all decisions have been made and subjects are well aware of this fact. Therefore, this modification allows us to answer our first research question (i.e., the effect of inequality on trust in other individuals in a group; first mover) and the second research question (i.e., the greediness of individuals as a function of stage-1 income; second mover), within the same context.

We also measure participants' beliefs regarding the behavior of the other player in this stage. Specifically, we ask subjects to indicate whether they believe the other player in the dyad transferred her endowment when acting as a first mover (yes/no), and to indicate how much they think the other player sends back when acting as second mover (in six ranges: €0 to €3.00; €3.01 to €6.00; ...; €15.01 to €18.00). We do not incentivize beliefs because the preclusion of hedging opportunities would have required rather complex randomizations. Given the randomization in the implementation of the strategy method we did not want to complicate matters further.

We implemented two variations of the trust game stage. In condition *Piece Rate* and *Tournament*, stage-1 dyads remain intact and proceed together to stage 2. We emphasized at the very beginning of the experiment that subjects will interact with the same partner throughout the whole experiment. At the start of stage 2, subjects are reminded of this fact. They also receive a reminder of their own and the other person's stage-1 earnings before making any choices in the trust game. In contrast, in condition *Tournament-New* the dyads are re-matched in stage 2, such that each person will play with a person with whom she did not interact in stage 1. Again, we made clear at the beginning of the

experiment that they interact with different, randomly determined subjects in the two stages. At the beginning of stage 2, they were informed about the new match and they received information on their own and the other persons' (the new partner in the dyad) earnings from stage 1. This design precludes attributions of responsibility for each other's stage-1 outcomes. Moreover, as only earnings (but not effort) are communicated, is it not possible to attribute high or low stage-1 earnings to luck or effort.

4.2.4 Procedural details and variable definitions

In total, 636 subjects took part in the experiment that was programmed using z-Tree (Fischbacher, 2007): 160 in condition *Piece Rate*, 134 in condition *Tournament*, and 342 in condition *Tournament-New*. While we conducted *Piece Rate* and *Tournament* in parallel, we added *Tournament-New* after completing the other conditions to scrutinize the generality of the results. The first two conditions were run on a subject pool at the Universities in Heidelberg and Mannheim (balanced across conditions). For condition *Tournament-New* we used the same subject pool and recruited 202 new subjects. In addition we ran sessions at the laboratory at the Technical University Berlin with a total of 140 subjects to increase power, given the larger number of subgroups in matching stage-1 winners and losers. Participants were undergraduate students from a wide range of different majors, who were recruited with ORSEE (Greiner, 2015) in Berlin and Mannheim and with Hroot (Bock et al., 2014) in Heidelberg.

Final payoffs were determined by adding payoffs from both the real-effort stage and the trust game. A typical session lasted about 50 minutes, and subjects earned, on average about €13.40 (approximately $14.70 at that time), with final payoffs ranging from €1.20 to €30. There was no show-up fee in addition to the incentivized payoffs; that is, incentives were very salient.

At the beginning of a session we matched participants in equal-gender dyads, with one mixed dyad if there was an uneven number of (fe)males. This was done based on the information about each subjects' gender from the initial questionnaire. The matching procedure was anonymous and in particular subjects were not aware of the exact match-

ing procedure. We implemented this matching procedure to control for possible gender differences in the performance in the multiple-round slider task (Gill & Prowse, 2014) and in the behavior in the trust game (Bellemare & Kröger, 2007).

In the presentation of the results we use the following conventions. In the fixed dyads conditions *Piece Rate* and *Tournament* we will call the person with the higher income in a dyad "successful" and the person with the lower income "unsuccessful". In the *Tournament-New* condition, participants encounter new partners, leading to various matches based on the stage-1 income. In the presentation, we denote subjects as "successful" if stage 1 income equals €12.00 and as "unsuccessful" if stage-1 income equals €1.20. This definition reflects the typical payoff pattern for the successful and unsuccessful in condition *Tournament* (results are robust to alternative definitions). In our analysis using the successful-unsuccessful denomination, we drop observations with equal income (in *Piece Rate* and *Tournament*, N=12) and unclassified subjects with an income between €12.00 and €1.20 (in *Tournament-New*, N=54).

4.3 Results: Income Inequality Manipulation

We first provide evidence on effort levels, i.e., the number of correctly positioned sliders, in the different conditions. The *Piece Rate* and *Tournament* conditions did not result in different levels of effort with an average number of correctly solved sliders of 75 in *Piece Rate* and 76 in *Tournament* in all four rounds ($p = 0.795$, two-sided t-test). Effort in *Tournament-New* was somewhat higher at 81 compared to *Tournament* ($t = 2.28$, $p = 0.023$). Importantly, the average difference in effort levels between the two players in a dyad in the first slider task does not differ in all three conditions (3.93 in *Piece Rate*, 4.33 in *Tournament*, and 4.54 in *Tournament-New*, two-sided t-tests, all $p > 0.28$).

Table 4.1: Stage-1 Earnings

	Piece-Rate	*Tournament*	*Tournament-New*
Earnings: mean	5.77	6.60	6.60
Earnings: median	5.75	6.60	6.60
Earnings: 10% percentile	4.93	1.20	1.20
Earnings: 90% percentile	6.70	12.00	12.00

Notes: Entries are in €.

Table 4.1 displays stage-1 earnings and shows that the tournament condition has the intended effect on inequality. While average earnings are comparable across the different treatments, the variation in earnings is much larger in Tournament and Tournament-New than in Piece Rate. That is, small initial differences in effort translate into vast income inequality in *Tournament* and *Tournament-New*, but not in *Piece Rate*.

It is conceivable that subjects perceive the high reward for the tournament winner as justified, taking a meritocratic perspective and focus on incentives for performance (see e.g., Cappelen et al., 2007). This is not what happens in the current context.

Table 4.2: Fairness Evaluation of Payment Mechanism

Point of evaluation	Evaluators	*Piece Rate*	*Tournament*	*Tournament-New*
Before experience	All	7.17	3.69^{***}	3.91^{***}
		(n=160)	(n=134)	(n=342)
After experience	All	6.78^{+++}	$2.44^{+++,***}$	$2.90^{+++,***}$
		(n=160)	(n=134)	(n=342)
After experience	Successful	7.32	2.98^{***}	3.57^{***}
		(n=78)	(n=63)	(n=144)
	Unsuccessful	$6.36^{\#\#}$	$1.92^{\#\#,***}$	$2.00^{\#\#\#,***}$
		(n=78)	(n=63)	(n=144)

Notes: Entries are fairness ratings ranging from 0 (perceived as very unfair) to 10 (perceived as very fair). Significance levels: 10%, 5%,1% (two-sided t-test); pairs with equal earnings excluded in analyses of successful and unsuccessful. *,**,*** indicates significant difference between Piece Rate and Tournament conditions. #,##,### indicates significant difference between successful and unsuccessful. +,++,+++ indicates significant difference between evaluation before and after experience.

Table 4.2 shows that participants perceive the tournament mechanism as substantially less fair than the piece-rate mechanism. We observe strong treatment differences both

before and after the experience of the task and for both the successful and the unsuccessful: the piece-rate scheme always receives much higher fairness evaluations than the two tournament schemes. Experiencing the task leads to lower evaluations compared to the mere verbal description for all three conditions. In all three conditions, the unsuccessful perceive the task as less fair than the successful.

We conclude that the stage-1 manipulation succeeded in inducing strong differences in income inequality and fairness perceptions across piece rate and tournament conditions. Moreover, successful and unsuccessful subjects strongly differ in their fairness perceptions, reflecting a self-serving bias that might have lead the successful to perceive the procedures and resulting positional differences as more justifiable than the unsuccessful.

4.4 Results: Social Interaction Effects for Fixed Dyads

4.4.1 Main Effects

We now turn to the analysis of whether the strong differences in payoff inequality and fairness perception between *Piece Rate* and *Tournament* affect behavior in the stage-2 trust game. Figure 4.1 and Table 4.3 show our main results. We observe strong treatment effects, with the share of trusting participants (i.e., transferring their endowment to the second mover) being almost 20 percentage points lower in *Tournament* than in *Piece Rate* (top panel, Table 4.3). Trust is significantly lower in *Tournament* for both the successful and the unsuccessful. However, we do not detect significant differences in trust between these subgroups in either treatment.

Figure 4.1: Trust rates and returns (in €) in *Piece Rate* and *Tournament*

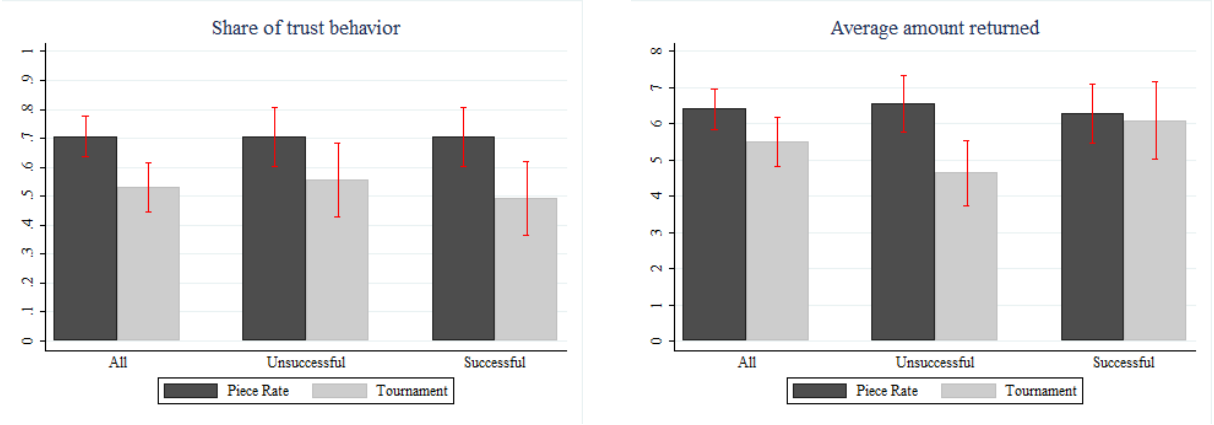


Table 4.3: Social interaction effects of payment mechanism

Point of evaluation	Evaluators	*Piece Rate*	*Tournament*
Trusting	All	71% (n=160)	53%*** (n=134)
	Successful	71% (n=78)	49%*** (n=63)
	Unsuccessful	71% (n=78)	56%*** (n=63)
Amount returned	All	€6.41 (n=154)	€5.50** (n=134)
	Successful	€6.30 (n=78)	€6.10 (n=63)
	Unsuccessful	€6.55 (n=78)	€4.65##,*** (n=63)

Notes: Significance levels: 10%, 5%,1% (two-sided t-test); pairs with equal earnings excluded in analyses of successful and unsuccessful. *,**,*** indicates significant difference between Piece Rate and Tournament conditions. #,##,### indicates significant difference between successful and unsuccessful.

Result 1: *Unjust inequality in stage 1 is detrimental for social trust in stage-2 interaction for fixed dyads.*

The bottom panel of Table 4.3 shows the amounts returned by the second mover. Remember that there are no strategic considerations at this stage and that these amounts are conditional on the trust decision of the first mover resulting in a budget of €18 for the second mover and €0 for the first-mover. We observe that amounts returned are almost €1 lower in the Tournament than in the *Piece Rate* condition (6.4 vs 5.5). Thus, transferring the budget implies an expected loss for the first mover in *Tournament*. This effect is mainly driven by the behavior of the unsuccessful stage-1 subjects. While there

is no difference in the amounts returned across conditions for the successful, the stage-1 losers strongly reduce these amounts in *Tournament*. Consequently, amounts returned are significantly lower for the unsuccessful than for the successful in *Tournament*.

Despite the higher amounts returned in *Tournament* by the successful, we may argue that they still fall strongly short of relevant normative benchmarks. First, they give less than the unsuccessful relative to their wealth. Second, in spite of having typically earned €12 in stage 1 (vs. €1.20 for their partner), they are far from sharing the stage-2 income (return €9), or overall income (return €15.60) equally. However, failure to meet such normative criteria is not restricted to the successful. In *Piece Rate*, stage-1 payoff differences are modest in most dyads, and both the successful and the unsuccessful fail to share their income equally (return €9). It seems that in general, stage-1 income is not taken into consideration when deciding about how much to return to the trustor. The observer's higher normative expectations towards the stage-1 winners make this behavior look less acceptable for the successful in *Tournament*.

Result 2: *Unjust inequality in stage 1 is detrimental for generosity in stage-2 interaction for fixed dyads.*

Result 3: *In the low-inequality environment (Piece Rate) both the winners and the losers are equally generous; in the high-inequality environment (Tournament) the winners are more generous in absolute terms, and less generous relative to their wealth.*

While reduced trustworthiness (generosity) affects the distribution of trust game earnings resulting in a higher variance and skewness, reduced trust affects overall welfare because of the inefficiency of forgoing the tripled payoffs after transfer. Indeed, we observe that the welfare effects are substantial. Expected trust game earnings are €1.08 lower in the *Tournament* condition (€7.26 vs. €6.18), a 15% loss compared to the *Piece Rate* condition.

4.4.2 Underlying Mechanism

The previous analysis has illustrated that there are substantial differences in trust and trustworthiness in the fixed-dyad design of the *Tournament* vs. *Piece Rate* condition. Our controlled laboratory context allows us to shed more light on the underlying mechanisms of this effect. We discuss the role of beliefs, the effect of pure inequality (not necessarily perceived as unjust), and the case of random losses in dyads with equal performance in *Tournament*.

Beliefs. In stage 2 we measured subjects' beliefs regarding the other player's behavior as a trustor and as a trustee in a dyad. In the Appendix (Table A.1), we show that the *Tournament* condition induces more pessimistic beliefs regarding both trust and amounts returned. These effects are significant for the whole sample, but only significant for the successful subgroup when differentiating by stage-1 outcome. That is, the stage-1 condition affects subjects' beliefs. In tables 4.4 and 4.5 we investigate whether these beliefs can explain the treatment effects on trust and trustworthiness. The tables provide four specifications: Specifications 1 and 2 verify the raw comparisons discussed above including various controls. Specifications 3 and 4 include beliefs about trust and trustworthiness.

We find a clear correlation between beliefs and behavior. For trust, beliefs about the other person's trust and her trustworthiness relate to higher trust. The latter effect makes sense from a strategic point of view (expecting lower returns on trust), while the former effect suggests a conditionally-cooperative or reciprocal view (conditioning on behavior if the other person were in the trustor's position). Results on trustworthiness support the reciprocal view as well. Higher beliefs on amounts returned by the other player relate to higher amounts returned. Because strategic aspects are absent for the second mover, beliefs about the other person's returns can only play a role in terms of reciprocal thinking. Note that while beliefs play a role for both trustor and trustee behavior, the main treatment effects of the *Tournament* condition remain substantial when including the beliefs. That is, beliefs cannot fully explain the effect of unjust inequality on social interactions.

Table 4.4: Determinants of trust

Dependent variable: Transfer (yes/no) to second mover

	(1)	(2)	(3)	(4)
Tournament	-0.178**	-0.147	-0.133*	-0.131
	(-3.05)	(-1.74)	(-2.06)	(-1.45)
male	-0.082	-0.097	-0.059	-0.071
	(-1.42)	(-1.61)	(-0.94)	(-1.10)
Successful)		0.010		0.070
		(0.13)		(0.81)
Tournament x Successful		-0.076		-0.006
		(-0.63)		(-0.05)
Belief in trust by other			0.428***	0.411***
			(6.64)	(6.20)
returnbelief			0.046***	0.047***
			(3.27)	(3.33)
Observations	294	282	294	282
Joint effect of tournament variable		χ^2=9.67, $p < 0.01$		χ^2=4.05, $p = 0.132$

Notes: Marginal effects from probit regression with robust z statistics in parentheses. All regressions control for session size and location. Linear regressions support the sign of the interaction term in the probit regressions. Belief in amount returned by other scaled to 100 cents.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.5: Determinants of amounts returned

Dependent variable: Amount returned in cents

	(1)	(2)	(3)	(4)
Tournament	-101	-198**	-40	-166**
	(-1.90)	(-2.70)	(-0.85)	(-2.68)
male	-215***	-207***	-170***	-157***
	(-4.05)	(-3.77)	(-3.63)	(-3.33)
Succesful		-8		17
		(-0.13)		(0.31)
Tournament_successful		167		253**
		(1.53)		(2.80)
Belief in trust by other			13	22
			(0.26)	(0.44)
returnbelief			77***	84***
			(7.42)	(8.62)
Observations	294	282	294	282
Joint effect of tournament variable		$F = 3.70$, $p = 0.026$		$F = 4.59$, $p = 0.011$

Notes: Tobin regressions with robust t statistics in parentheses. All regressions control for session size and location. Linear regressions support the sign of the interaction terms in the tobit regressions. Belief in amount returned by other scaled to 100 cents

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Pure inequality. While our design does not aim at disentangling the different aspects of unjust inequality and the subsequent erosion of trust and trustworthiness, we can use within-treatment variation in stage-1 earnings differences to obtain some insights into the effects of pure inequality, i.e., inequality that is not necessarily perceived as unjust. We define the earnings difference as the difference between a participant's own and the partner's stage-1 earnings. We use the same specification as in the regressions in Tables 4.6 and 4.5 and include the earnings difference, or alternatively its absolute value. We do this in the *Piece* Rate and *Tournament* conditions separately, and in the combined set of observations. Note that in the *Piece Rate* condition, we can study the effect of pure inequality absent the unjust and competitive allocation mode in *Tournament*. Although inequality is less severe than in *Tournament*, in *Piece Rate* there were still 78 dyads with a nonzero earnings difference, ranging from €0.05 to €4.10.

Table 4.6: Effect of pure inequality

	Piece Rate	*Tournament*	All
Trust			
Earnings difference	.01	-.003	-.003
	(.31)	(.78)	(.75)
Earnings difference	.1	-.013	-.017
(absolute value)	(2.11)**	(.82)	(2.83)***
Amounts returned			
Earnings difference	-5	8	7
	(.19)	(1.79)*	(1.76)*
Earnings difference	39	-14	-12
(absolute value)	(.91)	(1.19)	(2.09)**

Notes: Each cell reports the coefficient from a separate regression. Marginal effects from probit regressions for Trust with robust z-stats in parenthesis. Tobit regressions for Amounts Returned with robust t-stats in parenthesis. Amounts are in cents and Earnings and Earnings differences are scaled to 100 cents. All regressions control for session size, location and gender. *,**,*** indicates significant difference from zero at 10%, 5% and 1% level.

Table 4.6 shows the coefficients for the earnings difference variables (each entry refers to one separate regression). We do not find evidence of any negative effects of inequality on stage-2 behavior within either the *Piece Rate* or the *Tournament* conditions. When

combining the observations from the two treatments, the absolute value of the earnings difference becomes significant and negative for both trust and amounts returned, capturing the treatment effects between *Piece Rate* and *Tournament*. In sum, there is no evidence that pure inequality is driving the observed negative social-interaction effects.

Equality of outcome versus equality of opportunity. In the *Tournament* condition, 8 dyads ended up with an equal performance in the first round of the slider task. In this case, a random draw determined the player who received the high payoff and the time bonus (vs. low payoff and time penalty). Comparing random winners and losers, we find that random winners tend to trust less but return more money, albeit the differences are insignificant possibly due to low number of observations. Controlling for a random loss or win in the regressions in tables 4.6 and 4.5 by including a dummy for bad and good luck, we find that all results are qualitatively unaffected. There are no significant effects for random winners and losers compared to other successful and unsuccessful. That is, the treatment effects are not merely driven by a potential perception of the random draw (equality of opportunity) being unfair compared to, for example, an equal split of the payment (equality of outcome).

4.5 Results: Social-Interaction Effects in New Dyads

The comparison between *Piece Rate* and *Tournament* has revealed strong detrimental effects on social interactions. In this section, we test the boundaries of this effect by re-matching subjects into new dyads in stage 2. While the experience and perception of competition and unjust inequality is identical to the *Tournament* condition (see results of section 4.3), a direct attribution of "responsibility" for the mutual stage-1 outcomes is absent in this condition. A negative attribution of high stage-1 earnings for the successful to undeserved luck also becomes more difficult as effort information on the stage-1 dyad is not available.[57] The rematching of dyads allows us to distinguish between the role of

[57] König-Kersting et al. (2017) find that outcome information biases the perception of the underlying process ("outcome bias"). They find that the bias is mainly driven by positive random outcomes being falsely attributed to the decision maker's skill. If this effect transfers to the current setting, we expect that good stage-1 outcomes should more likely be attributed to skill, rather than luck, by stage-2 players.

a player's own income and the income of the matched partner: this was impossible in *Tournament* because these incomes were perfectly correlated.

We first run simple probit/tobit regressions with treatment dummies (and controls) to compare average behavior over all groups in *Tournament-New* (trust = 65%; amount re-turned = €6.61) to *Piece Rate* (trust = 71%; amount returned = €6.41) and *Tournament* (trust = 53%; amount returned = €5.50). The results show that *Tournament-New* does not differ significantly from *Piece Rate*, but leads to significantly larger trust and generosity than *Tournament* ($x \in \mathcal{X}^2 = 4.82$, $p = .028$ and $x \in \mathcal{X}^2 = 9.52$, $p = 0.02$).

Table 4.7: Social interaction effects - *Tournament-New*

	Participants	vs. all (1)	vs. successful (2)	vs. unsuccessful (3)
Trusting	All	65% (n=342)	64% (n=144)	64% (n=144)
	Successful	68% (n=144)	71% (n=56)	69% (n=67)
	Unsuccessful	62% (n=144)	61% (n=67)	55% (n=56)
Amount returned	All	€6.61 (n=342)	€6.49 (n=144)	€6.51 (n=144)
	Successful	€7.37 (n=144)	€7.96 (n=56)	€6.98$^{\#}$ (n=67)
	Unsuccessful	€5.74*** (n=144)	€5.48*** (n=67)	€5.52** (n=56)

Notes: *,**,*** indicates significant difference between successful and unsuccessful; #,##,### indicates significant difference between successful partner and unsuccessful partner at the 10%, 5%, 1% level, test of proportion for trust, and two-sided t-test for amounts returned. Unclassified participants (n=54, i.e., those with an income between €12.00 and €1.20) are excluded when conditioning on successful and unsuccessful decision maker or successful and unsuccessful partner. This leads to different number of observations across cells, depending on stage-2 matches with unclassified subjects.

Next, Table 4.7 shows detailed results for Trust and for Amounts Returned, separately for successful and unsuccessful decision makers, and successful and unsuccessful partners in the dyad. The upper panel of Table 4.7 shows trust behavior. There are no significant raw differences in trust between the successful and the unsuccessful (column 1), and neither between situations interacting with a successful partner (column 2), and an unsuccessful

partner (column 3). However, there is a tendency to trust the stage-1 losers less and also for the losers to trust less. Accordingly, trust within dyads of unsuccessful participants is lower than trust within dyads of successful participants (55% vs. 71%, $z = 1.77$, $p = 0.08$). Regressions reveal that winners are 12.4 percentage points more likely to trust others than losers, which is a significant effect (see Table 4.8).

The lower panel of Table 4.7 shows that stage-1 winners are significantly more generous as second movers than stage-1 losers are. This holds for interactions with other winners and for interactions with losers. In fact, the successful in *Tournament-New* behave more generously on average than the successful under the Piece-rate condition (7.37 vs 6.3, $p = 0.02$, two-sided t-test). When matched with another stage-1 winner, winners give even more to the partners in the dyad than when matched with a stage-1 loser (€7.96 vs. €6.98). As in the case of trust, these effects lead to an overall large difference of generosity within the group of unsuccessful people versus the group of successful people (€5.52 vs. €7.96, $t = 3.56$, $p < 0.001$).[58]

The result that dyads of stage-1 losers perform worst in terms of trust and trustworthiness suggests that the detrimental effect of inequality on trust and trustworthiness is not driven by inequality within dyads per se. Moreover, because of the reduced trust and trustworthiness within the group of dyads of stage-1 losers, stage-2 inequality is larger, and stage-2 welfare is lower in this group compared to the winner dyads. The expected welfare loss of the loser dyads amounts to €0.96, a 13% loss compared to the winner dyads. As in the case of trust, a regression analysis shows that the winners return significantly higher amounts in the trust game (Table 4.8).

Result 4: *The detrimental effects of unjust inequality on social interactions are dampened in newly assembled dyads. Negative effects derive mainly from interactions among the unsuccessful.*

A closer look at the participants' beliefs explains the differences in trust game behavior between *Tournament* and *Tournament-New*. Table 4.8 shows that the effect of beliefs on

[58] We can compare behavior in mixed dyads of successful matched with unsuccessful in *Tournament-New* to the respective group in *Tournament*. We find that the successful are more trusting in *Tournament-New* than in *Tournament* ($p = 0.02$) and equally trustworthy. The behavior of the unsuccessful does not differ significantly between *Tournament* and *Tournament-New*.

Table 4.8: Determinants of trust and amounts returned-*Tournament-New*

	Trust	Trust	Amounts Returned	Amounts Returned
Successful	0.124*	0.141	234***	160***
	(2.00)	(1.83)	(4.18)	(3.53)
Successful Partner	0.050	-0.062	56	-35
	(0.79)	(-0.82)	(1.00)	(-0.83)
male	-0.127*	-0.130	-88	-82
	(-2.00)	(-1.75)	(-1.52)	(-1.79)
Belief in trust by other		0.581***		122*
		(7.42)		(2.08)
returnbelief		0.620***		76***
		(4.51)		(6.08)
Observations	246	246	246	246

Marginal effects from probit regression for Trust with robust z statistics in parentheses. Tobit regressions for Tobit regressions for Amounts Returned with robust t statistics in parentheses. Amounts are coded in cents. All regressions control for session size and location. Belief in amount returned by other scaled to 100 cents.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

trust and amounts returned emerge in *Tournament-New* just as in *Tournament*. However, while in *Tournament* there were substantial negative effects of the stage-1 interaction on beliefs, especially for the winning partners, there are no such negative effects in *Tournament-New* (see Appendix A.2). Moreover, in *Tournament-New* the successful stage-1 players hold more positive views than the unsuccessful ones, especially when paired with another successful person.

We also observe that social aspects must be relevant for the observed effects. That is, the negative effects for the unsuccessful stage-1 dyads cannot simply derive from higher risk aversion caused by their lower income. We observe negative effects for the loser in both trust (potentially affected by risk attitude) and the non-strategic behavior as second mover. Moreover, in the comparison between *Piece Rate* and *Tournament*, where unsuccessful players were always matched with successful ones, there were no differences between the two groups. In contrast, recent literature suggests that, if relative position is salient, inequality may lead the poor to take higher levels of risk (see Payne et al., 2017,

and references therein). We therefore interpret our results in terms of reduced levels of social capital within groups of unsuccessful subjects, rather than in terms of risk attitudes.

4.6 Discussion

Our experiment investigates the potential negative effects of unjust economic inequality on social interactions and focuses, in particular, on the role of the economically successful in harming the social fabric. Our finding that unjust inequality arising in a competitive environment has substantial effects on trust and trustworthiness supports the view that such an environment might be detrimental to social interactions, well-being, and more generally to social capital (Kawachi et al., 1997; Verhaeghe, 2014; Buser & Dreber, 2015).[59] Increased pessimism about others' willingness to cooperate and thus a lower willingness to take the social risk of trusting a stranger is also indicative for a decline in social capital. Indeed, we not only find that beliefs are correlated with behavior but also that they are significantly more pessimistic if inequality is unjust. As a consequence, a vicious cycle of decreasing trust and cooperation may result, leading to a substantial loss of social capital.

Importantly, we find that the decline in trust and trustworthiness is mostly driven by the less well-off. Thus, we find no evidence for the hypothesis that the behavior of the successful is mainly responsible for the erosion of the social fabric. This is consistent with recent findings of Camera et al. (2016). They report that the worse-off subjects discriminate against better-offs by cooperating less with them in a repeated helping game, even when wealth is determined by chance, leading to an overall efficiency loss in the long run. Zheng (2018) similarly reports a higher degree of selfish behavior in a team production setting for low status subjects, where status is endowed in non-monetary terms (public praise). In Table A3 in the appendix, we summarize a larger set of experimental studies that relate to the question of the impact of inequality and competition on cooperation

[59] Besides negative economic consequences, limited social interaction between the poor and rich may also increase the cultural gap between them. New evidence by Bertrand and Kamenica (2018) suggests that media consumption, consumer behavior, and time use of the rich and poor in the US have not diverged much since the 1960s despite the tremendous increase in income inequality, while social attitudes did diverge.

and trust. Although these studies greatly differ in terms of design, the overall picture is consistent with negative social capital effects being more likely. However, the table shows a rather mixed picture about which social status group may drive the observed effects. That is, differences in implementation of inequality may be important for the relevant channel driving social capital effects.

If the, arguably modest, degree of competition and unjust inequality in a lab setting can induce strong effects on social behavior, we may expect the consequences to be even more severe in more significant situations outside the lab. However, our results also hint to the boundaries of such effects. Negative effects on trust and trustworthiness are overall reduced if the interaction partner has not directly contributed to the existing income inequality within a dyad. This happens despite the fact that subjects perceive the tournament as equally unfair in the two *Tournament* conditions. At a first glance, this result contradicts results in Buser and Dreber (2015) who report negative effects of competition on cooperation even in newly assembled groups. In contrast toBuser and Dreber, however, subjects in our new-dyads condition were aware of their own and the other player's in-come situation. It seems likely that the apparent uncertainty about outcomes in Buser and Dreber induces a behavior closer to our condition of fixed dyads. Indeed, positive trust game effects emerge in the new-dyads condition especially in interactions between two stage-1 winners, i.e., in a situation with high income and income equality. If information about other's income is absent, positive effects on trust (and trustworthiness) may not be easily realized.

The observed differences between the fixed dyads and the newly assembled dyads hint at the volatility of the subtle psychological effects caused by inequality or fairness cues. Moreover, our manipulation combined strong inequality with a competitive and perceived unjust payment scheme. We have argued that this key feature of our setup is relevant in many contexts outside the lab such as in educational systems, labor markets or one's social environment (e.g., Chetty et al., 2011; Hanushek & Woessmann, 2006; Lemieux et al., 2009). The more modest inequality emerging in condition Piece Rate is perceived as fair and allows players to maintain a high level of trust and trustworthiness. The perceived

justice of the institution from which unequal outcomes derive thus seems to constitute an essential aspect. Our results lend support to Starmans, Sheskin, and Bloom (2017), who argue that it is not inequality per se that bothers people in life, but economic unfairness. Indeed, dyads of unsuccessful participants in Tournament-New score low on trust and trustworthiness despite having equal outcomes; their experience of disadvantages caused by unfair economic allocations seems to affect behavior, rather than inequality per se.

The finding of low social capital among the poor is consistent with field data on deprived neighborhoods in the UK. Compared to wealthy neighborhoods, social capital is lower in deprived neighborhoods, measured by interactions among people in the same neighborhood and thus social class (Nettle et al., 2011). Our results suggest that these field data may not simply caused by selection of people in or out of certain neighbor-hoods. Nevertheless, selection and upbringing may be important in the field. For example, in contrast with our and with Nettle et al.'s finding, Martinsson et al. (2015) report that Colombian university students from a wealthy university are less cooperative among each other than those at a lower social status university. The differences in upbringing and life experiences seem to have an opposite effect in this sample compared to Nettle et al.'s UK data.

A large literature in psychology has argued that rich, high-status individuals are less generous in *absolute* terms than poor, low-status individuals (e.g. Piff et al., 2010; Piff et al., 2012; Guinote et al., 2015). In particular, this literature makes the causal claim that increasing wealth induces less social behavior. In correlational field data, the existence of a negative correlation between status and prosocial behavior has been questioned (Trautmann et al., 2013), and various studies have recently shown that wealthy individuals are often more prosocial and more generous in absolute terms (e.g. Andreoni et al., 2017; Smeets et al., 2015), and also relative to their wealth position (Korndörfer et al., 2015). A negative causal effect of increased wealth and status on prosociality may still exist, dampening an otherwise positive correlation between wealth and prosocial behavior through a selection effect if the prosocial are economically more successful.

In contrast to the results found in the above cited psychological literature, in our experiment the better-off stage-1 winners are always more generous than the worse-off in the second stage of the trustgame. Arguably, stage-1 losers should thus be more trusting than the winners, expecting higher returns from trust. Yet, this is not the case. Moreover, in the *Tournament-New* condition we observe that unsuccessful when matched with an-other unsuccessful subject are less trusting and less trustworthy than the successful when matched with another successful. That is, overall welfare is reduced and a higher degree of inequality emerges within their group of stage-1 losers. These results suggest that negative effects of unjust inequality are driven by the behavior of the poor, rather than the behavior of the rich.

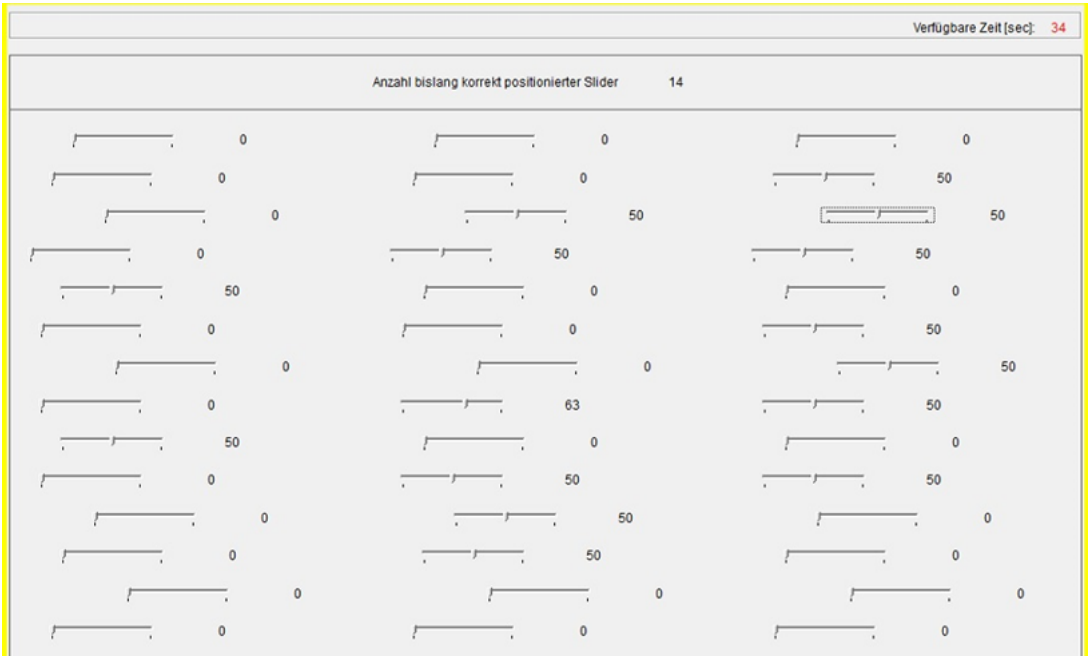
Some qualifications need to be made with respect to the last point. Despite their higher degree of generosity in absolute terms, the successful players still fall substantially short of obvious normative benchmarks for second movers, such as equal sharing of the stage-2 payoffs, or even equal sharing of total experimental payoffs; they give a lower share of their income compared to the poor. That is, while the successful are more prosocial, they fall short of the potential normative expectations we may hold with respect to their behavior (in contrast to the empirical expectations as measured in the experiment, which may turn out more consistent with actual behavior). This is not the case for the poor, for whom no such expectations exist in the current setup. The same is probably true in larger contexts outside the lab. Such an expectation-behavior gap for the rich may explain the appeal of picturing elites as immoral and selfish in popular discourses, which were eager to pick up the results by Piff et al. (2012) and others supporting the view of the selfish elite.

Appendix

A.1 Instructions and Screen Shots

An English translation of the original instructions can be found online at `https://www.dropbox.com/s/21c7unjcko336ck/Merged%20Instructions%20%28English%29.pdf?dl=0` (to prevent the current document from becoming excessively large). The instructions also contain relevant screen shots with explanations. Here we present the screenshot of the real-effort task as referred to in the main text.

Figure 4.2: Screen shot: slider task (42 sliders per round)



A.2 Effects of Stage-1 Condition on Beliefs

Tables A 4.9 and A 4.10 show beliefs in treatments Piece Rate and Tournament, and Tournament-New, respectively. Treatment comparisons find no significant differences between Piece rate and Tournament-New beliefs about trust (63% vs. 59%, $p = 0.386$) and about amounts returned (€5.85 vs. €5.74, $p = 0.676$).

Table 4.9: Effects of stage-1 condition on trust game beliefs

	Participants	*Piece Rate*	*Tournament*
Belief in Trust by Other	all	63% ($n = 160$)	50%** ($n = 134$)
	successful	58% ($n = 75$)	43%* ($n = 63$)
	unsuccessful	68% ($n = 75$)	56%* ($n = 63$)
Expected Amount Returned by Other	all	€5.85 ($n = 160$)	€5.08** ($n = 134$)
	successful	€5.69 ($n = 78$)	€4.36*** ($n = 63$)
	unsuccessful	€5.96 ($n = 78$)	€5.60## ($n = 63$)

Notes: *,**,*** indicates significant difference between treatment; #,##,### indicates significant difference between successful and unsuccessful; at the 10%, 5%,1% level, two-sided t-test for amounts returned, test of proportion for trust; pairs with equal earnings excluded in analyses of successful and unsuccessful.

Table 4.10: Effects of stage-1 condition on trust game beliefs - *Tournament-New*

	Participants		*Piece Rate*	*Tournament*
Belief in Trust by Other	all	59% ($n = 342$)	63% ($n = 144$)	54% ($n = 144$)
	successful	56% ($n = 144$)	66%* ($n = 56$)	51%$^{\#}$ ($n = 67$
	unsuccessful	61% ($n = 144$)	63%* ($n = 67$)	59% ($n = 56$
Belief in Amount Returned by Other	all	€5.74 ($n = 342$)	€5.90 ($n = 144$)	€5.35** ($n = 144$)
	successful	€6.04 ($n = 144$)	€6.91 ($n = 56$)	€5.40$^{\#\#\#}$ ($n = 67$)
	unsuccessful	€5.33** ($n = 144$)	€5.44*** ($n = 67$)	€4.93 ($n = 56$)

Notes: *,**,*** indicates significant difference between treatment; #,##,### indicates significant difference between successful and unsuccessful; at the 10%, 5%,1% level, two-sided t-test for amounts returned, test of proportion for trust; pairs with equal earnings excluded in analyses of successful and unsuccessful.

A.3 Experimental literature on inequality and competition

Table A3 presents laboratory experiments that study questions regarding the effect of competition and inequality on social interaction. We concisely summarize the key study aspects and the social interaction effect. If there exist any such effects, we indicate whether they are driven by the behavior of the successful/rich or the unsuccessful/poor.

Table 4.11: Overview of experimental studies

	Treatment	**Stage 1**	**Stage 2**	**Social Interaction Effect**
Anderson et al. (2006)	Public/private show up fee	High/low show-up fees as inequality "*priming*"	Trust game	Private: Trust (-), driven by the *successful*; Public: Trust (=).
Sadrieh and Verbon (2006)	High skewed treatment	Randomly endowed "*earnings*" to create inequality	Public goods game	Contribution (+), driven by the *unsuccessful*.
Brandts and Riedl (2017)	Rivalry/non-rivalry treatment	Prisoner's dilemma game (with a competitive setting in rivalry treatment) that creates inequality as "*priming*"	The circle test (similar to a dictator game)	"Generosity" towards others who they interacted before (-), driven by *unsuccessful*.
Harbring (2010)	Competition game	Inequality as "priming" via a competitive game.	Trust game	Trust (-), unclear who drives the results.
Heap et al. (2013)	High/low inequality	Randomly endowed "*earnings*" to create inequality.	Trust game (standard trust game or a labor market setting)	Trust and trustworthiness (-), driven by *both* the *successful* and the *unsuccessful*.
Smith (2011)	Inequality	Randomly endowed "*earnings*" to create inequality	Trust game	Trust and trustworthiness (=) because the *successful* trust less but return more, while the *unsuccessful* do the opposite.
Zheng (2018)	Baseline/random/ true status	Painting evaluation, as status "*priming*"	Similar to a dictator game over losses.	Less selfish behavior (+), driven by the *successful*, those who earned their high status.

Table 4.12: Overview of experimental studies (*continued*)

	Treatment	**Stage 1**	**Stage 2**	**Social Interaction Effect**
Greiner et al. (2012)	High/low inequality	Randomly endowed "*earnings*" to create inequality	Modified trust game (1st period only)	Trust (-), driven by *both*
Nishi et al. (2015)	Visible/non-visible wealth difference under three levels of inequality.	Randomly endowed "*earnings*" to create inequality	Cooperation game	Cooperation (-), driven by the *successful* when inequality is visible. Inequality itself is not sufficient to drive this result, visibility is the key driver.
Buser and Dreber (2015)	Feedback on slider task	Competitive sliders task tournament, as "*priming*"	Public goods game	Contribution (-), driven by *both* but more so by the *unsuccessful*.
Camera et al. (2016)	Info on wealth and role vs. no info	Helping game, payoffs as "*earnings*" to create inequality	Helping game (cont'd)	The act of helping others (-), driven by *both*.
Heap et al. (2013)	Direct, indirect, and no competition	Competitive double-auction market, payoffs as "*earnings*" to create inequality. In the absence of competition, randomly endow subjects.	Public goods game	Contribution (+) when no direct competition in stage 1, driven by the *successful*. Contribution (-) with direct competition, driven by *both*
Falk (2017)	High vs. low social status.	Relative status info revealed, as "*priming*" of social status.	Electric shocks to others for personal gain	Incidence of shocking others is higher when high/low status group interact, driven by *both*.

Table 4.13: Overview of experimental studies (*continued*)

	Treatment	**Stage 1**	**Stage 2**	**Social Interaction Effect**
Friedrichsen (2017)	Inequality	Randomly endowed "*earnings*" to create inequality	Consumers with different initial wealth choose between socially responsible product and a cheaper alternative.	The unsuccessful choose the socially responsible products significantly more than the successful; no base-line available to compare overall effect due to inequality.
Lotito et al. (2017)	High vs. low inequality	Competitive real-effort task (admin tasks), as inequality "*priming*"	Public goods game	Contribution (-): partial info on income/performance. Contribution (+) if full info. No competition effect, results driven by information about inequality.
Bejarano et al. (2018)	Inequality as a result of random shock or endowed inequality vs. equality as a baseline.	Randomly endowed "*earnings*" to create inequality vs. inequality as a result of a random shock.	Trust game	Trust (-): the successful trust less if inequality is due to a random shock that makes the second-mover poor. Trustworthiness (-): the unsuccessful return less regardless of the source of inequality.

Notes: In column Stage 1, "priming" indicates that stage-1 game payoffs either prime a winner/economically successful or loser/economically unsuccessful mindset and that they are not used as an endowment for the stage-2 game (i.e., payoffs in the two games are independent). "Earnings" indicate that the amount of money earned/randomly assigned to the subjects in the stage-1 game is used as the endowment of the stage-2 game. In column Social interaction effect, "(-)", "(=)", and "(+)" denotes a decrease, no effect, and increase of socially desirable interaction such as trust, cooperation, and contribution to public goods. Entries with n/a indicates not applicable because no relevant information is available.

Discussion and Conclusion

We now briefly summarize our results in view of the influence social preferences have on decision making in our studies. More specific conclusions of the individual studies have already been discussed in the respective chapters. We assess the impact of social preferences by the degree of pro- and antisocial behavior observed. Concerning prosocial (antisocial) behavior we refer to subjects, who accept a lower[60] payoff for themselves to increase (decrease) the payoff of the other person, they are interacting with. In our first study in chapter 1, subjects are given the opportunity to costly destroy another participant's endowment. We can directly assess the degree of antisocial behavior by the frequency of positive destruction decisions across all treatments. Decisions in the bargaining task in chapter 2 could be influenced by strategic considerations and risk preferences. It is thus not possible to disentangle effects on decisions based solely on social preferences from those based on other factors. Therefore we do not consider it for this analysis. In the study in chapter 3, subjects rank eight different payment pairs. We categorize a subject as prosocial, if at least in one case she prefers[61] an allocation where she receives a lower payment and the other player a higher one compared to an alternative allocation. This categorization applies the other way round also for antisocial behavior. The mixed cases, in which both kinds of behavior are observed, are not counted for either of the two categories. They mainly correspond to subjects, showing inequality-averse behavior. In our last study in chapter 4, subjects play a Trust Game. Prosocial behavior

[60] Concerning the cases, in which subjects do not face own costs, it could either be that subjects are indifferent with respect to the payoff of the other person, or they have social preferences as a tiebreaking rule. As it is hard to distinguish these cases, both are considered as belonging to the category of selfish preferences.

[61] This classification includes the case of indifference, as when being indifferent one potentially is willing to forgo some own payoff.

is measured by the number of cases, in which the trustee chooses to return a positive amount of money to the first player. Results of all studies are summarized in the table below:

Overview: Pro- and antisocial behavior

	Number of decisions	**Percentage of decisions being compatible with**	
		Prosocial Behavior	Antisocial Behavior
Chapter 1	119	X	23,1%
Chapter 3	336	43,2%	2,2%
Chapter 4	342	84,2%	X

Taken together, results indicate that social preferences have a substantial influence on individual decision making. There is though great heterogeneity in the fractions across studies. This suggests that behavior depends a lot on the exact framework and parameters. The degree of antisocial behavior observed in chapter 1 is much higher than in chapter 3. There are two possible explanations: Firstly, the costs for decreasing another subject's payoff are considerably lower in chapter 1 than in chapter 3. Secondly, in chapter 1 this is the only decision subjects can make. If in chapter 3 we classify the cases where the reduction of others' payoffs has no cost as antisocial behavior, this fraction would be much higher. In chapter 4, prosocial behavior is very high, despite involving high costs. But all of these cases correspond to situations in which the trustee has been transferred a considerable amount of money beforehand. Thus, reciprocity seems to have a strong impact. If we only count the cases as prosocial where the trustee returns more money than the initial amount sent by the first mover, the percentage of subjects exhibiting prosocial behavior would be much lower.Concerning the degree of social preferences underlying subjects' decisions, observed behavior can be seen as a lower bound. As there are specific costs involved for each decision, subjects who under these conditions neither show pro- nor antisocial behavior do not necessarily have selfish preferences. It could be the case that they would conform with either of these types of behavior when there would be lower costs involved.[62].

[62]Results of study 3 suggest, a certain amount of subjects have pro- and antisocial preferences as tie breaking mechanism. But as discussed beforehand, they are counted as belonging to the category of selfish preferences.

Finally, we discuss how to interpret the results of our lab experiments in assessing the degree of social preferences in the real world. A lot has already been said about the external validity of lab experiments. While Charness and Fehr (2015) take a rather optimistic point of view, Levitt and List (2007) are more skeptical about the transferability of results into the real-world. Levitt and List note several potential confounds on behavior, such as the degree of scrutiny, representativeness of samples, low stakes and the decision-making context. In contrast, Charness and Fehr, argue that all these aspects can in principle be accounted for: Lab experiments have also been conducted with non-student subjects and with high stakes. Moreover, many decisions in the field are observed by other people as well and the context there is even harder to control. Taken together, lab findings should be considered with caution as behavior in various studies does not necessarily translate to similar results in the field. But it is nevertheless important to investigate such questions in the lab to offer us some benchmark of understanding such behavior.

Concerning social preferences in particular, Levitt and List provide evidence that due to scrutiny the degree of prosocial behavior could be overestimated by lab experiments. The extent of prosocial behavior usually is negatively correlated with the degree of anonymity. But as people and institutions are strongly motivated by reputation effects, differences can go in either direction, depending on the observability of one's actions.[63] In contrast to experiments, in most situations in the real world it is legally not possible to reduce someone else's income or destroy someone else's possessions. Thus, antisocial behavior is probably much less pronounced in the field. Still, one could argue that this aspect could be captured in the lab by an adequate cost function reflecting legal punishments. For at least moderately high costs antisocial behavior is observed rather rarely in both environments. However, one still observes a considerable degree of antisocial behavior in experiments when costs are very low or zero.[64] Fortunately, overall the number of people exhibiting prosocial behavior is much higher than those engaging in antisocial behavior.

[63] See for example Hoffman et al., 1994 as well as Hoffman et al., 1996 for an analysis of giving behavior in Dictator Games and Harbaugh (1998) for a study about the prestige motive in donating to charities.

[64] See the results of chapter 1 or those of other comparable experiments as for example Zizzo and Oswald (2001),

List of Figures

List of Tables

References

Abbink, K., Masclet, D., & van Veelen, M. (2011). Reference point effects in antisocial preferences. *Working Paper*.

Abbink, K., & Sadrieh, A. (2009). The pleasure of being nasty. *Economics Letters*, *105*(3), 306–308.

Adams, J. S. (1963). Towards an understanding of inequity. *The Journal of Abnormal and Social Psychology*, *67*(5), 422–436.

Alesina, A., & Perotti, R. (1996). Income distribution, political instability, and investment. *European Economic Review*, *40*(6), 1203–1228.

Alvaredo, F., Chanel, L., Piketty, T., Saez, E., & Zucman, G. (2017). *World Inequality Report, Paris.*

Amanatullah, E. T., & Tinsley, C. H. (2013). Punishing female negotiators for asserting too much or not enough: Exploring why advocacy moderates backlash against assertive female negotiators. *Organizational Behavior and Human Decision Processes*, *120*(1), 110–122.

Anderson, L. R., Mellor, J. M., & Milyo, J. (2006). Induced heterogeneity in trust experiments. *Experimental Economics*, *9*(3), 223–235.

Andreoni, J., Nikiforakis, N., & Stoop, J. (2017). Are the rich more selfish than the poor, or do they just have more money? A natural field experiment. *Working Paper*.

Attanasi, G., Battigalli, P., & Nagel, R. (2016). Disclosure of Belief-Dependent Preferences in a Trust Game. *Working Paper*.

Aumann, R., & Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica*, 1161–1180.

Autor, D. H., Katz, L. F., & Kearney, M. S. (2008). Trends in US wage inequality: Revising the revisionists. *The Review of Economics and Statistics*, *90*(2), 300–323.

Azmat, G., & Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics*, *30*, 32–40.

Babcock, L., & Laschever, S. (2009). *Women don't ask: Negotiation and the gender divide*. Princeton University Press.

Bartling, B., Grieder, M., & Zehnder, C. (2017). Competitive pricing reduces wasteful counterproductive behaviors. *Journal of Public Economics*, *156*, 34–47.

Bear, J. B., & Babcock, L. (2012). Negotiation topic as a moderator of gender differences in negotiation. *Psychological Science*, *23*(7), 743.

Bejarano, H., Gillet, J., & Rodriguez-Lara, I. (2018). Do negative random shocks affect trust and trustworthiness? *Southern Economic Journal*, *85*(2), 563–579.

Bellemare, C., & Kröger, S. (2007). On representative social capital. *European Economic Review*, *51*(1), 183–202.

Benabou, R. (1996). Inequality and growth. *NBER Macroeconomics Annual*, *11*, 11–74.

Bertrand, M., & Kamenica, E. (2018). *Coming apart? Cultural distances in the United States over time* (Tech. Rep.). National Bureau of Economic Research.

Bjoernskov, C. (2018). *The Political Economy of Trust. In: The Oxford Handbook of Public Choice; Congleton, Roger D and Grofman, Bernard N and Voigt, Stefan (eds.)* (Vol. 2). Oxford University Press.

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, *55*(3), 789–865.

Bloom, N., Sadun, R., & Van Reenen, J. (2012). The organization of firms across countries. *The Quarterly Journal of Economics*, *127*(4), 1663–1705.

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, *63*(2), 131–144.

Bock, O., Baetge, I., & Nicklisch, A. (2014). Hroot: Hamburg registration and organization online tool. *European Economic Review*, *71*, 117–120.

Boll, C., Rossen, A., & Wolf, A. (2017). The EU gender earnings gap: Job segregation and working time as driving factors. *Jahrbücher für Nationalökonomie und Statistik*, *237*(5), 407–452.

Bolle, F., Tan, J. H., & Zizzo, D. J. (2014). Vendettas. *American Economic Journal: Microeconomics*, *6*(2), 93–130.

Bolton, G. E., Brandts, J., & Ockenfels, A. (2005). Fair procedures: Evidence from games involving lotteries. *The Economic Journal*, *115*(506), 1054–1076.

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

Bolton, G. E., & Ockenfels, A. (2006). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review*, *96*(5), 1906–1911.

Bowles, H. R., Babcock, L., & Lai, L. (2007). Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask. *Organizational Behavior and Human Decision Processes*, *103*(1), 84–103.

Brandts, J., & Riedl, A. (2017). Market Interaction and Efficient Cooperation. *Working Paper*.

Breza, E., Kaur, S., & Shamdasani, Y. (2017). The morale effects of pay inequality. *The Quarterly Journal of Economics*, *133*(2), 611–663.

Buser, T., & Dreber, A. (2015). The flipside of comparative payment schemes. *Management Science*, *62*(9), 2626–2638.

Camera, G., Deck, C., & Porter, D. (2016). Do Economic Inequalities Affect Long-Run Cooperation? *Working Paper*.

Cappelen, A. W., Eichele, T., Hugdahl, K., Specht, K., Sørensen, E. Ø., & Tungodden, B. (2014). Equity theory and fair inequality: A neuroeconomic study. *Proceedings of the National Academy of Sciences*, *111*(43), 15368–15372.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*(3), 818–827.

Cappelen, A. W., Konow, J., Sørensen, E. Ø., & Tungodden, B. (2013). Just luck: An experimental study of risk-taking and fairness. *American Economic Review*, *103*(4), 1398–1413.

Card, D., Mas, A., Moretti, E., & Saez, E. (2012). Inequality at work: The effect of peer salaries on job satisfaction. *American Economic Review*, *102*(6), 2981–3003.

Charness, G., & Fehr, E. (2015). From the lab to the real world. *Science*, *350*(6260), 512–513.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, *99*(1), 431–57.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593–1660.

Conrath, D. W. (1972). Sex role and "cooperation" in the game of chicken. *Journal of Conflict Resolution*, *16*(3), 433–443.

Cooper, D. J., & Dutcher, E. G. (2011). The dynamics of responder behavior in ultimatum games: A meta-study. *Experimental Economics*, *14*(4), 519–546.

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

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–74.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*(7137), 794–796.

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

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

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female

leaders. *Psychological Review*, *109*(3), 573–598.

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

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

Eckel, C., & Grossman, P. J. (2008a). Gender and negotiation in the small: Are women (perceived to be) more cooperative than men? *Negotiation Journal*, *24*(4), 429–445.

Eckel, C., & Grossman, P. J. (2008b). Men, women and risk aversion: Experimental evidence. *Handbook of Experimental Economics Results*, *1*, 1061–1073.

Eichberger, J., & Kelsey, D. (2000). Non-additive beliefs and strategic equilibria. *Games and Economic Behavior*, *30*(2), 183–215.

Eichberger, J., & Kelsey, D. (2014). Optimism and pessimism in games. *International Economic Review*, *55*(2), 483–505.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 643–669.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, *94*(4), 857–869.

Exley, C. L., Niederle, M., & Vesterlund, L. (2016). *Knowing when to ask: The cost of leaning in* (Tech. Rep.). National Bureau of Economic Research.

Falk, A. (2017). Status inequality, moral disengagement and violence. *Working Paper*.

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

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, *326*(5952), 535–538.

Faravelli, M. (2007). How context matters: A survey based experiment on distributive justice. *Journal of Public Economics*, *91*(7-8), 1399–1422.

Fehr, D. (2018). Is increasing inequality harmful? Experimental evidence. *Games and*

Economic Behavior, *107*, 123–134.

Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, *7*(2), 235–266.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Fehr, E., Naef, M., & Schmidt, K. M. (2006). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review*, *96*(5), 1912–1917.

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

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

Fisman, R., Jakiela, P., Kariv, S., & Markovits, D. (2015). The distributional preferences of an elite. *Science*, *349*(6254), aab0096.

Frank, R. H., & Cook, P. J. (2010). *The winner-take-all society: Why the few at the top get so much more than the rest of us*. Random House.

Friedrichsen, J. (2017). Is Socially Responsible Production a Normal Good? *Working Paper*.

Gaertner, W., & Schokkaert, E. (2012). *Empirical social choice: Questionnaire-experimental studies on distributive justice*. Cambridge University Press.

Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, *102*(1), 469–503.

Gill, D., & Prowse, V. (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, *5*(2), 351–376.

Gill, D., Prowse, V., & Vlassopoulos, M. (2013). Cheating in the workplace: An experimental study of the impact of bonuses and productivity. *Journal of Economic Behavior & Organization*, *96*, 120–134.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics*, *115*(3), 811–846.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.

Greiner, B., Ockenfels, A., & Werner, P. (2012). The dynamic interplay of inequality and trust: An experimental study. *Journal of Economic Behavior & Organization*, *81*(2), 355–365.

Grossman, P. J., Komai, M., et al. (2013). Within and across class envy: Anti-social behaviour in hierarchical groups. *Working Paper*.

Guinote, A., Cotzia, I., Sandhu, S., & Siwa, P. (2015). Social status modulates prosocial behavior and egalitarianism in preschool children and adults. *Proceedings of the National Academy of Sciences*, *112*(3), 731–736.

Guiso, L., Sapienza, P., & Zingales, L. (2004). The role of social capital in financial development. *American Economic Review*, *94*(3), 526–556.

Hanushek, E. A., & Woessmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, *116*(510), 63–76.

Harbaugh, W. T. (1998). The prestige motive for making charitable transfers. *American Economic Review*, *88*(2), 277–282.

Harbring, C. (2010). On the effect of incentive schemes on trust and trustworthiness. *Journal of Institutional and Theoretical Economics*, *166*(4), 690–714.

Harsanyi, J. C. (1967). Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Management Science*, *14*(3), 159–182.

Healy, P. J. (2011). Epistemic foundations for the failure of Nash equilibrium. *Working Paper*.

Heap, S. P. H., Tan, J. H., & Zizzo, D. J. (2013). Trust, inequality and the market. *Theory and Decision*, *74*(3), 311–333.

Heinz, M., Normann, H.-T., & Rau, H. A. (2016). How competitiveness may cause a gender wage gap: Experimental evidence. *European Economic Review*, *90*, 336–349.

Hernandez-Arenaz, I., & Iriberri, N. (2018). Gender Differences in Alternating-Offer

Bargaining: An Experimental Study. *Working Paper*.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, *7*(3), 346–380.

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review*, *86*(3), 653–660.

Hu, A., & Zou, L. (2015). Sequential auctions, price trends, and risk preferences. *Journal of Economic Theory*, *158*, 319–335.

John, L. K., Loewenstein, G., & Rick, S. I. (2014). Cheating more for less: Upward social comparisons motivate the poorly compensated to cheat. *Organizational Behavior and Human Decision Processes*, *123*(2), 101–109.

Kawachi, I., Kennedy, B. P., Lochner, K., & Prothrow-Stith, D. (1997). Social capital, income inequality, and mortality. *American Journal of Public Health*, *87*(9), 1491–1498.

Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *The Quarterly Journal of Economics*, *112*(4), 1251–1288.

König-Kersting, C., Pollmann, M., Potters, J., & Trautmann, S. T. (2017). Good decision vs. good results: Outcome bias in the evaluation of financial agents. *Working Paper*.

Konow, J. (2001). Fair and square: The four sides of distributive justice. *Journal of Economic Behavior & Organization*, *46*(2), 137–164.

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature*, *41*(4), 1188–1239.

Korndörfer, M., Egloff, B., & Schmukle, S. C. (2015). A large scale test of the effect of social class on prosocial behavior. *PloS one*, *10*(7), e0133193.

Langer, A., Stewart, F., Smedts, K., & Demarest, L. (2017). Conceptualising and Measuring Social Cohesion in Africa: Towards a perceptions-based index. *Social Indicators Research*, *131*(1), 321–343.

Lehrer, E. (2012). Partially specified probabilities: Decisions and games. *American Economic Journal: Microeconomics*, *4*(1), 70–100.

Leibbrandt, A., & List, J. A. (2014). Do women avoid salary negotiations? Evidence from a large-scale natural field experiment. *Management Science*, *61*(9), 2016–2024.

Lemieux, T., MacLeod, W. B., & Parent, D. (2009). Performance pay and wage inequality. *The Quarterly Journal of Economics*, *124*(1), 1–49.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, *21*(2), 153–174.

Li, S., Qin, X., & Houser, D. (2018). Revisiting gender differences in ultimatum bargaining: Experimental evidence from the US and China. *Journal of the Economic Science Association*, *4*(2), 180–190.

Lo, K. C. (1996). Equilibrium in beliefs under uncertainty. *Journal of Economic Theory*, *71*(2), 443–484.

Lotito, G., Migheli, M., Ortona, G., et al. (2017). Competition, Information and Cooperation. *Working paper*.

Lutzker, D. R. (1961). Sex role, cooperation and competition in a two-person, non-zero sum game. *Journal of Conflict Resolution*, *5*(4), 366–368.

Marianne, B. (2011). New perspectives on gender. In *Handbook of labor economics* (Vol. 4, pp. 1543–1590). Elsevier.

Martinsson, P., Villegas-Palacio, C., & Wollbrant, C. (2015). Cooperation and social classes: Evidence from Colombia. *Social Choice and Welfare*, *45*(4), 829–848.

Mazei, J., Hüffmeier, J., Freund, P. A., Stuhlmacher, A. F., Bilke, L., & Hertel, G. (2015). A meta-analysis on gender differences in negotiation outcomes and their moderators. *Psychological Bulletin*, *141*(1), 85–104.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, *10*(1), 6–38.

Miller, L., Montero, M., & Vanberg, C. (2015). Legislative bargaining with heterogeneous disagreement values: Theory and experiments. *Working Paper*.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, 286–295.

Nash, J., et al. (1950). Equilibrium points in n-person games. *Proceedings of the National*

Academy of Sciences, *36*(1), 48–49.

Nettle, D., Colléony, A., & Cockerill, M. (2011). Variation in cooperative behaviour within a single city. *PloS one*, *6*(10), e26922.

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

Nishi, A., Shirado, H., Rand, D. G., & Christakis, N. A. (2015). Inequality and visibility of wealth in experimental social networks. *Nature*, *526*(7573), 426–429.

Oxoby, R. J., & Spraggon, J. (2008). Mine and yours: Property rights in dictator games. *Journal of Economic Behavior & Organization*, *65*(3-4), 703–713.

Piff, P. (2013). Does money make you mean? *TedxMarin. TEDtalks. Video retrieved from http://www. ted. com/talks/paul_piff_does_money_make_you_mean*.

Piff, P., Kraus, M. W., Côté, S., Cheng, B. H., & Keltner, D. (2010). Having less, giving more: The influence of social class on prosocial behavior. *Journal of Personality and Social Psychology*, *99*(5), 771-784.

Piff, P., Stancato, D. M., Côté, S., Mendoza-Denton, R., & Keltner, D. (2012). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences*, *109*(11), 4086–4091.

Piketty, T., & Saez, E. (2003). Income inequality in the United States, 1913-1998. *The Quarterly Journal of Economics*, *118*(1), 1–41.

Piketty, T., & Saez, E. (2014). Inequality in the long run. *Science*, *344*(6186), 838–843.

Piketty, T., Saez, E., & Zucman, G. (2017). Distributional national accounts: Methods and estimates for the United States. *The Quarterly Journal of Economics*, *133*(2), 553–609.

Polak, B. (1999). Epistemic conditions for Nash equilibrium, and common knowledge of rationality. *Econometrica*, *67*(3), 673–676.

Putman, R. D. (2000). *Bowling alone: The collapse and revival of American community.* Simon & Schuster New York.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 1281–1302.

Rigdon, M. L. (2012). An experimental investigation of gender differences in wage negotiations. *Working Paper*.

Roemer, J. E. (1998). *Theories of distributive justice*. Harvard University Press.

Roemer, J. E. (2008). *Equality of opportunity*. Springer.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, 97–109.

Sadrieh, A., & Schröder, M. (2012). The desire to influence others. *Working Paper*.

Sadrieh, A., & Verbon, H. (2006). Inequality, cooperation, and growth: An experimental study. *European Economic Review*, *50*(5), 1197–1222.

Säve-Söderbergh, J. (2007). Are women asking for low wages? Gender differences in wage bargaining strategies and ensuing bargaining success. *Working Paper*.

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

Smeets, P., Bauer, R., & Gneezy, U. (2015). Giving behavior of millionaires. *Proceedings of the National Academy of Sciences*, *112*(34), 10641–10644.

Smith, A. (2011). Income inequality in the trust game. *Economics Letters*, *111*(1), 54 - 56.

Solnick, S. J. (2001). Gender differences in the ultimatum game. *Economic Inquiry*, *39*(2), 189–200.

Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour*, *1*(4), 0082.

Stiglitz, J. E. (2012). *The price of inequality: How today's divided society endangers our future*. WW Norton & Company.

Stuhlmacher, A. F., & Linnabery, E. (2013). Gender and negotiation: A social role analysis. *Handbook of Research on Negotiation Research*, 221–248.

Trautmann, S. T., van de Kuilen, G., & Zeckhauser, R. J. (2013). Social class and (un) ethical behavior: A framework, with evidence from a large population sample. *Perspectives on Psychological Science*, *8*(5), 487–497.

Veblen, T. (2017). *The theory of the leisure class*. Routledge.

Verhaeghe, P. (2014). *What about me?: The struggle for identity in a market*. Scribe

Publications, London.

Wilkinson, R., Pickett, K., & Cato, M. S. (2009). *The spirit level. why more equal societies almost always do better.* Penguin, London.

Wolff, I. (2014). When best-replies are not in equilibrium: Understanding cooperative behaviour. *Working Paper*.

Zhang, L., & Ortmann, A. (2013). On the interpretation of giving, taking, and destruction in dictator games and joy-of-destruction games. *Working Paper*.

Zheng, J. (2018). High Social Status Induces Pro-Social Behavior. *Working Paper*.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98.

Zizzo, D. J., & Oswald, A. J. (2001). Are people willing to pay to reduce others' incomes? *Annales d'Economie et de Statistique*, 39–65.