Cooperation in Social Dilemmas

Five Essays in Experimental Economics

Inaugural-Dissertation

zur Erlangung des Grades

Doctor oeconomiae publicae (Dr. oec. publ.)

an der Ludwig-Maximilians-Universität München

2012

vorgelegt von

Dominik Matzat

Referent:	Prof. Dr. Martin G. Kocher
Korreferent:	Prof. Dr. Joachim K. Winter
Promotionsabschlussberatung:	7. November 2012

Datum der mündlichen Prüfung: 22. Oktober 2012

Namen der Berichterstatter:

Prof. Dr. Martin G. Kocher, Prof. Dr. Joachim K. Winter, Prof. Dr. Klaus M. Schmidt

Acknowledgements

I am grateful to numerous people that provided me with excellent knowledge and support during the years in which this thesis evolved.

First of all, I would like to thank Martin Kocher both in his role as my first supervisor and as a co-author. I benefited a lot from so many constructive discussions and valuable suggestions. It was him who introduced me into the field of experimental economics and I owe him great thanks for all his patience and encouragement.

Furthermore, I thank Joachim Winter and Klaus Schmidt for agreeing to serve as second and third supervisors of my thesis as well as for many valuable comments in internal seminars throughout the years.

Special thanks go to my co-authors Nadja Furtner, Peter Martinsson, Gerhard Riewe and Conny Wollbrant for a pleasurable and fruitful collaboration. I enjoyed it a lot working with all of them.

Many other friends, colleagues and participants at conferences and seminars contributed to this thesis. In particular, I thank Bianca Bauer, Eva van den Broek, Amrei Lahno, Julius Pahlke, Sebastian Strasser and participants of the MELESSA Brown Bag Seminar and the Micro Workshop of the University of Munich, the ESA conferences in Innsbruck 2009 and Luxembourg 2011, the annual meeting of the Verein für Socialpolitik in Frankfurt 2011 and the PhD-Workshop in Behavioral and Experimental Economics in Innsbruck 2011. I further thank the entire staff involved in conducting my experiments at MELESSA and the Seminar of Behavioral and Experimental Economics for a very inspiring research environment.

Financial support from the Ideenfonds of the University of Munich and the German Science Foundation through GRK 801 is gratefully acknowledged.

Finally, I am grateful to Lucia and my parents for their inexhaustible support and patience. I am deeply indebted to them.

Dominik Matzat

Contents

List of Tables

List of Figures

Preface

Understanding cooperation of individuals in social dilemmas is one of the key issues in modern societies as plenty of our daily life situations share this feature. Social dilemmas describe situations in which it is rational for a selfish actor to free-ride although cooperation would be optimal from a social perspective. Examples range from effort decisions in work or sports teams, the private provision of local public goods, the extraction of common pool resources, voter participation in elections to up-to-date international problems like climate protection, debt discipline or the regulation of financial markets. In all these examples a rational and selfish individual, state or organization will choose an action that is beneficial for herself but not for the society as a whole. We can interpret this as a kind of market failure and economists are interested in efficient ways to overcome the problem.

The social dilemma problem has for a long time been recognized in economics. However, in the past two decades the discipline developed a powerful instrument to study its importance in greater detail: laboratory experiments arose. Experimental research showed that the assumption of purely rational and selfish individuals is at odds with much of the empirical evidence. Indeed, it is a robust finding that many individuals cooperate in social dilemmas. This result may not be too surprising for the interested reader; however, it breaks with the most famous assumption of economic theory: the "homo oeconomicus" approach. Thus, a very exciting research area evolved in economics focusing on observed behavior and behavioral theories that try to explain why individuals are willing to cooperate in social dilemmas. Among those concepts are inequity aversion (Fehr and Schmidt, 1999), preferences for social welfare (Charness and Rabin, 2002) or conditional cooperation (e.g. Fischbacher et al., 2001).

In particular, recent research concentrates on three aspects. First, it takes the heterogeneity of individuals as a starting point and searches for common features of cooperative individuals. Second, it studies environmental determinants that foster the general willingness to cooperate in order to recommend institutional designs that raise group cooperation and efficiency. This point is especially important for economic policy as it is one of the main aims of governmental intervention to generate welfare-enhancing structures. Third, it focuses on the endogenous choice and self-selection of individuals into different institutional settings accounting for the fact that people actively shape their personal environment. This dissertation contributes to each of these aspects by presenting five

laboratory experiments. They have in common that they investigate how personal and institutional determinants - separately and in interaction - affect cooperation in social dilemmas, connections we still know too little about.

Laboratory experiments are important for the analysis of social dilemmas as they allow a clean identification of causal effects which otherwise would be difficult or even impossible to obtain. In particular, their high degree of control is a major advantage compared to other empirical methods. They can serve to directly test and improve behavioral theories or be explorative in nature by inspiring new theoretical approaches due to unexpected discoveries. Moreover, laboratory methods can be used to test institutional arrangements in a kind of wind channel. If these institutions work in the supposed way they can be more trustfully recommended for implementation in real world settings. All experiments in this dissertation were conducted with undergraduate students of the University of Munich who took decisions under complete anonymity and real monetary incentives.

To pin down the social dilemma to the laboratory, the following simple linear public goods paradigm (based on Isaac et al., 1985; Isaac and Walker, 1988) is used as a workhorse. We randomly match a group of experimental participants and endow each group member with a fixed amount of money. The participant can split the endowment between her private and a public account. Investments into the private account are paid out in a one-to-one relationship while contributions to the public account are summed up over all group members, multiplied by an exogenously given and commonly known factor and divided equally (i.e. independent from one's own contribution; the public good property) among the group members. This simple voluntary contribution mechanism constitutes a social dilemma if the multiplying factor is smaller than the group size but larger than one. Because the factor is smaller than the group size, the return from investing one unit of money into the public account is smaller than the return from investing it into the private account and hence no rational and selfish individual should contribute any positive amount. On the other hand, as the factor is larger than one there is an efficiency gain for the group as a whole from investing into the public account. It is exactly this trade-off that is inherent in all social dilemma situations. Hence, the public goods paradigm allows us to study social dilemmas in the laboratory and the following chapters use this approach and extend it appropriately.

Chapters 1 and 2 focus on individuals' endogenous choice between different sanctioning institutions that can be made available in social dilemmas. While participants in Chapter 1 decide between formal (centralized) and informal (decentralized) punishment institutions, participants in Chapter 2 cast their vote for informal punishment, informal reward or a no-

sanction environment. Chapter 3 studies the role of hierarchy within teams facing a social dilemma by giving one group member distribution rights over the group benefit. Chapters 4 and 5 finally focus on the behavioral foundations of cooperation. In Chapter 4 cooperative behavior is connected to beliefs, trust and natural risk. At last, Chapter 5 concentrates on gender effects by linking cooperative behavior both to underlying preferences for cooperation and to a person's expectation on the behavior of others. The chapter concludes with a cross-country overview of data from related studies.

Chapter 1 studies individual's willingness to dispense with informal peer-to-peer punishment by assigning punishment rights to a formal authority. This research is motivated by the real life observation that organizations in social dilemmas can usually choose between setting up formal membership rules and relying on informal sanctions. As an example one might think about soccer clubs which decide on introducing a punishment catalogue for misbehaving players (e.g. for being late). Moreover, the willingness to delegate punishment rights is a fundamental justification for the foundation of nations as they are able to provide centralized police and jurisdictional systems. The main aim of this study is to figure out conditions under which individuals are more willing to assign their rights to a formal authority. In a way, this experiment can therefore be seen as searching for conditions under which Thomas Hobbes' (2008 [1651]) claim that citizens want to overcome the state of nature of a *bellum omnium contra omnes* is more prevalent.

While there are many experiments in economics focusing on peer-to-peer punishment in social dilemmas (starting with Fehr and Gächter, 2000, 2002) and also a few studying formal punishments (e.g. Tyran and Feld, 2006), this chapter is a novel approach in combining both punishment institutions by directly connecting their costs and effectiveness levels. Participants face a binary public goods game (i.e. an all-or-nothing contribution decision) and repeatedly cast their vote either for informal peer-to-peer punishment which relies on costly mutual sanctioning within the group or for an automatic punishment of free-riders by a costly external authority. The decision within the group is then taken by majority rule. The design ensures that costs and effectiveness levels in the informal institution sum up to those in the formal institution. Experimental treatments vary two important dimensions of real life punishment conditions: the strength of the punishment instrument (correspondingly in both institutions) and the possibility to counter-punish in the peer-to-peer punishment environment.

The results show that the formal institution is much more demanded in case of strong punishment instruments. I find that this is in particular due to the fact that weak formal

institutions raise cooperation but not far enough to compensate for the fixed costs of the organization. On the other hand, there is no effect of counter-punishment on individual's average voting decision which can be explained by the observation that counter-punishment reduces contributions and punishment activities in the informal institution to the same extent. Finally, I add two control treatments using unanimity instead of majority rule to control for the possibility that the first effect is driven by the voting procedure. However, unanimity rule does not increase cooperation in the weak formal institution. Hence, I conclude that individual's willingness to assign punishment rights to an external authority increases in the strength of the available punishment instrument but does not hinge on potential counter-punishment threats in the informal setting.

In Chapter 2, which is a joint work with Martin Kocher, we analyze preferences over informal punishment and reward mechanisms in comparison to a sanction-free environment. While there is already some literature on this issue (especially Sutter et al., 2010) the findings are still inconclusive and necessitate further investigations. The endogenous choice between these institutions is important as many groups in social dilemmas decide themselves on the rules that govern their interaction and, in particular, on a possible introduction of one of those enforcement mechanisms. For example, consider the formation of social norms in work or sports teams that specify how shirking and hard-working team members are treated. Or, if we focus on the level of international politics, consider representatives in the United Nations Security Council who form guidelines on how to deal with countries that violate global security.

Precisely, experimental participants in this chapter choose repeatedly between an informal punishment institution, an informal reward institution and a standard linear public goods game. Moreover, they do not only decide on the institution but also on the strength of the enforcement mechanism in case of punishment or reward. With this setup we extend the study of Sutter et al. (2010) by two features. First, as in Chapter 1, the voting decision appears in each period which enables us to control for learning effects during the course of the experiment. Second, we are the first who combine a vote on institutions with a choice on the strength of the enforcement mechanism. The latter is an important property of many real-life solutions to social dilemmas as groups usually do not only agree on the type of sanction but also on how strong the sanction should be. To account for the fact that the reward leverage is typically more limited than the one of punishment, we implement an asymmetric choice interval comprising punishment mechanisms that are stronger than the highest possible

reward mechanism. Furthermore, we take explicit care for the fact that groups often interact repeatedly in real world situations which suggests that reputation effects might influence institutional preferences. One could, for instance, argue that sanctioning institutions are more attractive in a fixed group structure as the application of sanctions is then more credible due to strategic motivations. To control for this assertion we conduct two treatments that vary group composition: a partner and a stranger matching. While groups stay constant over all periods in the first treatment, participants are randomly re-matched every period in the latter.

Our results, however, show that group composition has no substantial impact on individual's voting decision indicating that strategic considerations are not important for institutional choice. In both treatments we find approximately one half of our subjects voting for the reward institution and one quarter each for the punishment institution and the standard public goods game. This distribution is astonishingly stable over time. Moreover, almost all participants who vote for a sanctioning institution prefer an extreme value of the enforcement mechanism. This suggests that sanction supporters do not rely on weak instruments but rather try to impose strong cooperation incentives. We show that the predominant preference for the strong reward mechanism is more in line with the social welfare argument of Charness and Rabin (2002) than with predictions based on inequity aversion (Fehr and Schmidt, 1999). From an ex post point of view this is justified because the reward institution is the most efficient institution. This is true although the maximal reward leverage in our setup is rather low and although the punishment institution leads to higher cooperation. Finally, we find a strong and henceforth notable gender effect in the voting behavior. Whereas men show a much larger preference for the punishment institution, women more likely vote for the standard public goods game. However, both groups have in common that a majority of their members prefers the reward institution.

In Chapter 3, a joint work with Martin Kocher and Gerhard Riewe, we combine the classic cooperation dilemma with a realistic asymmetry often found in real world applications: an unequal distribution of property rights over the team output. Obviously, these are situations in which the team output is not a pure public good, i.e. its benefits have to be dividable in an unequal way. In such cases hierarchy levels can arise. Consider for example small work teams in consultancies in which a team leader or manager has the responsibility to split some monetary fund or work order among the team members. Other "teams" that often show a natural or exogenously imposed hierarchy structure are for instance sports teams, political parties, military units or families. Surprisingly, hierarchy structures in the

distribution of the team output are neglected in the literature. This is the more astonishing as they comprise a costless implicit sanctioning mechanism because team leaders can both reward contributors and punish non-contributors to the public account, the latter due to the exclusion from the group benefit. We therefore hypothesize that such a hierarchy structure leads to more cooperation than a situation without hierarchy.

To test this hypothesis we consider a repeated public goods game with fixed group composition. Within this setup we implement two experimental treatments: one with and one without a "team allocator". In the team allocator treatment, we design the simplest possible hierarchy mechanism: One person per group is chosen randomly and put in the role of the team allocator. This group member has the authority to distribute the entire revenues from the public account among the group members including herself. The allocator is completely free in her distribution decision and takes the decision in each period of the multi-period setup. In the control treatment, there is no team allocator and the public account is split equally among the group members as in a standard public goods game. However, to adjust the different contribution incentives across treatments - a rational and selfish team allocator will invest into the public account - both the team allocator and one randomly chosen member in the control treatment are forced to contribute their full endowment to the public account.

Our results support the hypothesis that hierarchy leads to more group cooperation. Indeed, ordinary team members in the team allocator treatment contribute 25% more than the respective subjects in the control treatment. We can show that this is due to a large fraction of team allocators who use the implicit sanctioning mechanism in exactly the way described above. Even in the last period in which pretending cooperativeness is pointless, many team allocators behave pro-socially which we show is explainable both by preferences for social welfare (Charness and Rabin, 2002) and inequity aversion (Fehr and Schmidt, 1999). Hence, hierarchical structures mitigate the social dilemma problem. This result is even more remarkable as the mechanism comes without any monetary costs in contrast to punishment or reward mechanisms considered so far in the literature. Furthermore, note that we find this effect already for our *randomly* determined allocator. Recent literature (e.g. Baldassarri and Grossman, 2011; Levy et al., 2011) suggests that this effect is even stronger if team members are allowed to elect their team allocator.

Chapter 4 is a joint work with Martin Kocher, Peter Martinsson and Conny Wollbrant. In this piece of research we connect cooperative behavior to beliefs, trust and natural risk. In contrast to the previous chapters we focus on individual's cooperation in a social dilemma

that is neither manipulated by an institutional mechanism nor influenced by repetition. We know from many experiments that in such a standard environment there is usually a majority of persons who do not free-ride on their group members. However, till today the personal determinants of cooperation in such a situation are unclear. Why do some people cooperate and others not? We focus on three potential driving forces behind cooperation. First, an individual's belief about other group members' contributions could influence her own contribution decision. Second, trust in other's reliability might be associated with positive contributions to the public good as the latter involves a certain degree of confidence in others' cooperation. Third, natural risk attitudes might matter due to the fact that unknown contribution levels of the other group members generate a risky decision. Note that social dilemmas, strictly speaking, involve rather social than natural risk, i.e. risk that stems from human decisions instead of a random event. However, previous research has found clear connections between both aspects suggesting that it is reasonable to concentrate on the more common notion of natural risk. While all these definitely related concepts were separately linked to cooperation before, we are the first who provide a complete and fully incentivized analysis of these concepts.

We do so by presenting an experimental design consisting of three parts. In the first part, participants play a one-shot public goods game based on the design introduced by Fischbacher et al. (2001). Precisely, they first enter their unconditional (i.e. "standard") contribution, second they fill in a conditional contribution schedule stating their contribution for each potential average contribution level of the other group members (applying the so-called strategy method), and third, additionally, they enter a guess (i.e. belief) about others' average unconditional contribution. In Part II, they complete the Holt and Laury (2002) test on natural risk and in Part III they play a trust game similar to Berg et al. (1995).

We find that both beliefs and trust are positively associated with contributions to the public good. This holds not only for the unconditional contribution but also for the likelihood of being classified as a conditional cooperator rather than a free-rider, according to the conditional contribution schedule. However, natural risk surprisingly neither matters for the contribution nor for the trust decision. Indeed, the combination of beliefs, conditional and unconditional contributions provides evidence that our participants do not interpret the contribution decision as risky at all. The link between trust and cooperation is especially remarkable as it comprises an important message for economic policy. If we can form trust-improving societies, this will ceteris paribus reduce the social dilemma problem without the necessity of any further intervention.

Chapter 5, finally, is a joint work with Nadja Furtner, Martin Kocher, Peter Martinsson and Conny Wollbrant and studies gender differences in cooperative behavior. As there is a bunch of previous research on this topic yielding contradicting results, we try to investigate it more closely by eliciting both a person's underlying cooperative preferences and her expectation on others' behavior and connecting these results to actual cooperation. With this setup we can disentangle two different reasons for potential gender differences in cooperation. First, it could be that both sexes have the same cooperative preferences but simply differ in their expectation on group members' average contribution. Assuming the plausible relation that a lower belief leads to lower contributions - an association true for a large group of individuals, the so-called "conditional cooperators" - this would create a gender gap in observed behavior. Second, it could also be the case that both sexes share the same belief about group members' behavior but differ in their underlying cooperative preferences which would also entail such a gender effect.

To study this question we implement the one-shot public goods game invented by Fischbacher et al. (2001), like we did in Chapter 4, and ask first for a person's unconditional contribution and second for her conditional contribution schedule. Thereafter, we again elicit the belief about her group members' average unconditional contribution. Finally, participants fill in a short post-experimental questionnaire that controls among other things for gender. It is important to emphasize that we do not manipulate experimental sessions with regard to the gender composition as we do not want to induce persons to believe that they take part in a gender study. This should strengthen the reliability of our findings.

Our results show that women's unconditional contribution is significantly higher than that of men. Moreover, women both hold a more optimistic belief about other group members' contributions and have a more cooperative underlying preference as obtained out of the conditional contribution schedule. Precisely, we can show that women are more often sorted into the group of "conditional cooperators". Hence, the results suggest that both beliefs and cooperative preferences are important determinants for the observed gender gap. In the last section we connect our results to cross-country data from related experimental studies. The overview shows the following: The gender difference in cooperative preferences with women being more often classified as conditionally cooperative is a robust finding. However, gender differences in the belief vary over studies and indicate that contradicting results regarding individuals' unconditional contributions could be caused by fluctuating beliefs. Thus, the chapter can present an important step in unraveling the confusion on gender effects present so far in the literature.

As a last point, let me remark that all chapters of the dissertation are independent from each other and contain their own introductions and appendices. Hence, they could be read in any order.

Chapter 1

Endogenous Choice of Formal and Informal Punishment Institutions in Social Dilemmas

1.1 Introduction

Social dilemmas are a widely discussed phenomenon. They describe situations in which it is individually rational for selfish actors to free-ride although cooperation would be optimal from a social perspective. The private provision of public goods is a classic example. Experimental results reveal that voluntary cooperation in such dilemmas is usually limited. One solution for this problem is the introduction of a punishment mechanism. Fehr and Gächter (2000, 2002) showed in two seminal papers that an informal (decentralized) punishment institution, i.e. peer-to-peer punishment after the observation of group members' contributions, is able to increase cooperation substantially. An alternative mechanism would be to implement a formal (centralized) punishment institution which instead of peer-to-peer punishment relies on rule-based punishment by an appointed or elected external authority. Such a mechanism is commonly used in reality; consider for example penalties by private organizations like employers' or homeowners' associations or governmental fines in case of tax evasion. Moreover, on an international level organizations like the United Nations or the European Union have at least some power to punish member countries for misbehavior.

Surprisingly, experimental literature started only recently to investigate the formation of formal punishment institutions. For example, studies by Kosfeld et al. (2009) and Putterman et al. (2011) show that formal punishment institutions are favored by experimental participants compared to a no-punishment environment and can increase cooperation as well as group efficiency. The role of endogenous choice is important here as formal institutions do not exist out of the nowhere but have to be formed by institutional members at some point in time. However, none of these studies allows for the comparison between a formal and an informal punishment institution. This is astonishing as real-world organizations can usually decide whether they want to stick to an informal mechanism (comprising social penalties like

ostracism or badmouthing) or whether they want to impose membership rules that include formal fines. As an example one might think about soccer clubs which decide about the implementation of a punishment catalogue for misbehaving players (e.g. for being late). Eventually, the question of formal punishments goes back to the British philosopher's Thomas Hobbes work "Leviathan" (2008 [1651]) in which he states people's wish to depart from the state of nature of a *bellum omnium contra omnes* by assigning punishment rights to a superior authority. This desire serves as a justification for the foundation of nations.

The present chapter addresses the issue of formal versus informal punishment with a controlled laboratory experiment in which participants can directly cast their vote for one of the two punishment institutions. Precisely, we use a repeated binary public goods game in the stranger design and include an endogenous choice stage at the beginning of *each* period. In this stage group members elect the institution. Institutional selection ensues due to majority rule (and in a control experiment we demand unanimity for the implementation of formal sanctions). Formal and informal institutions differ only in their punishment mechanism. In the formal institution punishment occurs only and with certainty to all free-riders, i.e. subjects that do not invest into the public good. The institution is costly and has to be paid a fixed amount irrespective of whether it exerts punishment or not. In the informal institution there exists a costly individual option to punish group members after the contribution decision. This includes the possibility to also punish cooperators and in the respective treatment to retaliate. The informal institution, however, is costless if the punishment instrument is not used. We connect both institutions by a novel approach making the informal institution at most as costly as the formal institution and at most as effective in punishing free-riders.

The main focus is to figure out conditions under which a formal punishment institution is more likely to be installed. In a way this experiment looks for institutional environments in which Thomas Hobbes' statement is more prevalent. Note that laboratory experiments are particularly suited for this question as they allow a clean identification of causal effects. Two dimensions are varied experimentally: the strength of punishment and the possibility to counter-punish. Both dimensions are important with respect to reality as the nature of available punishment instruments differs strongly by situation. The strength of punishment is either weak or strong (correspondingly varied in both the formal *and* the informal institution). Following Tyran and Feld (2006), parameters are set such that in the "weak" condition punishments are too weak to destroy free-riding incentives, i.e. free-riding is still a dominant strategy even if you know for sure that you will be punished. The "strong" condition on the other hand consists of a punishment instrument that is strong enough to ensure that free-riding

is not optimal once you know that you will be punished. Regarding the second dimension, counter-punishment in the informal institution will be made available or not. Recent research (Denant-Boemont et al., 2007; Nikiforakis, 2008) shows that the possibility to retaliate has a strong influence on subject's contribution and punishment behavior. We study whether the counter-punishment option also affects institutional vote. In all treatments, we control for self-selection of behavioral types using an incentivized social value orientation questionnaire and ask for a couple of socio-economic facts including gender.

Results show that formal punishment institutions are more often implemented if the available instrument is strong. While the willingness to establish formal sanctions also emerges under the weak condition, the lower contribution level if implemented reduces their attractiveness in later periods. Interestingly, we do not find an effect of counter-punishment on subject's average voting decision. However, it slightly reduces votes for the formal institution in the beginning. Self-selection of types does not occur in the absence of counter-punishment but plays some role when retaliation is possible in the informal institution. Socio-economic aspects like gender show no systematic effect on the observed voting pattern. Moreover, the underlying voting rule does not seem to drive our results.

As to our knowledge, there exist only two further working papers (Markussen et al., 2011; Kamei et al., 2011) that address the question of endogenous choice between formal and informal institutions. Differences and overlapping with these papers that developed independently from ours will be discussed in detail in the next section. While both papers do not look at counter-punishment they report evidence of an increased preference for the formal institution in case of strong and cheap formal punishments (holding the informal institution constant).

The remainder of this chapter is organized as follows. Section 1.2 gives a brief overview of the related literature. In Section 1.3 we provide the details of the experiment. Section 1.4 presents theoretical predictions and Section 1.5 reports the experimental results. Finally, Section 1.6 concludes the chapter.

1.2 Literature review

Starting with Fehr and Gächter (2000, 2002) a huge literature evolved on the impacts and success of *exogenously* imposed informal punishment institutions in social dilemmas.[1] This literature shows that costly informal punishment, although being irrational from a standard game theoretic point of view, is heavily used by experimental participants and can increase cooperation dramatically (Fehr and Gächter, 2000, 2002). Effects on cooperation depend on the costs of punishment (Anderson and Putterman, 2006; Carpenter, 2007a; Egas and Riedl, 2008) and the effectiveness (strength) of the instrument (Nikiforakis and Normann, 2008; Egas and Riedl, 2008). Even non-monetary punishment can increase cooperation (Masclet et al., 2003; Rege and Telle, 2004; Noussair and Tucker, 2005). However, costly informal punishment is not solely attributed to free-riders but also hits a non-negligible number of cooperators (Falk et al., 2005; Cinyabuguma et al., 2006; Herrmann et al., 2008; Gächter and Herrmann, 2011). Importantly, efficiency effects are ambiguous. While the punishment option tends to reduce efficiency in the short-run (Fehr and Gächter 2002; Egas and Riedl, 2008; Herrmann et al., 2008), positive effects are observed for longer time horizons (Gächter et al., 2008). Counter-punishment reduces punishment and contributions (Denant-Boemont et al., 2007; Nikiforakis, 2008) while third-party punishment, i.e. costly punishment by unaffected individuals, enforces the cooperation norm (Fehr and Fischbacher, 2004; Baldassarri and Grossman, 2011).[2] Further studies investigate for instance the role of network effects (Casari and Luini, 2009; Carpenter et al., 2010), group size (Carpenter, 2007b), income (Masclet and Villeval, 2008), feedback (Nikiforakis, 2010), communication (Bochet et al., 2006; Janssen et al., 2010), threats (Masclet et al., 2011), emotions (Joffily et al., 2011) or differences between punishment and reward (Sefton et al., 2007).

A newer strand of the literature concentrates on the *endogenous* formation of informal punishment institutions when there is the alternative to have a no-sanction environment. Botelho et al. (2005), extending the Fehr and Gächter (2000, 2002) design, report a strong reluctance to form informal punishment institutions. This finding is, for early periods, confirmed by Ertan et al. (2009) and a voting-by-feet experiment of Gürerk et al. (2006). However, the latter studies show that over time punishment institutions can gain support. Sutter et al. (2010) reveal an endogenous choice premium for the level of cooperation when

[1] There is some earlier work on punishment in social dilemmas by non-economists, consider for example Yamagishi (1986, 1988) and Ostrom et al. (1992).

[2] Note that there is a difference between third-party punishment and formal punishments. While both are exerted through unaffected parties, third-party punishment does not occur automatically due to rule violation. Instead, it relies on an individual choice of the observing party.

comparing endogenously and exogenously implemented informal institutions (the impact of elections is also studied by Dal Bó et al., 2010).[3]

Interestingly, only few and very recent papers address the question of endogenous formation of *formal* punishment institutions (all of them having a no-sanction environment as alternative).[4] Tyran and Feld (2006) study the endogenous choice of mild and severe sanctions and find that both are implemented and improve efficiency compared to the no-sanction case due to expectations of commitment and conditional cooperation. Kosfeld et al. (2009) allow subgroups of individuals to implement severe formal sanctions within their union while non-members can free-ride on them without the danger of being punished. The authors show that over time more and more formal punishment institutions arise and almost all consist of the grand coalition. Andreoni and Gee (2011) use a special formal punishment mechanism ("the hired gun") that punishes only the lowest contributor and also find strong support for the willingness to form formal institutions. Interestingly, they have one treatment in which they reveal a demand for formal punishment even *on top* of an informal punishment institution. This results in a crowding-out of peer-to-peer punishment.[5] Finally, Putterman et al. (2011) let subjects choose a formal punishment scheme out of a broad menu of options including the possibility of having no punishment at all. They find that subjects choose high and efficient levels of formal sanctions. Moreover, they report individual characteristics that foster the learning process (e.g. male gender, intelligence and cooperative orientation).

Most closely related to our work, however, are two recent companion working papers by Markussen et al. (2011) and Kamei et al. (2011). These papers are the only two that we are aware of that also allow for a direct endogenous choice between formal *and* informal punishment institutions. Markussen et al. (2011) investigate institutional preferences in a setup in which individuals vote for formal, informal or no-punishment environments in a series of distinct pairwise elections. More precisely, subjects play a linear public goods game 28 times in a row. While they have to play the no-punishment condition for the first four periods, from period 5 on there is an election in every fourth period determining the institution for the next four-period phase using majority rule. Choice is between no punishment and informal (periods 5 and 17), no punishment and formal (periods 9 and 21) as

[3] For a more detailed up-to-date survey on the literature of informal punishment, see Chaudhuri (2011).

[4] In a broader sense, one could add the experimental literature on tax evasion as far as it allows for endogenous parameter choices (e.g. Alm et al., 1999; Feld and Tyran, 2002). It shows that voting on tax, audit or fine rates changes tax compliance compared to exogenous enforcement. Tax evasion games are structurally similar but usually include specific features like income disclosure decisions or mid-level detection rates.

[5] Such a crowding-out effect is also reported in a recent experimental study by Kube and Traxler (2011). However, in their design both punishment mechanisms are given exogenously. Note that the combined effect of formal and informal sanctions is especially interesting in cases in which there is reasonable doubt that a formal authority can prevent the (additional) use of informal sanctions.

well as formal and informal (periods 13 and 25). The authors vary costs (low, high) and strength (weak, strong) of the formal punishment mechanism and find a higher preference for formal over informal sanctions when the punishment instrument is strong and cheap. Overall, however, they report astonishingly strong support for the informal institution and a quite stable picture when comparing the two voting decisions in periods 13 and 25.

Kamei et al. (2011) use almost the same design and concentrate only on the choice between the formal and the informal punishment institution. If the formal institution is chosen, subjects can further decide about whether the public or private account is sanctioned and at which rate (allowing for weak and strong sanctions). Treatment variations concern the number of elections (three vs. six) and whether or not there are fixed costs of the formal institution. In the 3-vote treatments, subjects play a couple of periods under exogenously given institutions beforehand, which are varied to control for order effects. The authors find that most subjects vote for the informal institution once fixed costs exist and they report a slight tendency towards choosing more formal institutions (with efficient parameters) over time.[6]

Importantly, and in stark contrast to our study, both above mentioned papers hold the informal punishment scheme constant over treatments. In our setup, both the formal and the informal institution are adjusted *correspondingly* under the weak and strong condition. This reflects the idea that depending on the social dilemma there might be strong punishment instruments available or not. If they exist, however, then they improve formal *and* informal punishment opportunities at the same time. Moreover, we model the informal institution as being at most as costly as the formal institution and at most as effective in terms of punishing free-riders to get a weaker and more comparable institution. This includes that, in contrast to Markussen et al. (2011) and Kamei et al. (2011), a single person in the informal institution has less punishing power towards a free-rider than the formal institution which is a quite reasonable assumption. Only if the group members work together, they can achieve the same punishment level as the formal institution. As Kamei et al. (2011) we consider learning aspects. While we do not have exogenous experience under different institutions, we cover more voting decisions (ten), two different voting rules and, most importantly, can control for different learning effects under weak and strong conditions. Furthermore, we tackle the role of counter-punishment in the informal institution. Experimental designs differ also in other points (group size, matching protocol, contribution and punishment space, institutional term,

[6] In addition, Markussen et al. (2011) and Kamei et al. (2011) implement exogenous treatments in which groups play the most frequent voting patterns of the endogenous treatments. They confirm Sutter et al.'s (2010) finding of an endogenous choice premium for both the case of informal and non-deterrent formal sanctions.

etc.). Thus we add important insights to Markussen et al.'s (2011) and Kamei et al.'s (2011) findings.

1.3 The experiment

1.3.1 Our public goods setup

We use a repeated binary public goods game with punishment and voting opportunities. Each period $t \in \{1,2,\dots,T\}$ consists of the following three [four] stages: (o) voting, (i) contribution, (ii) punishment [and (iii) counter-punishment] in which the members of group $I = \{1,2,\dots,n\}$ decide simultaneously.[7] In the contribution stage, a subject $i \in I$ decides whether to invest her endowment E into the public or her private account. It is not allowed to split the endowment on both accounts, so this is an either-or-decision.[8] Investments in the private account are paid out in a one-to-one relationship. Investments in the public account, denoted $c_{j,t}$, are summed up over all n group members, multiplied by a factor γ and divided equally among the group members. Hence, after stage 1 group member i receives the following payoff $\pi_{i,t}$:

$$\pi_{i,t} = E - c_{i,t} + \frac{\gamma}{n}\sum_{j=1}^{n} c_{j,t} \tag{1.1}$$

To ensure that a social dilemma exists, the condition $1 < \gamma < n$ has to hold. If this is true, it is individually rational for a selfish actor to invest E into the private account, i.e. to free-ride. However, contributing to the public account (cooperation) would be optimal from a social perspective.

In the punishment stage, payoffs from the contribution stage can change depending on the implemented institution. In case the informal institution was chosen by group voting, subjects receive information on their group members' contribution decisions and are allowed to punish them individually. Each subject i decides for each other group member separately whether she wants to punish this person or not. Note that this includes the possibility to also punish cooperators. Punishment is a simple binary decision; hence subjects cannot influence the intensity of the sanction. Punishing one person costs the punisher m and reduces the payoff of

[7] Counter-punishment is not allowed in all treatments. See the next subsection for treatment details.

[8] We choose a binary contribution decision because it is the simplest possible setup. More continuous contribution patterns need an additional assumption how the formal punishment scheme should deal with partial contributors. There are different possibilities, for example introducing a step-level or a linear decreasing fine (the latter is used by Markussen et al., 2011 and Kamei et al., 2011, the former for example in Tyran and Feld, 2006). As we do not expect any important insights from partial contributions, we exclude this option.

the punished subject by l. Thus, the stage-2-payoff of subject i in period t, for the case of informal punishments, can be described by

$$\pi_{i,t} = E - c_{i,t} + \frac{\gamma}{n}\sum_{j=1}^{n} c_{j,t} - m\sum_{\substack{j=1 \\ j\neq i}}^{n} p_{ji,t} - l\sum_{\substack{j=1 \\ j\neq i}}^{n} p_{ij,t}, \tag{1.2}$$

with $p_{ji,t}$ being a dummy variable equaling one if subject j is punished by subject i in period t, and zero otherwise. Accordingly, $p_{ij,t}$ describes punishment of subject i by subject j.

If the group voted for the formal institution, an automatic punishment mechanism replaces individual peer-to-peer punishment. This mechanism punishes each free-rider with a fine f but does not punish cooperators at all, i.e. the sanction $s_{i,t}$ is defined as $s_{i,t}(c_{i,t}) = f$ if $c_{i,t} = 0$ and $s_{i,t}(c_{i,t}) = 0$ if $c_{i,t} = E$. However, the institution is costly and reduces each group member's period payoff by z irrespective of whether and how many subjects free-ride. Stage-2-payoff in case of the formal institution is shown by

$$\pi_{i,t} = E - c_{i,t} + \frac{\gamma}{n}\sum_{j=1}^{n} c_{j,t} - z - s_{i,t}(c_{i,t}). \tag{1.3}$$

If counter-punishment is possible, an additional stage occurs only in the informal institution. In this stage, punished subjects get the chance to retaliate by counter-punishing their punishers. Importantly, it is not possible for subjects to sanction group members that did not punish them in stage 2 of the same period (to avoid delayed punishment). Subjects decide for each group member that punished them separately whether they want to exert counter-punishment on this person or not. As for the punishment instrument, counter-punishment is a binary decision and reduces payoffs of the punisher by m and of the punished person by l. Stage-3-payoffs of the informal institution will hence look as follows:

$$\pi_{i,t} = E - c_{i,t} + \frac{\gamma}{n}\sum_{j=1}^{n} c_{j,t} - m\sum_{\substack{j=1 \\ j\neq i}}^{n} p_{ji,t} - l\sum_{\substack{j=1 \\ j\neq i}}^{n} p_{ij,t} - m\sum_{\substack{j=1 \\ j\neq i \\ p_{ij,t}=1}}^{n} q_{ji,t} - l\sum_{\substack{j=1 \\ j\neq i \\ p_{ji,t}=1}}^{n} q_{ij,t} \tag{1.4}$$

The parameter $q_{ji,t}$ equals one if subject j is counter-punished by subject i in period t and zero otherwise ($q_{ij,t}$ captures counter-punishment of subject i by subject j). On the contrary, the formal institution is not affected by counter-punishment. Thus, equation (1.3) fully describes stage-3-payoffs of the formal institution.

As already mentioned, the voting decision in stage 0 determines whether the formal or informal institution is implemented and hence influences the punishment and counter-punishment possibilities within the group. Note that the voting stage appears every period and the voting outcome therefore applies only to the respective period. This allows us to study the impact of learning effects on institutional preferences. Voting itself is costless, mandatory and anonymous and subjects are not allowed to be indifferent. Aggregation of preferences ensues due to majority rule in our main treatments (including a random draw in case of a tie) and due to unanimity rule in the control treatments, where the informal institution is always implemented except all subjects voted for the formal institution. Subjects receive feedback on the chosen institution before the start of stage 1 but do neither know the exact voting outcome nor who voted for which institution.

1.3.2 Experimental design and procedure

In our experiment we set $E = 18$ tokens[9], $\gamma = 1.6$ and $n = 4$. Hence, we use a standard marginal per capita return (MPCR) of $\gamma/n = 0.4$ that meets the social dilemma requirements. We have $T = 10$ periods and randomly re-match subjects every period (i.e. stranger design). The re-matching is done to ensure that reputation effects and retaliation motives do not influence our results.[10] The latter is especially important as we explicitly study the effect of retaliation with our counter-punishment condition.

Regarding the punishment mechanisms we implement the following: Costs of the formal institution are set to $z = 3$ for each group member. Thus, the formal institution costs the group a total of twelve tokens whenever it is formed. Note that these fixed costs do not vary across treatments and that they are of perceptible size.[11] The sanctioning power of the formal punishment instrument differs across treatments. While the fine equals $f = 18$ in the strong condition, it is $f = 6$ in the weak condition. Hence, the whole investment into the private account is automatically taken away from the respective individual under the strong condition but only one third is confiscated under the weak condition. Whereas the former constitutes a

[9] Tokens are the experimental currency unit. They were converted into euro at the end of the experiment using the exchange rate of 30 tokens = 1 euro. Results from all periods were paid out.

[10] As we do not implement a perfect stranger design, both motives are, strictly speaking, not completely destroyed. However, the random re-matching combined with the fact that subjects never knew with whom they were playing should be strong enough to ensure that reputation and retaliation incentives do not affect behavior.

[11] Indeed, they engulf 27.78% of the maximal cooperation gain. In the literature we often find low or even no costs of formal sanctions (see for example Kosfeld et al., 2009; Putterman et al., 2011). As high fixed costs, e.g. for buildings, staff and equipment, are the main argument against the formation of such institutions in many real world applications, we concentrate on the case in which costs are perceptible.

deterrent punishment, the latter punishment is too weak to destroy free-riding incentives (the marginal rate of investing into the private account equals $1 - f/E = 2/3 > MPCR$).[12]

To make the informal institution comparable, we model it, in the absence of counter-punishment, as being at most as costly as the formal institution and at most as effective in terms of punishing free-riders. Subjects pay costs of $m = z/(n-1) = 1$ for each group member they want to punish. This means that individual (aggregate) punishment costs can rise up to a level of 3 (12) as under the formal institution if all punishment opportunities are exerted. However, there are no costs at all if no group member punishes. The strength of the punishment instrument is set to $l = f/(n-1)$ which equals 6 in the strong condition and 2 in the weak condition. Hence, parameter values ensure that punishment of free-riders can be at most as strong as under formal sanctions equaling the latter if a free-rider is punished by all other group members. If counter-punishment is possible, costs and punishment of free-riders in the informal institution can exceed those of the formal institution but only if subjects do actually engage in counter-punishment (parameters of the counter-punishment stage equal those of the punishment stage).

Due to the usage of punishment instruments, period earnings may become negative and subjects were warned that this possibility exists. Negative earnings have to be compensated with gains from other periods and with a lump-sum payment of 60 tokens that each subject receives before the start of period 1. We do not implement any limitation for (counter-) punishing in the informal institution once period earnings are negative. Furthermore, we try to keep the feedback as similar as possible across institutions and treatments. Thus, after the contribution stage, subjects always learn each group member's contributions to the public and private account as well as stage-1-payoffs. If the formal institution is chosen, subjects obtain thereafter individual information on each group member's costs of the institution, costs due to received punishment as well as the resulting period earnings. If the informal institution is chosen, subjects are told each group member's costs due to assigned punishment, costs due to received punishment and stage-2-payoffs. Additionally, subjects are informed about the persons that punished them.[13] The latter is done to make the feedback comparable to the case of counter-punishment in which this piece of information is necessary for retaliating. After the counter-punishment stage, subjects analogously learn each group member's costs due to assigned counter-punishment, due to received counter-punishment and stage-3-payoffs.

[12] Hence, there is no social dilemma in the strong condition but in the weak condition.
[13] Note that subjects are of course only informed about the group member's ID and not about the respective person in the room (i.e. anonymity is maintained).

By varying the strength of the punishment instrument (dimension 1) and the availability of counter-punishment in the informal institution (dimension 2) we obtain a 2 x 2 design of experimental treatments (see Table 1.1). However, we focus only on three of these treatments (S, W, S-CP) and do not implement the weak condition with counter-punishment possibilities. We expect the latter to be less interesting as weak punishment instruments should lead to low levels of individual peer-to-peer punishment and thus, per construction, leave only little room to counter-punish. Differences to the weak condition without counter-punishment should therefore be negligible. Instead we implement two control treatments for the case without counter-punishment (S-U, W-U) in which unanimity within the group is required for implementing the formal institution. Unanimity rule is a more realistic assumption for many real world applications due to the presence of exit options and it might increase cooperation in the formal institution, especially for the case of weak instruments, as it is common knowledge that all group members prefer formal punishments. Hence, we can check whether the voting rule affects individual's institutional choice.

Table 1.1: Main treatments

		Strength of punishment instrument	
		Strong	Weak
Counter-punishment possible	No	S	W
	Yes	S-CP	-

All treatments started with an incentivized social value orientation questionnaire, also known as ring test or the decomposed game technique, to control for behavioral types (see Appendix 1A for details of the questionnaire).[14] In this test subjects are randomly matched into groups of two and answer a set of decision tasks. In each task a subject chooses one out of two allocations that assign payoffs to herself and her matched partner. The partner stays the same over all tasks and answers the same set of questions (vice versa influencing the first person's payoff). The proposed allocations lie equally spaced on a circle around the origin and each task consists of choosing between two adjacent allocations. By summing up subject's choices for herself and her partner, we obtain a motivational vector whose angle to the origin is used to sort the subject into one out of eight behavioral types. Moreover, the length of the vector serves as a measure of consistency. This test gives us an independent

[14] The social value orientation questionnaire originally stems from psychology (see Liebrand, 1984 or earlier work by Griesinger and Livingston, 1973). In economic research it is used for example by Offerman et al. (1996), Park (2000), Brosig (2002), van Dijk et al. (2002) or Sutter et al. (2010).

measure of individual's general inclination to cooperate and thus allows controlling for self-selection of behavioral types in the public goods game.

Before the start of the questionnaire subjects received written instructions only for this part but they knew that there will be a second part of the experiment and that this part will be unrelated to the first one. After completion of the questionnaire which was given without any feedback to avoid biases due to income effects, subjects received instructions for the public goods game and a set of control questions. All instructions and control questions (to be found in Appendix 1B) were written in neutral language, read aloud and subjects were encouraged to ask if anything was unclear to them. Questions were answered privately and control exercises, after leaving enough time for studying, were solved aloud. At the end of the experiment, subjects learned their earnings from the ring test and answered a couple of socio-economic questions regarding gender, age, etc.

Overall, 144 students studying various disciplines took part in our main treatments (and 96 students in the control treatments). We had 48 participants per treatment, always distributed over two sessions with 24 persons. Within each session we randomly sorted subjects into units of twelve (without telling them) and only re-matched them within their matching group. Hence, we gained four statistically independent observations per treatment. Experiments were computerized using the software package z-tree (Fischbacher, 2007) and the recruiting system ORSEE (Greiner, 2004) and run in the MELESSA laboratory of the University of Munich in July 2011. Sessions lasted up to 75 minutes and subjects earned on average 15.1 euro including a 4 euro show-up fee.

1.4 Theoretical predictions

In this section we derive theoretical predictions according to standard preferences (Section 1.4.1) and discuss deviating behavioral aspects that are documented in the literature (Section 1.4.2). As we used a stranger design, reputation formation should not play any role in our treatments. Hence, we can interpret the finite length public goods game as consisting of a sequence of separate one-shot games. In the following, we derive predictions for our one-shot games in the different treatments.

1.4.1 Predictions based on standard preferences (homo oeconomicus - concept)

Assuming common knowledge of self-interest and rationality standard backward induction arguments apply within each one-shot game. If a group chooses the informal institution and counter-punishment is possible, no counter-punishment will occur as it is costly. Hence, the retaliation threat is not credible and does not influence prior behavior in the contribution or punishment stage. The cost argument also prevents any sanctioning in the punishment stage, irrespective of whether the punishment technology is weak or strong. Thus, whenever the informal institution is chosen, standard game theory predicts zero punishment [and zero counter-punishment]. Anticipating this, individuals have no incentive to contribute to the public account because the marginal per capita return is smaller than one. Full free-riding, i.e. $c_i = 0 \; \forall i$, is therefore the unique equilibrium behavior under informal punishment in all of our treatments yielding a payoff of $\pi_i = 18$ for each group member.

The situation differs for the formal institution because here the underlying punishment technology is crucial. If the punishment technology is strong, i.e. the entire investment into the private account is taken away by the mechanism, then it is a dominant strategy to contribute to the public account. Hence, $c_i = 18 \; \forall i$ and there is no punishment in equilibrium. Taking into account the costs of the institution, payoffs for each i equal $\pi_i = 0.4 \cdot 72 - 3 = 25.8$. If, however, the punishment technology is weak, i.e. only one third of the investment into the private account is confiscated, then it is a dominant strategy for each group member to free-ride and accept punishment. Choosing $c_i = 0$ is optimal for all i because the marginal rate of investing into the private account when being punished (2/3) is still larger than the marginal per capita return of the public account (0.4). Payoffs in this case are $\pi_i = 18 - 3 - 6 = 9$ for each group member.

How should individuals therefore vote in our treatments? If the punishment technology is weak as it is in treatments W and W-U, the formal institution leads to lower payoffs in equilibrium than the informal institution. Hence, there is a subgame perfect Nash equilibrium in which each group member votes for informal punishments. However, depending on the underlying voting rule, other vote combinations can also be part of equilibrium. For the majority rule, this includes a minority voting for formal punishments (resulting in the informal institution) and even the case in which each group member votes for formal punishments (the formal institution being implemented). Under unanimity rule all voting combinations with up to three votes for the formal institution (each resulting in the informal institution) can be part of equilibrium. Nevertheless, voting for the informal institution is always a weakly dominant strategy if the punishment technology is weak and thus, plausible

equilibrium refinements such as weak dominance or trembling-hand perfection (Selten, 1975) rule out equilibria that involve votes for the formal institution. The unique robust equilibrium, irrespective of whether we use majority or unanimity rule, implies that each group member votes for informal punishments and the informal institution is implemented. On the contrary, if the punishment technology is strong as it is in treatments S, S-CP and S-U, equilibrium payoffs are higher in the formal institution. Analogously, it is a weakly dominant strategy to vote for formal punishments and the unique robust equilibrium includes that each group member votes for formal punishments and the formal institution is implemented.[15]

To sum up, assuming standard preferences and a plausible equilibrium refinement, subjects vote for formal institutions if and only if the punishment mechanism is strong. Counter-punishment opportunities and the voting rule have no influence. Hence, theory predicts a difference in voting behavior between treatments S and W as well as S-U and W-U but no difference between S and S-CP.

1.4.2 Behavioral aspects

Predictions are less clear once behavioral components are considered. On the one hand, there are arguments why subjects might vote for the informal institution if punishment instruments are strong. Experimental evidence has shown that many individuals voluntarily contribute positive amounts to the public account and, moreover, do engage in costly informal sanctioning of free-riders making the punishment threat credible (e.g. Ledyard, 1995; Fehr and Gächter, 2000, 2002). Hence, the efficiency of the informal institution especially under the strong instrument should be higher than standard preferences predict. Theoretically, efficiency can even exceed the level under formal punishments, for example if enough group members are sufficiently inequity averse (see Fehr and Schmidt, 1999).[16] Other behavioral explanations for an increased efficiency in the informal institution include altruism (see e.g. Ledyard, 1995), warm glow (Andreoni, 1990), confusion (Andreoni, 1995; Palfrey and Prisbrey, 1997), preferences for social welfare (Charness and Rabin, 2002), reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004) or conditional cooperation (e.g. Fischbacher et al., 2001). There might also be an additional gain from endogenously choosing the informal institution (cf. Sutter et al., 2010) as participation rights can raise group

[15] Other subgame perfect Nash equilibria involve a minority or all group members voting for informal sanctions (majority rule) and two or more individuals voting against formal sanctions (unanimity rule), respectively.

[16] Appendix 1C shows that with Fehr and Schmidt preferences and a reasonable distribution of the inequity aversion parameters it is possible to sustain an equilibrium with higher payoffs in the informal institution if the punishment instrument is strong. This means that voting for the informal institution is an equilibrium strategy if at least some subjects are sufficiently inequity averse.

cooperation. Thus, subjects could be driven by the feeling that they can achieve cooperation without implementing a formal institution. Furthermore, individuals might have a general reluctance to vote for an automatic punishment mechanism that they cannot influence after the contribution decision. Especially the perceivable amount of fixed costs could deter subjects. Consequently, it is not obvious that individuals prefer the formal institution in case of a strong punishment mechanism.

On the other hand, there exist reasons to vote for the formal institution if punishment instruments are weak. Not only is the punishment threat much weaker under informal sanctions and free-riding less attractive in the formal institution given that not all group members punish (i.e. weaker conditions for other-regarding preferences can sustain cooperation in the formal institution, see for example the Fehr and Schmidt preferences in Appendix 1C), voting for the formal institution can also serve as a cooperation signal (Tyran and Feld, 2006). Because of the stranger design, there is no possibility to signal cooperativeness before the contribution decision except of voting for formal punishments. This might give reason to believe that contribution levels are higher if the formal institution is chosen. The strength of this effect could perhaps be influenced by the voting rule as unanimity creates a stronger signal for group identity. Hence, subjects might be more willing to vote for the formal institution under weak instruments if unanimity rule is applied.

To sum up, behavioral arguments cast doubt on the difference in voting behavior between treatments S and W as well as S-U and W-U the way it is predicted by standard theory.

Moreover, counter-punishment possibilities might play a role and lead to differences between treatments S and S-CP. We know from previous experiments that many subjects do retaliate if they get the chance to do so. Thus, this threat is, in contrast to predictions from standard preferences, indeed credible and reduces punishment activities and contributions to the public account.[17] The literature has shown that the latter tends to outweigh efficiency gains due to reduced punishment (Denant-Boemont et al., 2007; Nikiforakis, 2008). If this is true, it might raise subjects' inclination to vote for the formal institution. Furthermore, endogenous choice literature reports a strong reluctance of subjects to vote for informal institutions (at least in the beginning). By adding another stage of punishment to the informal setup which introduces further potential losses to the payoff function (consider the additional terms in equation 1.4), subjects might be even more willing to switch to the formal institution to avoid personal feuds. Therefore, formal institutions might be more attractive in treatment S-CP than in S.

[17] Note that counter-punishment in our experiment cannot be explained by Fehr and Schmidt preferences (see Appendix 1C). Instead it can be interpreted as a reciprocity-driven phenomenon.

1.5 Results

The result section is divided into four parts: aggregated information on the voting decision (Section 1.5.1) and the contribution and sanctioning behavior (Section 1.5.2) in our main treatments, a detailed analysis of the voting behavior (Section 1.5.3) as well as behavioral differences in our control treatments using unanimity rule (Section 1.5.4). We start with a first look at how individuals vote in treatments S, W and S-CP.

1.5.1 The voting decision

Table 1.2 reports results on the absolute number and the percentage of votes for the formal institution in our three main treatments. It shows that the effectiveness of the punishment instrument influences subject's voting decision. While the formal institution is favored by individuals in more than 50% of cases in treatment S, only 32.92% of votes are allotted to formal punishments in treatment W. This difference is highly significant ($p < 0.05$, two-sided Mann-Whitney-U-test (MWU-test), $N = 8$) and confirms the prediction from standard theory that stronger punishment instruments raise the demand for formal institutions. However, the difference is much smaller than the 0%-100% difference that standard theory predicts. Hence, behavioral arguments like social preferences play a role and reduce (increase) the attractiveness of the formal institution under the strong (weak) instrument. Interestingly, votes for the formal institution do not increase in treatment S-CP compared to treatment S. If anything, they are even a bit lower but the difference between the treatments is not significant (p-value = 0.89, two-sided MWU-test, $N = 8$). Hence, the introduction of counter-punishment to the informal institution does not lead to a higher demand for formal punishments.

Table 1.2: Votes for the formal institution by treatment

Treatment	Absolute number (percentage) of votes for formal institution
S	247 (51.46%)*
W	158 (32.92%)*
S-CP	225 (46.88%)

Note: * Significant difference between S and W ($p < 0.05$).

Note that the percentages shown in Table 1.2 are lower than what is reported so far in the literature for the comparison between the formal institution and the standard public goods game. This has mainly two reasons: First, we assume higher fixed costs of the formal institution (some of the existing studies do not impose any costs at all) and second, we allow for informal punishments as the alternative mechanism. However, our results show that even

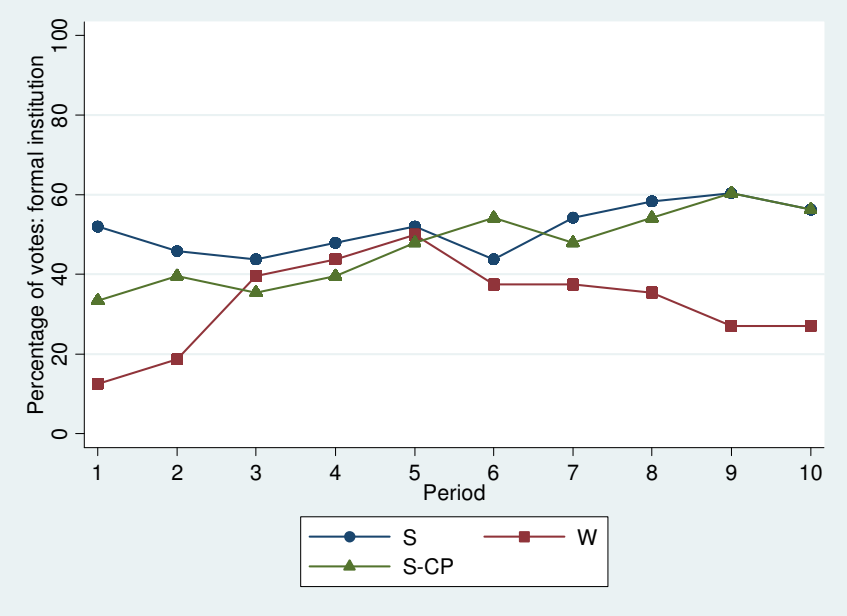
with this much more realistic setup, there is still a substantial demand for the formal institution. When comparing our results to Kamei et al.'s (2011) treatments with positive fixed costs (the size is comparable with ours), we find a higher support for formal punishments. In Kamei et al. (2011) only about 30% of votes favor the formal institution, the large majority of those preferring to implement severe sanctions. The difference can be explained by the fact that we use a weaker informal punishment mechanism.[18]

Result 1.1. *On average, formal institutions are significantly more attractive if the punishment instrument is strong than if it is weak. Counter-punishment has no effect on the average voting decision.*

If we consider the evolution of votes over time (see Figure 1.1), we find almost the same picture for treatments S and S-CP. Both graphs are slightly increasing and, moreover, converge to each other. While counter-punishment reduces votes for the formal institution in the first periods of S-CP compared to S (the difference is significant in period 1: $p = 0.06$, two-sided MWU-test, $N = 96$ and χ^2- test), this effect cancels out completely for later periods. Interestingly, the formal institution is far away from ever reaching full support. In fact, there is no single period in which more than 60% of subjects prefer the usage of formal punishments. Hence, we can claim that informal institutions do not die out over time, even if we would extend the number of periods in the experiment. Regarding treatment W, we find a hump-shaped pattern. In the beginning only very few subjects vote for formal punishments (significantly less than in treatment S: $p < 0.01$, two-sided MWU-test, $N = 96$ and χ^2- test). However, over time consent with the formal institution increases till it reaches the level of the other treatments in period 5. From this moment on more and more subjects switch back and vote for the informal institution again. This behavior raises the suspicion that subjects are unsatisfied with the formal institution once they experience it. We will come back to this issue in Section 1.5.3.

[18] Markussen et al. (2011) also have treatments with positive fixed costs that show even less support for formal punishments, but these treatments are less comparable to ours as the formal institution is more expensive and there are only two voting stages.

Figure 1.1: Votes for formal institution over time by treatment



Result 1.2. *Over time, we observe a hump-shaped demand for the formal institution in case of the weak punishment instrument. Both other treatments show an increasing trend but counter-punishment slightly reduces the attractiveness of the formal institution in the beginning.*

1.5.2 Contributions, sanctions and profits

So far, we concentrated on aggregated voting decisions and neglected the behavior within a chosen institution. Table 1.3 reports descriptive results regarding subjects' mean contribution and sanctioning behavior within the respective institution and adds information on the arising earnings. Results are shown for each treatment separately.

Looking at column 1 of Table 1.3, mean contributions in the formal institution of treatments S and S-CP are close to the maximum level of 18 showing that most subjects understand that contributing to the public account is the dominant strategy (we observe only 5 cases of zero contributions in treatment S and a slightly higher number of 15 in S-CP). In treatment W, contributions are significantly lower.[19] However, despite the zero contribution prediction from standard theory and in line with Tyran and Feld (2006), subjects contribute to the public account, on average, in more than 50% of cases. In the informal institution (cf. column 2), we find that mean contribution levels are quite high in treatment S but that introducing counter-punishment decreases mean contributions by three tokens and that there

[19] Significances in Table 1.3 are computed by random effects regressions (clustered for matching groups) that have the respective dummy variable as the only explanatory variable. Clustered probit regressions for contributions and sanctions yield the same results.

is almost no cooperation under the weak instrument in treatment W. The difference between S and W is highly significant whereas the difference between S and S-CP is not significant. Note that the latter effect is thus weaker than reported in Denant-Boemont et al. (2007) or Nikiforakis (2008).

Table 1.3: Contributions, punishment, counter-punishment and profits by institution and treatment

	Contribution			Punishment	Counter-Punishment	Profit		
	Formal	Informal	*All*	Informal	Informal	Formal	Informal	*All*
S	17.63*†	12.60*†	*15.11*†‡	22.22%$^{\dagger\#}$	-	25.20*†^	20.89*§	*23.05*†
	(N=240)	(N=240)	*(N=480)*	(N=720)		(N=240)	(N=240)	*(N=480)*
W	9.54*†	2.94*†	*4.31*†	11.14%†	-	17.90†	18.76§	*18.58*†
	(N=100)	(N=380)	*(N=480)*	(N=1140)		(N=100)	(N=380)	*(N=480)*
S-CP	16.59*	9.63*	*12.40*‡	9.14%$^{\#}$	58.23%	23.55*^	20.74*	*21.86*
	(N=192)	(N=288)	*(N=480)*	(N=864)	(N=79)	(N=192)	(N=288)	*(N=480)*

Notes: Mean contributions and profits presented. Punishment and Counter-Punishment are documented by percentages which equal the number of observed cases relative to all cases in which the sanction is possible. Significant difference between Formal and Informal: * $p < 0.01$; between S and W: † $p < 0.01$ and § $p < 0.10$; between S and S-CP: $^{\#}$ $p < 0.01$, ‡ $p < 0.05$ and ^ $p < 0.10$.

Overall, in each treatment mean contributions are significantly higher in the formal than in the informal institution. Higher contributions are even true for each single period in which formal institutions are observed (cf. Table 1D.1 in Appendix 1D for implementation frequencies over time and Figure 1D.1 in Appendix 1D for the evolution of contributions). Hence, formal punishments are always cooperation-improving. Column 1 of Table 1.4 shows a linear model for contributions which is estimated by OLS using robust standard errors clustered on the matching group level. We include a dummy variable for the formal institution (*Formal*), two treatment dummies (*W*, *S-CP*) and the respective interaction effects (*Formal * W*, *Formal * S-CP*) and find that both interaction effects are insignificant.[20] This means that we cannot claim that the formation of the formal institution has different effects in treatments W and S-CP compared to S. This is interesting as standard theory predicts a stronger effect in S and S-CP compared to W. On the other hand we also have no support for the behavioral argument that the effect in S-CP is larger than in S. Note further that we get the same results when we control for subject's behavioral type as obtained out of the social value orientation questionnaire. Hence, results on cooperation are not driven by self-selection of more cooperative subjects.

[20] The reference category is the informal institution in treatment S. We do not report a probit regression because interaction effects cannot be interpreted properly in nonlinear models (see Ai and Norton, 2003 for a discussion). Note that no single fitted value lies outside the interval [0, 18]. A random effects regression yields very similar results.

Table 1.4: Explaining contributions and profits

	Dependent variable:	
	Contribution	Profit
Formal (= 1)	5.025***	4.307***
	(1.074)	(1.339)
W (= 1)	-9.663***	-2.134*
	(1.000)	(1.098)
S-CP (= 1)	-2.975	-0.157
	(2.211)	(1.944)
Formal * W	1.578	-5.162***
	(1.188)	(1.436)
Formal * S-CP	1.944	-1.493
	(1.956)	(1.778)
Constant	12.600***	20.893***
	(0.860)	(1.007)
N	1440	1440
R^2	0.376	0.140

Notes: *** Significant at 1% level; ** significant at 5% level; * significant at 10% level. OLS regressions. Robust standard errors in parentheses (clustered on matching group level).

Figure 1D.1 in Appendix 1D reveals that contributions decline over time in both institutions of treatment W (very severe in the formal institution) as well as in the informal institution of S-CP. On the contrary, there is only a small decline in the informal institution of treatment S and full cooperation is a stable phenomenon in the formal institutions of treatments S and S-CP. Hence, in the long run a high level of cooperation is only achievable with the strong formal institution. Finally, taking endogenous choice into account, column 3 of Table 1.3 reports, on average, high cooperation in treatment S (83.94% of endowment invested), a significantly lower level in S-CP (68.89%) and very little cooperation in W (23.94%).[21]

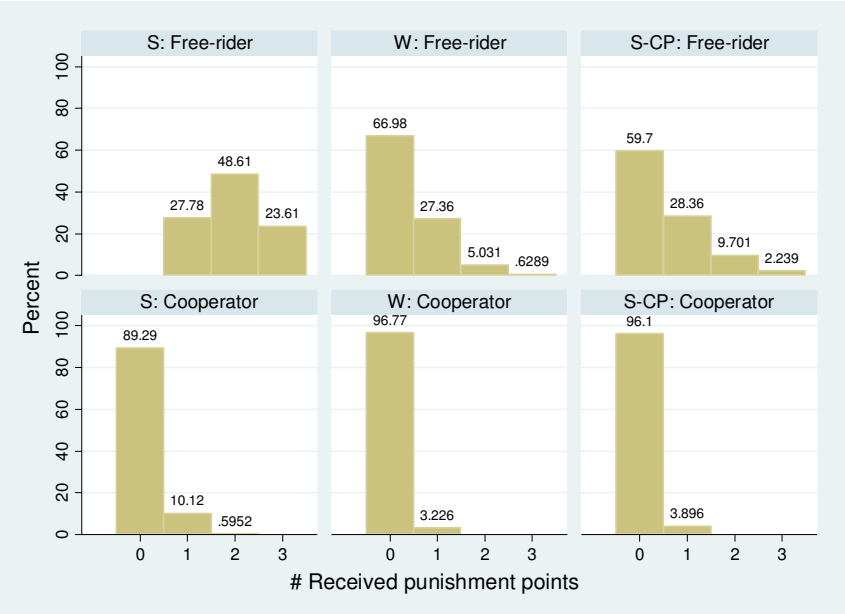
Result 1.3. *In all treatments, mean contributions are significantly higher in the formal than in the informal institution. However, the level in the formal institution of W is lower than in the other treatments and contributions decrease over time.*

Regarding subject's sanctioning behavior in the informal institution (cf. Table 1.3, columns 4 and 5 as well as Figure 1D.2 in Appendix 1D) we can, not surprisingly, show that punishing appears significantly less frequent in case of a weak punishment instrument

[21] A significant difference between treatments S and W can also be shown with a two-sided MWU-test ($p < 0.05$, $N = 8$). However, this test yields no significance for the difference between S and S-CP ($p = 0.25$, $N = 8$).

(treatment W) and in case of a retaliation fear (treatment S-CP) compared to treatment S. Moreover, counter-punishment is observed, on average, in almost 60% of cases and is thus a very credible threat for our subjects. The large percentage of retaliation cases is remarkable if one takes into account that almost all the counter-punishing subjects are free-riders. Figure 1.2 shows that cooperators in treatment S-CP are hardly ever punished and thus have little chances to counter-punish. Hence, counter-punishment in our experiment can be seen predominantly as punished free-riders sanctioning cooperators.[22] Owing to the stranger design, this is not caused by strategic reasons but has to be explained by reciprocal motivations. We also find very little evidence for perverse punishment in the other treatments (at least less than reported for instance in Herrmann et al., 2008 or Nikiforakis, 2008).

Figure 1.2: Number of punishment points received by treatment and contribution type



The probability with which a free-rider is punished varies widely across treatments. Figure 1.2 illustrates that there is no single case in our data in which a free-rider is not punished by at least one group member in treatment S. Moreover, the median free-rider receives two punishment points. This reduces her payoff by 12 tokens which is enough to make free-riding non-profitable. In treatment W, however, 2/3 of free-riders are not punished at all and two or more punishment points are unlikely. Almost the same picture appears in

[22] Note that they are not sanctioning other free-riders as the latter rarely punish in the punishment stage.

treatment S-CP where punishment frequencies for free-riders are only slightly higher.[23] This means that free-riding pays, on average, for individuals in the informal institutions of W and S-CP.

Result 1.4. *Punishment is less often observed in the informal institutions of W and S-CP compared to S. The counter-punishment instrument is used frequently although almost all punished subjects are free-riders.*

Finally, columns 6-8 of Table 1.3 show mean profits in the different treatments. Column 6 reports results for the formal institution revealing a significantly lower payoff in treatment W than in S as well as a small difference between S and S-CP. In the informal institution (column 7), profits are also significantly lower in W compared to S ($p < 0.10$) and this difference becomes much more significant once period 1 is excluded ($p = 0.03$; cf. Figure 1D.3 in Appendix 1D for the time trend). Interestingly, mean profits are almost equal in S and S-CP, i.e. the introduction of counter-punishment does not reduce earnings in the informal institution. Hence, there is no decline in contributions that outweighs efficiency gains due to reduced punishment activities. This might be an explanation why we do not observe more votes for the formal institution once counter-punishment possibilities exist. Moreover, Figure 1D.3 in Appendix 1D shows that profits in the informal institution of treatment S-CP are even higher than under S for the first two periods (although the difference is not significant). This phenomenon, if anticipated by individuals, can explain the lower percentage of votes for the formal institution in treatment S-CP in the beginning. It seems to be the case that subjects understand that the counter-punishment instrument decreases punishment activities but they do not foresee its consequences on contribution levels. Figure 1D.1 in Appendix 1D confirms that, in the beginning, contribution levels are not lower in the informal institution of S-CP than in S. As over time subjects learn the effect on contribution levels, the difference in institutional votes vanishes.

Furthermore, we get an idea why there is this difference in institutional votes between treatments S and S-CP on the one hand and W on the other hand. While in the first two treatments mean profits are significantly higher in the formal than in the informal institution this is not true for the treatment W. Here, the formal institution even leads to slightly lower profits because the average contribution level of 9.54 tokens is not enough to compensate for

[23] Note that a free-rider is slightly more heavily punished in treatment S-CP than in W although Table 1.3 reports lower punishment percentages. The difference is due to higher contribution levels and hence a lower number of free-riders in S-CP.

the fixed costs of the institution. Or put differently: the number of free-riders in the formal institution under a weak punishment mechanism is too large to make it a profitable institution. An OLS regression on profits using cluster-robust standard errors (cf. column 2 in Table 1.4) confirms that the formal institution has a worse effect on profits in treatment W than in S as the interaction effect *Formal * W* is significantly negative. On the contrary, *Formal * S-CP* is not significant meaning that there is no additional gain or loss in terms of profits of forming formal institutions in treatment S-CP.[24]

The time trend (see Appendix 1D) further reveals that profits in the formal institution of W decrease severely. The longer the time horizon is the worse formal punishments perform as contributions approach zero and each individual bears the fixed costs of the institution and suffers the automatic punishment. What remains unclear for the moment is why we never observe more than 60% of subjects voting for the formal institution in treatments S and S-CP although this institution is clearly more profitable. We will look at this question in the next subsection. Finally, column 8 of Table 1.3 reports overall profits in our three treatments showing significantly lower profits in treatment W compared to S (also confirmed by a two-sided MWU-test: $p < 0.05$, $N = 8$) but no significant difference between S and S-CP.

Result 1.5. *The formal institution is only profitable in treatments S and S-CP. In W, mean contributions are too low to outweigh institutional costs. We find no difference in profits between the informal institutions of S and S-CP.*

1.5.3 Explaining subject's voting behavior

Having analyzed the aggregated data, we now turn our attention more closely to the individual behavior and ask the question what drives a subject's voting decision within a given treatment. We distinguish between the voting decision in period 1 (without having any experience) and decisions in later periods that will depend on the history of the game.

Starting with period 1 one may wonder whether differences in subject's general willingness to cooperate (i.e. their behavioral type) can explain the decision for formal or informal institutions. To control for this, we included a social value orientation questionnaire (ring test) in the experiment that enables us to sort subjects into different behavioral categories. Figure 1D.4 in Appendix 1D shows the distribution of types for each treatment. Note that we focus only on subjects with a consistency ratio of at least 2/3 which excludes 7

[24] A random effects regression yields the same results.

subjects (4.86%) from the analysis.[25] It is obvious that, as usual in public goods games, only two motivations matter: individualism and cooperation. Indeed, there is only one subject choosing differently by acting altruistic. Moreover, the type distribution does not differ across treatments ($p = 0.87$, χ^2-test) and hence does not drive the observed treatment effects. But do individualistic (i.e. more selfish individuals) vote differently than cooperative subjects? Columns 1 and 2 of Table 1.5 confirm that this is not the case. The relative number of individuals that vote for the formal institution does not differ between types (χ^2- and Fisher's exact tests yield insignificant results for each treatment: all $p > 0.35$).[26] Hence, self-selection of behavioral types does not explain the voting behavior in period 1. There are as many individualistic as cooperative subjects who are willing to vote for the formal institution and thus, at least in treatments S and S-CP, willing to bind themselves to contribute. We can conclude that selfish individuals are not the general problem behind the non-implementation of formal punishment systems. This finding coincides with Putterman et al. (2011) who state that even completely selfish individuals have incentives to implement formal institutions to overcome free-riding motives and gain efficiency.[27]

Table 1.5: First period's voting separated for subject's behavioral type and gender

		Individualistic	Cooperative	*All*	Men	Women	*All*
S	Formal	15	8	*23*	12	13	*25*
	Informal	15	7	*22*	10	13	*23*
	All	*30*	*15*	*45*	*22*	*26*	*48*
W	Formal	3	3	*6*	4	2	*6*
	Informal	26	13	*39*	17	25	*42*
	All	*29*	*16*	*45*	*21*	*27*	*48*
S-CP	Formal	11	5	*16*	10	6	*16*
	Informal	21	9	*30*	11	21	*32*
	All	*32*	*14*	*46*	*21*	*27*	*48*

Note: 8 subjects excluded in the type classification part due to other motivations or inconsistent answers in the ring test.

[25] Subjects with a low level of consistency cannot be classified unambiguously. There is no standard practice of which consistency ratio to demand. While Park (2000) requires a ratio of 75%, Brosig (2002) already classifies subjects with 25% consistency. We use the common threshold of 2/3 but none of our results would change by using a higher or lower ratio.

[26] There is enough variation in subjects' motivational degrees (as obtained out of the ring test) between the two types to concentrate on categories. Nevertheless, insignificance ($p > 0.23$ in each treatment) can also be shown by probit regressions on subject's vote that use the exact degree of subject's motivation as explanatory variable.

[27] We can concentrate on purely selfish individuals in our data by looking only at subjects with a motivational degree of zero. Out of those we observe 5/15 voting for formal sanctions in treatment S, 1/14 in W and 7/16 in S-CP. Hence, we also observe votes for the formal institution within this subgroup. Again, no distribution is significantly different from the cooperative type (all p-values of χ^2- and Fisher exact tests > 0.23).

We also take a close look at gender as gender was found to have a strong impact on the endogenous choice between informal institutions and the standard voluntary contribution mechanism (see Chapter 2).[28] In this chapter, it is shown that women shy away from the informal punishment institution by preferring the standard voluntary contribution mechanism much more often than men do. Hence, one might expect a similar movement towards the formal punishment institution in our experiment as the formal institution resembles the standard public goods game in many respects: there is no punishment decision to take and you can avoid being punished for sure. However, if we look at the data for the first period, we do not find any gender effect in treatments S and W (cf. columns 4 and 5 of Table 1.5). In treatment S-CP, there is a difference but it is reversed as women *less* likely vote for the formal institution. Note that this difference is weakly significant ($p = 0.06$, χ^2- test; $p = 0.07$, probit regression). Hence, we do not find any evidence that women prefer formal sanctions more than men if they have no experience with the social dilemma. If anything, the tendency goes even in the opposite direction.[29]

Result 1.6. *First period's voting behavior is neither influenced by subject's behavioral type (as obtained out of the social value orientation questionnaire) nor in a systematic way by gender.*

In the next step, we want to have a close look at how often and in which direction individuals change their vote during the course of the experiment. Table 1.6 shows that in treatment S 14 out of 48 subjects (29.17%) do not change their vote at all, i.e. they vote for the same institution ten times in a row. Interestingly, only 5 subjects (10.42%) always prefer the informal institution. Put it the other way around: almost all subjects are willing to implement formal punishments in at least one period. This is clear evidence that even in the case of perceivable fixed costs there is no large group of subjects that completely denies the formal institution. Moreover, one half of the individuals in S switches their vote only once or twice and can therefore be considered as voting persistently. But Table 1.6 illustrates that these switches go in both directions. The reason why aggregate vote outcomes for the formal institution never exceed 60% is therefore due to a large group of individuals that vote for formal punishments somewhere during the experiment but switch (back) to prefer the

[28] Note that there is almost no correlation between gender and individual's type in our experiment ($\rho < |0.1|$ in each treatment). Independence between both variables cannot be rejected ($p > 0.50$ in each treatment, χ^2- tests).
[29] We checked also for other personal characteristics like age, field of study or experimental experience. However, none of them has a systematic influence on the voting decision, neither in period 1 nor in later periods. We therefore neglect them in the analysis.

informal institution afterwards. Finally, roughly a quarter of subjects (cf. column 4) switches more than two times. The average switching rate per person is 1.67.

Table 1.6: Number and direction of switches in individual's voting behavior over time by treatment

		No switch	One switch leading to…	Two switches leading to…	> Two switches leading to…	*All*
S	Formal	9	7	4	7	*27*
	Informal	5	5	7	4	*21*
	All	*14*	*12*	*11*	*11*	*48*
W	Formal	0	3	0	10	*13*
	Informal	14	3	10	8	*35*
	All	*14*	*6*	*10*	*18*	*48*
S-CP	Formal	11	8	2	6	*27*
	Informal	8	2	4	7	*21*
	All	*19*	*10*	*6*	*13*	*48*

In treatment W, a much larger fraction of subjects (14/48 = 29.17%) is never willing to vote for the formal institution and there are many subjects that switch back to the informal institution after having voted for formal punishments before. Almost all of the individuals that vote for the formal institution in the end have switched their opinion several times and thus cannot be interpreted as being entirely convinced by the institution. On average, the switching rate per person is 2.06 which is higher than in S. In treatment S-CP, almost 40% of subjects never switch vote but clearly prefer one institution (both institutions being observed). One quarter of subjects switches several times and the rest only once or twice. Note that switches go in both directions like in treatment S. The average switching rate of 1.65 is almost the same as in S.[30]

Result 1.7. *Over time, subjects switch their vote rather rarely. Interestingly, in treatments S and S-CP we do not only observe subjects switching from the informal to the formal institution as the latter is more profitable but also quite a number of subjects that switch the other way around.*

Finally, Table 1.7 explains individual's voting behavior in periods 2 to 10 by treatment variables, personal characteristics and experiences in the previous period. We report evidence

[30] We also checked for the timing of switches but no important trend can be observed. When categorizing switches in early (periods 1/2 - 3/4), middle (4/5 - 6/7) and late (7/8 - 9/10) we find a weak increase in treatments S (24, 27 and 29 switches, respectively) and W (29, 35, 35) and a hump-shaped pattern in S-CP (23, 34, 22).

from four linear probability models (cf. models 1 - 4) that are estimated by OLS using robust standard errors clustered on the matching group level.[31] Model 1 includes the treatment dummies *W* and *S-CP* (*S* serves as reference category), the gender dummy *Woman*, the type dummy *Type* (being 1 for cooperative and 0 for individualistic subjects) and *Period* as well as interaction terms between the latter three variables and treatment. Note that as before we exclude 8 subjects with other motivations or inconsistent answers in the ring test. Column 1 reveals that gender has no significant impact on institutional choice in periods 2 - 10. Even for treatment S-CP where we found a weak effect in period 1, the combined effect of *Woman* + *Woman * S-CP* is not significantly different from zero ($p = 0.38$). As signs indicate, on average it is still the case that more men than women vote for the formal institution in S-CP. While this is also true for treatment S, the fact is only slightly reversed for treatment W. Hence, there is no evidence that women shy away from informal punishment institutions when formal punishment is the alternative. Regarding the type classification we do not find any effect for treatments S and W but a small impact in treatment S-CP. The interaction effect *Type * S-CP* is weakly significant ($p = 0.06$) and the combined effect *Type* + *Type * S-CP* is positive and also significantly different from zero ($p < 0.10$). Both results become stronger if we focus only on periods 3-10. Hence, there is a tendency that cooperative individuals prefer the formal institution more than individualistic subjects in treatment S-CP. Remember that there was no effect at all in period 1. This self-selection effect only emerges over time as cooperative individuals contribute to the public account, punish free-riders and suffer from counter-punishment. It is this retaliation experience that drives those subjects more likely into the formal institution. Not astonishingly, model 1 also finds a significantly different time trend in treatment W compared to S and S-CP. This is in line with what we have seen in Figure 1.1.

Model 2 drops the treatment effects introduced by model 1 and instead focuses on six lagged variables that control for subject's experience.[32] We use a dummy variable that takes into account whether a subject voted for the formal institution in the previous period (*Institutional choice (t-1)*). Further, *Informal (t-1)* indicates whether a subject played the informal institution in the previous period or not. *AvgContribution_others (t-1)* captures the

[31] Again, we apply a linear probability model rather than a probit model to avoid the difficulties in interpreting the interaction effects (see Ai and Norton, 2003). Less than 5% of the fitted values lie outside the [0, 1] range in each of the regressions (in model 1 even 0%) and results are close to the probit specification. We further estimated all models as random effect specifications and find very similar results. Moreover, we tried a linear, dynamic panel data estimation method (Arellano and Bond, 1991) for the time-varying variables. Effects comparable to those of OLS can be found for most of the covariates including the lagged dependent variable. Hence, the linear probability model seems to be an appropriate choice. Finally, note that including up to three lags in models 2-4 does not change our results.

[32] See Fischbacher and Gächter (2010) for the use of OLS with respect to a dynamic model on beliefs.

mean contribution of the three group members a subject was matched with in the period before and *Informal (t-1) * AvgContribution (t-1)* is the respective interaction effect. Finally, sanctioning behavior in the informal institution is considered by including the interaction effects *Informal (t-1) * Punished (t-1)* and *Informal (t-1) * CounterPunished (t-1)* which equal one if a subject played the informal institution in the previous period and was sanctioned with the respective instrument by at least one group member.[33] Results show that *Woman* and *Type* do not play any role in the pooled model and that *Period* loses significance compared to model 1 which is not surprising because we do not account for different time trends across treatments. However, all of our lagged variables are highly significant and explain institutional choice to a large extend (see the quite high R^2). Much of the observed voting behavior can be explained simply by *Institutional choice (t-1).* There is a large amount of persistency in the data as many subjects only switch their vote rarely or even never at all (cf. also Table 1.6). Moreover, contribution behavior matters. While higher average contributions of the other group members in the formal institution increase the probability of voting for formal punishment (*AvgContribution_others (t-1)* is significantly positive) the contrary is true for the informal institution. Here, higher contribution levels of group members decrease the probability that a subject votes for the formal institution (see the large negative coefficient of *Informal (t-1) * AvgContribution_others (t-1)*).[34] Furthermore, the sanctioning behavior in the informal institution matters. Both being punished and being counter-punished significantly increases the probability that a subject votes for the formal institution in the next period. Coefficients show that the latter effect is even more important.[35]

Model 3 combines all the explanatory variables used in model 1 and model 2 and confirms previous results. Controlling for subject's experience even increases the significance of *Type * S-CP* and *Type* + *Type * S-CP* ($p < 0.05$ in both cases). Moreover, the interaction effect *Woman * W* becomes weakly significant ($p = 0.06$) confirming the result mentioned above that the direction of votes is slightly reversed in treatment W compared to S and S-CP. Note further that the time trends are still highly significant once we control for experience. This indicates that subjects do not only look at the previous period when forming their belief about future cooperation.

[33] Note that being counter punished is of course only possible in treatment S-CP.
[34] Indeed, the combined effect of *AvgContribution_others (t-1)* + *Informal (t-1) * AvgContribution_others (t-1)* is significantly negative ($p < 0.01$).
[35] This holds also if we focus only on treatment S-CP.

Table 1.7: Explaining individual's voting behavior

Dependent variable: Institutional choice (1 = Formal)				
	Model 1	Model 2	Model 3	Model 4
W (= 1)	-0.115	-	-0.043	-0.008
	(0.100)		(0.054)	(0.043)
S-CP (= 1)	-0.140	-	-0.061	0.004
	(0.162)		(0.056)	(0.054)
Woman (= 1)	-0.051	-0.016	-0.032	-0.029
	(0.082)	(0.022)	(0.022)	(0.020)
Type (1 = Cooperative)	-0.079	0.021	-0.038	-
	(0.080)	(0.028)	(0.029)	
Period	0.017**	0.004	0.008***	0.009***
	(0.007)	(0.003)	(0.001)	(0.002)
Woman * W	0.077	-	0.058*	0.057
	(0.097)		(0.031)	(0.036)
Woman * S-CP	-0.095	-	-0.030	-0.036
	(0.180)		(0.053)	(0.065)
Type * W	0.138	-	0.060	-
	(0.106)		(0.046)	
Type * S-CP	0.282*	-	0.152**	-
	(0.136)		(0.049)	
Period * W	-0.024**	-	-0.026***	-0.026***
	(0.011)		(0.006)	(0.007)
Period * S-CP	0.013	-	0.001	-0.001
	(0.009)		(0.005)	(0.005)
Institutional choice (t-1) (1 = Formal)	-	0.652***	0.636***	0.617***
		(0.031)	(0.031)	(0.035)
Informal (t-1) (= 1)	-	0.545***	0.462***	0.447***
		(0.087)	(0.076)	(0.077)
AvgContribution_others (t-1)	-	0.026***	0.015**	0.016**
		(0.004)	(0.005)	(0.005)
Informal (t-1) * AvgContribution_others (t-1)	-	-0.036***	-0.032***	-0.031***
		(0.006)	(0.005)	(0.005)
Informal (t-1) * Punished (t-1)	-	0.076**	0.085***	0.070***
		(0.026)	(0.025)	(0.022)
Informal (t-1) * CounterPunished (t-1)	-	0.184***	0.145**	0.103***
		(0.048)	(0.065)	(0.026)
Constant	0.465***	-0.315***	-0.094	-0.109
	(0.066)	(0.055)	(0.090)	(0.084)
N	1224	1224	1224	1296
R^2	0.057	0.427	0.443	0.405

Notes: *** Significant at 1% level; ** significant at 5% level; * significant at 10% level. OLS regressions using data from periods 2-10. Robust standard errors in parentheses (clustered on matching group level). 8 subjects excluded in models 1-3 due to other motivations or inconsistent answers in the ring test.

Model 4 finally replicates model 3 for the full data set, i.e. it excludes the type dummy. This specification shows that our results do not hinge on the restricted data set. The only difference is that *Woman * W* is insignificant (p = 0.11) and hence less robust.[36]

The regression results fit nicely some of the findings reported by Markussen et al. (2011) and Kamei et al. (2011). Both papers also control for gender and subject's general willingness to cooperate and find no significant effects. Note, however, that they do not use the ring test measure but apply either the first period's contribution in a standard public goods game or the conditional contribution schedule (Fischbacher et al., 2001). As they are not focusing on counter-punishment they naturally cannot observe self-selection to matter in the context of retaliation. Further, they also find that votes for the formal institution depend on past behavior, especially, on whether a person was punished in the informal institution before. This effect is even more obvious in their setting than in ours as they use a partner matching. In contrast to our study they do not report a significant time trend in the voting behavior. However, this is most likely caused by their low number of voting periods. We summarize our regression results as follows.

Result 1.8. *Some self-selection emerges over time in treatment S-CP where cooperative individuals vote more likely for the formal institution. There is no such effect in the other treatments and we also find no gender effect. Voting behavior can be explained by the history of the game. Especially, low average contribution levels in the informal institution and the experience of being punished or counter-punished motivates subjects to vote for the formal institution.*

Can our regressions explain the descriptive phenomenon that many subjects in treatments S and S-CP switch back from voting for formal punishments in period t to preferring the informal institution in period $t + 1$? The answer is yes and no. Such behavior is partly due to the (rare) experience of low average contribution levels in the formal institution and to a well-functioning of the informal institution (the latter can appear although a subject votes for formal punishments due to majority rule as decision criterion). However, there is also an explanation that goes beyond the regression models: if the formal institution is implemented subjects become more aware of the fact that they cannot punish group members deliberately.

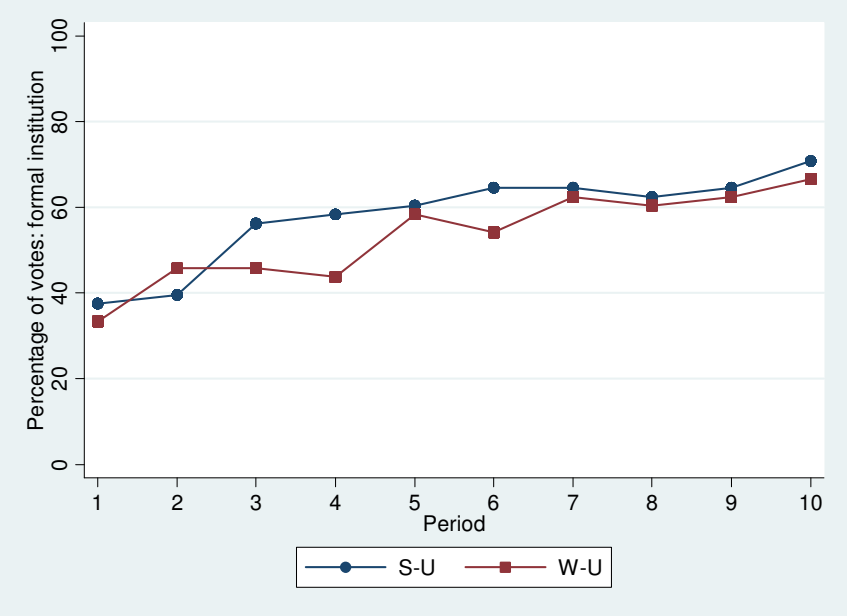
[36] As a minor change note that *Informal (t-1) * CounterPunished (t-1)* becomes more significant. This is due to the fact that we omit *Type * S-CP* which partly measures the same effect as both variables account for cooperative subjects behaving differently in treatment S-CP due to the experience of counter-punishment. Note, however, that *Type * S-CP* has a broader scope as it is not restricted to the previous period and can also account for observed retaliation within the group.

This lack of decision power seems to be an additional driver for subjects to switch back preferring the informal institution and is even more relevant for treatment S where there is no danger of being counter-punished.

1.5.4 Behavioral differences under unanimity rule

To control for the fact that our voting rule might drive the results on cooperation and, hence, the treatment effect between S and W we conduct two control treatments S-U and W-U using unanimity instead of majority rule. In the following we only report selected aggregate data for these treatments. As Figure 1.3 illustrates, the development in S-U is similar to what we have seen for treatment S (see Figure 1.1). However, the time trend in treatment W-U is clearly different from the one observed in W. While support for the formal institution already starts at a much higher level it, more importantly, does not decrease in later periods but proceeds parallel to the development in S-U (two-sided MWU-tests confirm that the difference between both treatments is not significant in any period).[37]

Figure 1.3: Votes for formal institution under unanimity rule



Nevertheless, Table 1.8 shows that average contributions in the formal institution of W-U are only slightly higher than in W (cf. again Table 1.3; the difference is not significant: p =

[37] The difference between W-U and W in period 1 is significant ($p < 0.05$ with two-sided MWU-test, N = 96 and χ^2-test) and can be explained by a subject's belief that cooperation in the formal institution is easier to achieve with unanimity rule because all group members have to agree on the institution. On the contrary, the difference between S-U and S in period 1 is not significant.

0.54) and therefore do not lead to significantly higher profits in the formal than in the informal institution.[38,39] Consequently, the lack of a decrease in votes for the formal institution in W-U is not due to a higher level of cooperation in the formal institution once unanimity is demanded but is rather driven by the fact that there is too little experience with formal punishments. Indeed, from the few observations we have 30% of subjects do not vote for the formal institution again after experiencing it, a number similar to what we find for subjects preferring formal punishments in the first periods of treatment W. However, we find a large demand for formal punishments caused by subjects without any experience with the formal institution. Hence, the unanimity rule itself does not seem able to stabilize a high demand for formal institutions in the long run when sanctioning instruments are weak.[40]

Result 1.9. *Using unanimity instead of majority rule does not lead to a higher level of cooperation when the weak formal institution is formed. The observed parallelism in voting behavior between the weak and the strong punishment scenario under unanimity rule is rather due to low implementation rates (little experience) with the formal institution when instruments are weak.*

Table 1.8: Contributions, punishment and profits by institution and treatment

	Contribution			Punishment	Profit		
	Formal	Informal	*All*	Informal	Formal	Informal	*All*
S-U	17.63*†	8.79*†	*9.68*†	29.86%†	25.20*†	17.00*§	*17.82*
	(N = 48)	(N = 432)	*(N = 480)*	(N = 1296)	(N = 48)	(N = 240)	*(N = 480)*
W-U	10.50*†	2.54*†	*3.34*†	8.64%†	18.80†	18.75§	*18.75*
	(N = 48)	(N = 432)	(N = 480)	(N = 1296)	(N = 48)	(N = 432)	*(N = 480)*

Notes: Mean contributions and profits presented. Punishment is documented by percentages which equal the number of observed cases relative to all cases in which the sanction is possible. Significant difference between Formal and Informal: * $p < 0.01$; between S-U and W-U: † $p < 0.01$ and § $p < 0.10$.

[38] Significances in Table 1.8 are computed by random effects regressions (clustered for matching groups) that have the respective dummy variable as the only explanatory variable. Clustered probit regressions for contributions and punishments yield the same results. The difference in column 3 of Table 1.8 is also significant using a two-sided MWU-test ($p < 0.05$, $N = 8$).

[39] Contributions also decrease over time which means that accounting for different implementation frequencies there is most likely no difference in contributions at all.

[40] One further aspect of the voting rule is noteworthy: As Table 1.8 reveals, overall profits in treatment S-U are much lower than in treatment S (cf. Table 1.3) and this difference is highly significant using either a random effects regression ($p < 0.01$) or the MWU-test ($p < 0.05$, $N = 8$). Consequently, the unanimity rule destroys welfare compared to the majority rule as it complicates the formation of formal institutions. This effect does not matter for the weak instrument because in this case there is no efficiency gain in forming the formal institution.

1.6 Conclusion

We report experimental evidence on subject's voting decision between informal and formal punishment institutions in social dilemmas. More precisely, we let subjects vote repeatedly in the course of a multi-period game and exogenously vary the strength of the available punishment instrument (correspondingly in both institutions) and whether or not counter-punishment is possible in case the informal institution is implemented.

Our results show that formal punishment institutions are more often implemented if the available instrument is strong. While the willingness to establish formal sanctions also emerges under the weak condition, the lower contribution level if implemented reduces their attractiveness in later periods. Interestingly, we do not find an effect of counter-punishment on subject's average voting decision. The retaliation option influences voting behavior only in the beginning by slightly decreasing votes for the formal institution. In all treatments individual's voting behavior shows a high level of persistency and switches can be explained by group members' past contribution and sanctioning behavior. Self-selection of behavioral types does not play any role in the absence of counter-punishment but over time cooperative subjects vote more likely for the formal institution when retaliation is possible. Socio-economic facts like gender do not explain the voting pattern systematically. Finally, we report two control treatments using unanimity instead of majority rule to show that contributions in the formal institution do not depend on the voting procedure and, hence, no difference should be expected in the long run.

This study is among the first that tries to understand the formation of formal punishment institutions in a more realistic setup using informal punishment environments as the alternative institution and not the standard linear public goods game. In contrast to parallel papers that use this approach (Kamei et al., 2011 and Markussen et al., 2011) we create more comparable institutions by directly connecting costs and effectiveness of the informal and formal punishment mechanism. As this design is well-balanced, other aspects that might influence subject's institutional choice (e.g. communication, feedback, group size) could be studied within the framework.

To sum up, in line with Thomas Hobbes' (2008 [1651]) theory we find a preference for the assignment of punishment rights to a formal authority. Moreover, this chapter provides evidence that the institutional environment is a critical determinant of how likely formal institutions are formed. As an application our research might explain why formal institutions in the international arena (such as the Kyoto protocol or the EU Stability and Growth Pact)

have such a hard standing. Attempts to implement global rules often suffer from the fact that it is difficult to find strong but appropriate sanctioning instruments. Our results show that this drawback can severely hamper the formation process. Furthermore, this chapter suggests that the counter-punishment threat present at the international level does not foster the formation of formal institutions.

Appendix

1A Social value orientation questionnaire (ring test)

The social value orientation questionnaire consists of 24 different allocation tasks. In each task, a subject chooses among two payoff allocations, called options A and B (see Table 1A.1). Each option allocates money, in experimental currency units, to the subject herself (*own payoff* x) and an anonymous recipient (*other's payoff* y). The recipient stays the same in all 24 allocation tasks and answers herself the same set of questions (thereby, vice versa, influencing the first person's payoff). It is common knowledge that both persons receive the same set of tasks. No feedback about the other person's decisions is given during the questionnaire to avoid any strategic considerations.

All used payoff allocations lie, equally distributed, on a circle with radius $r = 15$ that is centered at the origin of an x- y-coordinate system, i.e. $r^2 = 15^2 = x^2 + y^2$ holds. Note that it is possible to represent these allocations by vectors in a Cartesian plane. Tasks are designed such that subjects always decide between two adjacent payoff allocations. By assuming that subjects have a preferred motivational vector $\vec{M}$ somewhere in the Cartesian plane, it is optimal for them to always choose the allocation that is closer to $\vec{M}$.

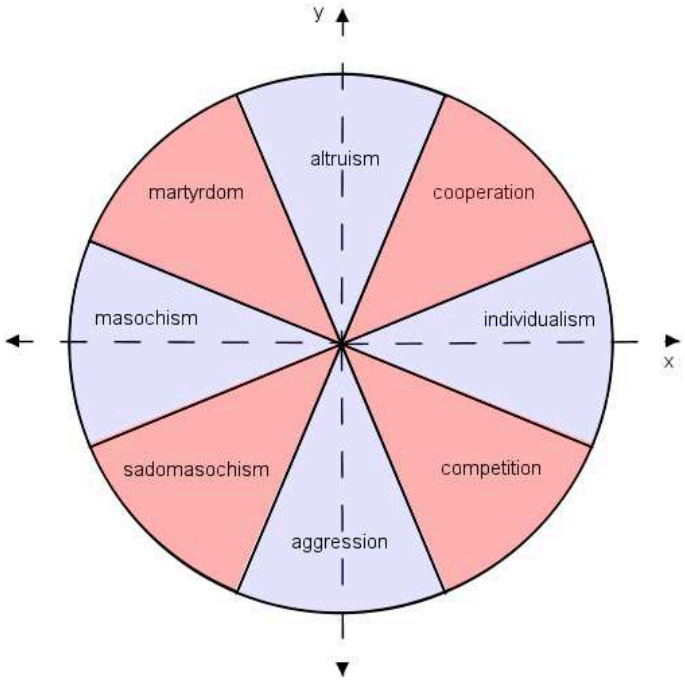
Adding up subject's x and y separately across all decisions yields a total sum of money allocated to the subject herself (X) and to the recipient (Y). The point (X, Y) determines the vector $\vec{A}$ which is used to estimate a subject's social orientation. This is done by computing the angle α between $\vec{A}$ and the x-axis using $\tan\alpha = Y/X$. The size of the angle specifies in which out of eight behavioral types a subject is sorted (see Figure 1A.1). Subjects with an angle α between 337.5° and 22.5° are classified as individualistic, subjects with an angle between 22.5° and 67.5° as cooperative. The other categories are: altruism (between 67.5° and 112.5°), martyrdom (between 112.5° and 157.5°), masochism (between 157.5° and 202.5°), sadomasochism (between 202.5° and 247.5°), aggression (between 247.5° and 292.5°), and competition (between 292.5° and 337.5°).

Additionally, the length of vector $\vec{A}$ can be used as a consistency measure. If a subject decides consistently over all 24 allocation tasks, the length will be 30 while perfect random choice will result in a vector of zero length. The greater the length of the vector the more consistent is a subject's decision. The questionnaire is fully incentivized since subject's earnings are determined by the sum of her decisions for *your payoff* and the sum of the recipient's decisions for *other's payoff*.

Table 1A.1: The 24 allocation tasks

	Option A		Option B	
Question number	your payoff (x)	other's payoff (y)	your payoff (x)	other's payoff (y)
1	15	0	14.5	-3.9
2	13	7.5	14.5	3.9
3	7.5	-13	3.9	-14.5
4	-13	-7.5	-14.5	-3.9
5	-7.5	13	-3.9	14.5
6	-10.6	-10.6	-13	-7.5
7	3.9	14.5	7.5	13
8	-14.5	-3.9	-15	0
9	10.6	10.6	13	7.5
10	14.5	-3.9	13	-7.5
11	3.9	-14.5	0	-15
12	14.5	3.9	15	0
13	7.5	13	10.6	10.6
14	-14.5	3.9	-13	7.5
15	0	-15	-3.9	-14.5
16	-10.6	10.6	-7.5	13
17	-3.9	-14.5	-7.5	-13
18	13	-7.5	10.6	-10.6
19	0	15	3.9	14.5
20	-15	0	-14.5	3.9
21	-7.5	-13	-10.6	-10.6
22	-13	7.5	-10.6	10.6
23	-3.9	14.5	0	15
24	10.6	-10.6	7.5	-13

Figure 1A.1: Classification of behavioral types



1B Experimental instructions (originally in German)[41]

A warm welcome to an experiment on decision making!

Thank you for participating!

During the experiment you and all other participants will be asked to make decisions. Your decisions as well as the decisions of the participants you are matched with determine your earnings from the experiment according to the following rules. Please stop talking to other participants from now on.

The whole experiment is computerized and will last approximately **75 minutes**. All your decisions and answers as well as your earnings remain anonymous. You will not find out with whom you are matched in each of the experiment's parts and how much each of the other participants earns. We evaluate data from the experiment on aggregate level only and never link names to data from the experiment. At the end of the experiment, you will be asked to sign a receipt for your earnings. This has accounting purposes only.

The experiment consists of **two** parts. Your decisions in Part I of the experiment **do not** have any effects on Part II. At the beginning of each part, you will receive the corresponding instructions for this part. The instructions will be read out loud and you will get time to ask questions. Please, do not hesitate to ask if anything is unclear to you. If you have any questions, please raise your hand, and one of the experimenters will come to you and answer your questions privately. In case the question is relevant for all participants, its answer is repeated aloud. In the interest of clarity, we will only use male terms in the instructions. They should be interpreted as being gender-neutral.

While taking your decisions at the PC, there will be a clock counting down in the right upper corner of the screen. The clock serves as a guide for how much time you should need. You may exceed the time. The input screens will **not** be turned off when time has run out. However, the pure information screens will be turned off when time has run out. Once you have taken a decision or have read through a screen, please confirm by pressing the "OK" button. For means of help, you will find a pen on your table.

Your earnings in the experiment will be calculated in **"tokens"**. At the end of the experiment, the "tokens" get converted into euro at the exchange rate announced in the respective part. In addition, you receive 4 euro for your arrival on time. Your total earnings from the experiment will be paid out to you privately and in cash at the end of the experiment.

[41] Baseline instructions describe treatment S. Differences in the other treatments are indicated by [**WEAK**]: weak punishment instrument, [**COUNTER**]: counter-punishment and [**UNANIMITY**]: unanimity rule.

Part I

In Part I of the experiment all participants are randomly assigned into groups of two. You will not find out with whom you form a group – not during the experiment and not after the experiment either. Correspondingly, the other person in your group will receive no information about your identity. You have to take 24 decisions in this part of the experiment. In each decision you can choose between 2 options, A and B. Each option allocates a positive or negative payoff (earning) in tokens to you and to the other person in your group. The other person answers exactly the same questions. Your total payoff from Part I depends on your decisions *and* on the decisions taken by the other person in your group.

A decision example:

	Option A	Option B
Your payoff	10.00	7.00
Other's payoff	-5.00	4.00

- If you choose Option A you receive 10 tokens, and the other person loses 5 tokens. If the other person also chooses Option A, he, too, receives 10 tokens and you lose 5 tokens. In total, you therefore earn 5 tokens (10 tokens from your choice minus 5 tokens from the other person's choice). The other person earns 5 tokens (10 tokens – 5 tokens), too.
- In case you choose Option B and the other person chooses Option A, you earn 2 tokens (7 tokens from your choice minus 5 tokens from the other person's decision). The other person earns 14 tokens (10 tokens + 4 tokens).
- The remaining combinations (you choose A and the other person chooses B, or both persons choose B) are analogous to these two examples.

Overall you take 24 decisions like the one described above. Your total payoff is computed as follows: The 24 values for "your payoff" are summed up over your decisions. The 24 values for "other's payoff" are summed up over the other person's decisions. The sum of these two sums determines your total payoff from this part and is converted into euro at the end of the experiment as follows: **12 tokens = 1 euro**. This exchange rate is valid only for Part I of the experiment.

Note that you are not receiving information on each single decision taken by the other person in your group. Rather, you will find out only the sum of your decisions for "your payoff", the sum of the other person's decisions for "other's payoff" and your total payoff from Part I at the very end of the experiment. Note that you do not get any feedback immediately after Part I.

Part II

Every point you earn in Part II is converted into euro at the exchange rate of **30 tokens = 1 euro**. Your **initial endowment** is **60 tokens**. You receive this endowment only once at the very beginning of Part II.

Part II of the experiment consists of **10 identical periods** during which you will interact in **groups of 4**. In **each period** the **groups are randomly assigned**, i.e. you are matched with different persons. Neither during nor after the experiment will you be informed about the identity of the other persons with whom you are in one group in the respective period. The other participants won't be given this information either. In addition you will receive a **"group-membership number"**: 1, 2, 3, or 4. This number is also randomly assigned each period.

Timing of a period

Each period consists of **three** [**COUNTER**: **four**] **stages**. In stage 0 each group chooses one of the alternatives A or B. This choice affects later the course of stage 2 [**COUNTER**: and stage 3]. The setting in stage 1 is identical independent of the chosen alternative. We therefore start with the description of stage 1.

Stage 1:

In stage 1 of each period every group member receives **an endowment of 18 tokens**. These tokens can either be fully invested into a project or fully be transferred to the private account. Dividing the tokens between both options or saving the tokens for subsequent periods is not possible.

Option 1:

You invest all 18 tokens into the **project**. Your earnings from the project are equal to the total investment of all four group members into the project *multiplied* by the factor 0.4. If you invest your 18 tokens into the project, the total investment rises by 18 tokens and your earnings increase by 18*0.4 = 7.2 tokens. At the same time the earnings of each other group member also increase by 18*0.4 = 7.2 tokens. Thus, the total earnings of the group increase by 28.8 tokens. Your investment into the project therefore also increases the earnings of the other group members. This also holds vice versa: Your earnings increase if the other group members invest into the project. Every other group member who invests into the project increases your earnings by 18*0.4 = 7.2 tokens (irrespective of whether you invest yourself or not).

Option 2:

You transfer all 18 tokens to your **private account**. These tokens turn solely and in a one-to-one relationship into your earnings. Hence, if you transfer the 18 tokens to your private account your earnings increase by exactly 18 tokens. The other group members do not receive any earnings from your transfer to your private account. Vice versa, you do not receive any earnings from the other group members' transfers to their private accounts.

Your earnings after stage 1 of a period amount to the sum of your earnings from the project and your earnings from your private account:

Earnings after stage 1 = Earnings from project + Earnings from private account

The other group members' earnings are calculated in the same way.

On the screen you will be asked whether you want to invest your 18 tokens into the project. If you choose "YES" your endowment will be fully invested into the project. If you choose "NO" your endowment will be fully transferred to your private account.

Stage 2:
In stage 2 the stage-1-earnings can be changed. How the earnings can change depends on whether your group has chosen alternative A or B in stage 0.

Alternative A:
If your group has chosen alternative A, you will be informed of every group member's investments into the project, transfers to the private account, as well as earnings after stage 1. Then you can change the stage-1-earning of a group member by assigning a **"deduction point"**. You decide for **every** group member separately whether you want to assign a deduction point to him or not. Every deduction point you assign costs you **1 token** and decreases the earnings of the respective group member by **6 tokens** [**WEAK: 2 tokens**]. You can assign at most one deduction point to each group member. The other group members take the same decisions.

At the end of stage 2 [**COUNTER**: At the beginning of stage 3] you will be informed of every group member's costs due to assigned deduction points, costs due to received deduction points as well as earnings after stage 2. You will also be informed from which group members you have received deduction points.

Alternative B:
If your group has chosen alternative B, you will also be informed of every group member's investments into the project, transfers to the private account, as well as earnings after stage 1. However, you are not able to change the earnings of the other group members, as their earnings will be changed **automatically** by the following mechanism: Every group member who transferred the endowment of 18 tokens to his **private account** will get allocated 3 deduction points, such that his earnings from stage 1 will be reduced by **18 tokens** [**WEAK: 6 tokens**]. This means that in stage 2, your complete earnings [**WEAK**: a third of your earnings] from your private account will be deducted. On the contrary, group members who invested into the project will not get allocated any deduction points. The earnings from the project thus will not be reduced.
However, alternative B comes at a cost: If the group decides on alternative B, **costs of 3 tokens** accrue to **every** group member. These costs accrue independently from whether and how many group members do not invest into the project.
In case alternative B is chosen, there are no decisions to take in stage 2. Thus, on the screen you will directly be shown for every group member the costs due to choosing alternative B, the costs due to automatically received deduction points as well as earnings after stage 2.

Your earnings after stage 2 of a period are, hence, calculated depending on the chosen alternative:

In case of alternative A:

Earnings after stage 2 = Earnings after stage 1 – possible costs due to assigned deduction points - possible costs due to received deduction points

In case of alternative B:

Earnings after stage 2 = Earnings after stage 1 – costs due to the choice of alternative B - possible costs due to automatically received deduction points

The other group members' earnings are calculated in the same way.

[**COUNTER**: **Stage 3:**

Stage 3 is only reached if your group chose alternative A. If your group chose alternative B, stage 3 will be skipped and the earnings after stage 2 will automatically be the earnings after stage 3.

If your group chose alternative A, you will be informed, as already mentioned, of every group member's costs due to assigned deduction points, costs due to received deduction points as well as earnings after stage 2. You will also be informed from which group members you have received deduction points. You can then change the stage-2-earnings of **those and only those** group members by assigning to them a **counter-deduction point.** For every group member from whom you have received a deduction point you decide separately whether you want to assign to him a counter- deduction point. Every counter-deduction point you assign to someone costs you **1 token** and reduces the earnings of the respective group member by **6 tokens** [**WEAK: 2 tokens**]. You can at most assign one counter-deduction point to a group member. Please note also that you cannot assign a counter-deduction point to a group member from whom you have not received a deduction point. If you have not received a deduction point from any group member you therefore will not be able to assign any counter-deduction points. In this case just press the "OK"-button on the screen. The other group members take the same decisions.

At the end of stage 3, you will be shown for every group member the costs due to assigned counter-deduction points, the costs due to received counter-deduction points as well as the resulting earnings after stage 3.

Your earnings after stage 3 of a period are, hence, calculated depending on the chosen alternative:

In case of alternative A:

Earnings after stage 3 = Earnings after stage 2 – possible costs due to assigned counter-deduction points - possible costs due to received counter-deduction points

In case of alternative B:

Earnings after stage 3 = Earnings after stage 2

The other group members' earnings are calculated in the same way.]

Stage 0:

In stage 0 your group can choose whether alternative A or B should be implemented for the group. The voting process is the following: All group members decide simultaneously whether they prefer alternative A or B. Alternative B will be implemented if **more than** 2 group members decide on alternative B. If **more than** 2 group members vote for alternative A, alternative A will be implemented for the group. If **exactly** 2 group members vote for each alternative, one of the two alternatives will be chosen randomly and with equal probability. [**UNANIMITY**: Alternative B will be implemented if **all** 4 group members decide on alternative B. If **at least one** group member votes for alternative A, alternative A will be implemented for the group.] At the end of stage 0 you are informed of the alternative that will be implemented in your group. Thus, you can condition your behavior in stage 1 on the chosen alternative. Please note that the chosen alternative is only relevant for the current period. In the next period you will interact with different persons and there is a new vote.

Overall you will be taking decisions in 10 identical periods. Every period consists of the three [**COUNTER**: four] stages described above. At the end of the experiment, **the earnings after stage 2** [**COUNTER**: **after stage 3**] of all periods will be summed up and converted into euro. Please note: Both by choice of alternative A and B, negative earnings after stage 2 [**COUNTER**: stage 3] may, perhaps, arise in single periods. Such losses will be compensated by gains from other periods and the initial endowment.

The End

After 10 periods the experiment ends. After filling out a short questionnaire, you will be informed of your total earnings from Part I and Part II. Then you will be paid out your final earnings.

Control Exercises

Please solve the following exercises. They serve you as a control whether you understood the instructions correctly. All exercises are based on arbitrary examples. Please note the hints at the end of this section.

Exercise 1:

a) Assume that in a given period one group member opts for alternative A and three group members opt for alternative B. Which alternative will be implemented in this group during this period?

b) Assume that in a given period three group members opt for alternative A and one group member opts for alternative B. Which alternative will be implemented in this group during this period?

c) [**UNANIMITY**: Assume that in a given period all four group members opt for alternative B. Which alternative will be implemented in this group during this period?]

Exercise 2:

Assume that **alternative A** is implemented in a group. Group members 1, 2 and 3 invest their 18 tokens into the **project**:

a) What are the earnings **after stage 1** for persons 1, 2, 3, and 4 if person 4 also invests his 18 tokens into the project?

b) What are the earnings **after stage 1** for persons 1, 2, 3, and 4 if person 4 by contrast transfers his 18 tokens to his private account?

c) In case of situation b): What are the group members' earnings **after stage 2** if person 1 assigns a deduction point to person 4?

d) [**COUNTER**: In case of situation c): What are the group members' earnings **after stage 3** if person 4 assigns a counter-deduction point to person 1?]

Exercise 3:

Assume that **alternative A** is implemented in a group. Group members 1, 2 and 3 transfer their 18 tokens to their **private accounts**:

a) What are the earnings **after stage 1** for persons 1, 2, 3, and 4 if person 4 also transfers his 18 tokens to his private account?

b) What are the earnings **after stage 1** for persons 1, 2, 3, and 4 if person 4 by contrast invests his 18 tokens into the project?

c) In case of situation b): What are the group members' earnings **after stage 2** if person 4 assigns a deduction point to all his group members and, additionally, person 1 assigns a deduction point to both person 2 and person 4?

d) [**COUNTER**: If person 2 wants to assign counter-deduction points, whom can he assign counter-deduction points to?]

Exercise 4:

Assume that **alternative B** is implemented in a group. Group members 1, 2 and 3 transfer their 18 tokens to their **private accounts.**

a) What are the earnings **after stage 2** [**COUNTER**: **after stage 3**] for persons 1, 2, 3, and 4 if person 4 also transfers his 18 tokens to his private account?

b) What are the earnings **after stage 2** [**COUNTER**: **after stage 3**] for persons 1, 2, 3, and 4 if person 4 by contrast invests his 18 tokens into the project?

Exercise 5:

Assume that **alternative B** is implemented in a group. Group members 1, 2 and 3 invest their 18 tokens into the **project.**

a) What are the earnings **after stage 2** [**COUNTER**: **after stage 3**] for persons 1, 2, 3, and 4 if person 4 also invests his 18 tokens into the project?

b) What are the earnings **after stage 2** [**COUNTER**: **after stage 3**] for persons 1, 2, 3, and 4 if person 4 by contrast transfers his 18 tokens to his private account?

Computational hints:

$18 \cdot 1 = 18$	$18 \cdot 0.4 = 7.2$
$18 \cdot 2 = 36$	$36 \cdot 0.4 = 14.4$
$18 \cdot 3 = 54$	$54 \cdot 0.4 = 21.6$
$18 \cdot 4 = 72$	$72 \cdot 0.4 = 28.8$

1C Fehr and Schmidt (1999) preferences

The model

The inequity aversion model of Fehr and Schmidt (1999) states that a subject i's utility function (see 1C.1) consists of her own monetary payoff π_i (first term) *and* an inequity aversion component (terms 2 and 3). Inequity is modeled by taking the individual differences between π_i and the payoffs π_j of each of the other $n-1$ members of i's reference group and averaging over it. The model considers disadvantageous and advantageous inequity by forming two subgroups and including the parameters α_i and β_i that weight both possibilities, respectively. It is assumed that $0 \leq \beta_i < 1$ and $\beta_i \leq \alpha_i$, i.e. subjects weakly dislike inequity and the utility loss from disadvantageous inequity is not smaller than the one from advantageous inequity.[42]

$$U_i(\pi) = \pi_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{\pi_j - \pi_i, 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{\pi_i - \pi_j, 0\} \qquad \text{(1C.1)}$$

Implications for the formal institution

For the formal institution, Fehr and Schmidt preferences can easily be calculated as there is no punishment decision to take for individuals. Hence, Proposition 4 of Fehr and Schmidt (1999, p. 839) for the standard public goods game matters but has to be adjusted to account for automatic punishment. The relevant condition that corresponds to Proposition 4(a) of Fehr and Schmidt (1999) is

$$\frac{\gamma}{n} + \left(1 - \frac{f}{E}\right) \cdot \beta_i < 1 - \frac{f}{E}. \qquad \text{(1C.2)}$$

This means that free-riding is a dominant strategy for player i if the sum of the marginal per capita return (γ/n) and the marginal non-monetary benefit due to reduced advantageous inequity (at most $(1 - f/E) \cdot \beta_i$) is strictly lower than the marginal monetary gain from free-riding ($1 - f/E$). On the other hand, if the inequality sign is reversed player i will contribute if all other group members contribute as well. Note that the introduction of formal punishment (f/E) has two effects: it reduces both the monetary gain and the inequity in case of free-riding. It is obvious that for the strong instrument ($f/E = 1$), free-riding is, independent of β_i never optimal. This is not astonishing as standard preferences assuming rational and selfish

[42] To justify $\beta_i < 1$, mind that $\beta_i = 1$ implies that a subject is willing to burn one unit of money in order to reduce advantageous inequity. This is rather unlikely. Note further that the model nests the standard preferences for $\alpha_i = \beta_i = 0$.

subjects already yield that contributing to the public account is the dominant strategy. The full cooperation equilibrium is preserved under Fehr and Schmidt preferences as deviations cannot decrease inequity (actually they do not affect inequity at all). In case of the weak instrument ($f/E = 1/3$), inserting our experimental parameters yields $\beta_i < 2/5 = 0.4$. Hence, no individual with $\beta_i < 0.4$ will contribute to the public account. On the contrary, if all subjects fulfill $\beta_i \geq 0.4$ full cooperation is an equilibrium outcome.[43] Assuming for the purpose of illustration a uniform distribution of β_i, the probability of having four subjects with $\beta_i \geq 0.4$ in one group is $0.6^4 = 12.96\%$.[44] Furthermore, in line with Proposition 4(b) of Fehr and Schmidt (1999) we can show that no equilibrium with positive contributions exists if at least one group member satisfies $\beta_i < 0.4$.[45]

Implications for the informal institution

Regarding the informal institution, we deal with Proposition 5 of Fehr and Schmidt (1999, p. 841) as individuals can punish each other individually to reduce inequity. Note that Fehr and Schmidt preferences cannot explain the usage of counter-punishment in our setup because exerting retaliation reduces the monetary payoff *and* increases inequity.[46] Hence, counter-punishment is not credible and we can ignore the counter-punishment stage in the following. Further, in order to be able to compare results more directly to the formal institution, we do not make the assumption that there are only players with $\alpha_i = \beta_i = 0$ or $\beta_i \geq 1 - \gamma/n$ but also allow for intermediate types. Define a group of $n' \leq n$ "enforcers" who are willing to punish a defector out of the remaining $n - n'$ group members whereas the latter do not punish.[47] Consider the following strategies derived from Proposition 5: Each player contributes $c_i = E$. If each player does so, there is no punishment. If one of the non-enforcers deviates by contributing $c_i = 0$, she will be punished by each enforcer while all other players

[43] Of course, there is also an equilibrium in which no individual contributes. However, according to Fehr and Schmidt (1999), a plausible equilibrium refinement concept – pareto optimality – would eliminate it. We do not focus on the description of the whole set of equilibria in this appendix but concentrate on requirements to sustain positive contributions. Note that multiple equilibria also arise in the informal institution which is discussed next.

[44] We choose this simple uniform distribution because it at least roughly corresponds with the parameter distribution given in Fehr and Schmidt (1999, p. 844). As Fehr and Schmidt (1999) do not provide an explicit threshold of 0.4, we cannot directly infer a probability from their paper. Of course, the probability that full cooperation can be sustained in equilibrium increases in the weight you put on observing $\beta_i \geq 0.4$.

[45] This can easily be checked by verifying free-riding incentives for a contributor i in each of the three intermediate cases in which only one, two or three group members contribute. Positive contributions require that the utility of contributing is at least equal to the utility of not contributing, i.e. $21.6 - 4\alpha_i \geq 26.4 - 8\beta_i$ for three contributors, $14.4 - 8\alpha_i \geq 19.2 - 4\beta_i$ for two contributors or $7.2 - 12\alpha_i \geq 12$ for one contributor. However, none of these inequalities can be fulfilled for given parameter restrictions.

[46] The latter is due to our mechanism that ensures that free-riders, after being punished, never have payoffs below those of the punishing cooperators. Note further that free-riders have no reason to punish and cooperators will not be punished.

[47] An enforcer is not necessarily "conditionally cooperative" as we do not impose $\beta_i \geq 1 - \gamma/n$.

do not punish. If one of the enforcers deviates by choosing $c_i = 0$ or if more than one player deviates from contributing E, one Nash equilibrium of the punishment game is played. For this to be a subgame perfect equilibrium, we have to check the following three conditions:

(i) Free-riding does not pay for a non-enforcer: the monetary gain from free-riding ($E \cdot (1 - \gamma/n)$) plus the non-monetary utility loss due to suffering from advantageous inequity towards the enforcers ($-n'/(n-1) \cdot \beta_i \cdot (E - n'l + 1)$) plus the utility loss towards the contributors who do not punish ($-(n - n' - 1)/(n-1) \cdot \beta_i \cdot (E - n'l)$) has to be smaller or equal to the monetary loss from being punished ($n'l$). Setting up the inequality condition, simplifying and inserting $E = 18$, $\gamma = 1.6$ and $n = 4$ yields

$$\beta_i \geq \frac{(10.8 - n'l)}{\left(18 + \frac{n'}{3} - n'l\right)}. \tag{1C.3}$$

For the strong instrument ($l = 6$), any $\beta_i \geq 0$ satisfies the condition if $n' = 2$ or 3, i.e. even a person who does not care at all about inequity will contribute to the public account if she fears the punishment of at least two group members. Furthermore, if $n' = 1$, we demand $\beta_i \geq 0.39$ approximately and if $n' = 0$ we need $\beta_i \geq 0.6$. For the weak instrument ($l = 2$) and $n' = 3$, the necessary requirement is already about $\beta_i \geq 0.37$. Note that the latter condition almost equals the condition of the formal institution per construction.[48] Requirements on β_i increase further if n' is reduced and finally also demand $\beta_i \geq 0.6$ for $n' = 0$.

(ii) The punishment threat a non-enforcer faces is credible: enforcer i's utility change in the punishment stage from punishing assuming that all other $n' - 1$ enforcers punish, consists of the monetary costs (-1), the disadvantageous inequity towards contributing but not punishing group members ($-\alpha_i/(n-1) \cdot (n - n' - 1)$) and the disadvantageous inequity towards the defecting member who gets punished ($-\alpha_i/(n-1) \cdot (1 - n'l)$). This has to be at least as good as i's utility change if she does not exert punishment (given the other $n' - 1$ enforcers punish), which results in disadvantageous inequity towards the defecting member ($-\alpha_i/(n-1) \cdot (-l \cdot (n' - 1))$) and advantageous inequity towards the group members who punish ($-\beta_i/(n-1) \cdot (n' - 1)$). Setting up the inequality condition, simplifying and inserting $n = 4$ leads to

[48] The fine in the formal institution is three times the individual fine in the informal institution. The small difference between both conditions is solely due to the fact that punishments in the informal institution additionally decrease punishers' payoffs and, thus, slightly more increase inequity towards the defector.

$$l \geq 4 - n' + \frac{1}{\alpha_i} \cdot [3 - \beta_i \cdot (n' - 1)]. \qquad (1C.4)$$

For the strong instrument ($l = 6$) and $n' = 1$ this implies $\alpha_i \geq 1$, for $n' = 2$: $\alpha_i \geq 0.75 - 0.25\beta_i$ and for $n' = 3$: $\alpha_i \geq 0.6 - 0.4\beta_i$. For the weak instrument ($l = 2$) the requirements are much more restrictive: for $n' = 3$ we need $\alpha_i \geq 3 - 2\beta_i$ which is at least larger than 1. Moreover, for $n' < 3$ punishment is never credible.

(iii) None of the enforcers profits from free-riding: conditions (i) and (ii) on at most $n' - 1$ have to hold for the respective subgroup as there is one enforcer less to punish a defecting enforcer. Note that $\beta_i \geq 0.6$ is always sufficient.

To sum up our results for the informal institution, we find that in case of the strong punishment instrument full cooperation can be sustained in equilibrium under rather weak assumptions. Indeed, up to two group members that do not care at all about inequity can be motivated to contribute if the other group members are sufficiently averse to disadvantageous inequity (and have an incentive to contribute themselves). Using the uniform distribution for β_i and assuming both that 40% of individuals satisfy $\alpha_i \geq 1$ and that there is a perfect positive correlation between α_i and β_i (both as it is done in Fehr and Schmidt, 1999), all 40% of individuals with $\beta_i \geq 0.6$ fulfill $\alpha_i \geq 1$ and can therefore credibly threaten free-riders. Hence, having at least two group members with $\beta_i \geq 0.6$ is sufficient to sustain full cooperation as an equilibrium outcome. As the following calculation shows, the probability that there are at least two such group members in a four-person group is quite high: $\binom{4}{2}0.4^2 0.6^2 + \binom{4}{3}0.4^3 0.6^1 + \binom{4}{4}0.4^4 0.6^0 = 52.48\%$. Note further that full cooperation can also be sustained if there is only one or even no group member who satisfies $\beta_i \geq 0.6$. Thus, with Fehr and Schmidt preferences full cooperation is in many cases an equilibrium outcome if the punishment instrument is strong.

On the contrary, if the punishment instrument is weak requirements for an equilibrium with full cooperation are rarely satisfied in the informal institution. Not only do we demand $\beta_i \geq 0.37\ \forall i$ (which is almost equal to the condition $\beta_i \geq 0.4\ \forall i$ in the formal institution), we also need either three enforcers with $\alpha_i \geq 3 - 2\beta_i$ *and* $\beta_i \geq 0.6$, or four enforcers with $\alpha_i \geq 3 - 2\beta_i$ or, alternatively, all four group members have to satisfy $\beta_i \geq 0.6$. Note that these restrictions are much more demanding than the requirement in the formal institution. For the purpose of illustration, once again, assume that β_i is uniformly distributed, 40% of individuals satisfy $\alpha_i \geq 1$ and there is a perfect positive correlation between α_i and β_i. Moreover, assume a decreasing density function for $\alpha_i \geq 1$ that allows 30% of individuals to

satisfy $\alpha_i \geq 3 - 2\beta_i$ *and* $\beta_i \geq 0.6$.[49] Then, the probability of observing a group structure that allows for full cooperation as an equilibrium outcome is approximately: $0.4^4 0.6^0 + 4 \cdot 0.3^3 \cdot (0.6 - 0.37) = 5.04\%$. This probability is lower than in the formal institution. Furthermore, in line with Proposition 4(b) of Fehr and Schmidt (1999) partial cooperation cannot occur under the weak instrument as there is no credible punishment within a subgroup and one free-rider is already enough to completely wipe out cooperation in the absence of punishment.

Implications for voting behavior

Fehr and Schmidt preferences can therefore explain why subjects may vote for the informal institution in treatments S, S-U and S-CP and for the formal institution in treatments W and W-U. The former is due to the existence of a subgame perfect equilibrium with higher payoffs in the informal institution once group structure allows that full cooperation is an equilibrium outcome. The reason is the saving of institutional fixed costs. The latter is caused by the higher probability to sustain an equilibrium with full cooperation in the formal institution. Whenever full cooperation can be an equilibrium outcome in the formal but not in the informal institution, subjects have an incentive to vote for formal punishments to maximize payoffs (the fixed costs are too low to offset the advantageousness due to increased cooperation). This case appears, roughly speaking, when all group members satisfy $\beta_i \geq 0.4$ but not $\beta_i \geq 0.6$.[50] Hence, Fehr and Schmidt preferences can account for a voting behavior that differs from standard theory.

[49] Looking at the distribution of α_i reported in Fehr and Schmidt (1999, p. 844), it makes sense to assume decreasing probabilities for $\alpha_i \geq 1$. Precisely, we assume the following: there are 10% of individuals with $1 \leq \alpha_i < 1.5$, 10% with $1.5 \leq \alpha_i < 2.5$, 10% with $2.5 \leq \alpha_i < 4$ and 10% with $\alpha_i \geq 4$. Connecting this to the uniform distribution of β_i by assuming perfect positive correlation it follows that all individuals with $\beta_i \geq 0.7$ have $\alpha_i \geq 1.5$ which is (roughly) in line with the requirement $\alpha_i \geq 3 - 2\beta_i$. However, 10% of individuals have $0.6 \leq \beta_i < 0.7$ and $\alpha_i < 1.5$ which fails to meet the condition. Hence, only 30% of individuals fulfill both $\alpha_i \geq 3 - 2\beta_i$ *and* $\beta_i \geq 0.6$. Note further that the probability of having four enforcers with $\alpha_i \geq 3 - 2\beta_i$ in one group and at least one of them satisfying $\beta_i < 0.6$ is zero given our assumptions.

[50] In the real world inequity aversion can be more complex than the model assumes. If, for example, β_i decreases in the amount of inequity, i.e. individuals are more averse to generating small payoff differences, this would even more favor votes for the formal institution in treatments W and W-U.

1D Further results

Frequency of formal institutions over time

Table 1D.1: Number of groups playing the formal institution over time across treatments

Period	1	2	3	4	5	6	7	8	9	10	Sum (Percentage)
S	7	7	4	5	6	4	6	6	8	7	60 (50.00%)
W	0	2	2	4	6	3	2	4	2	0	25 (20.83%)
S-CP	1	3	4	4	4	5	5	7	8	7	48 (40.00%)

Contributions, sanctions and profits over time

Figure 1D.1: Average contributions over time across institutions and treatments

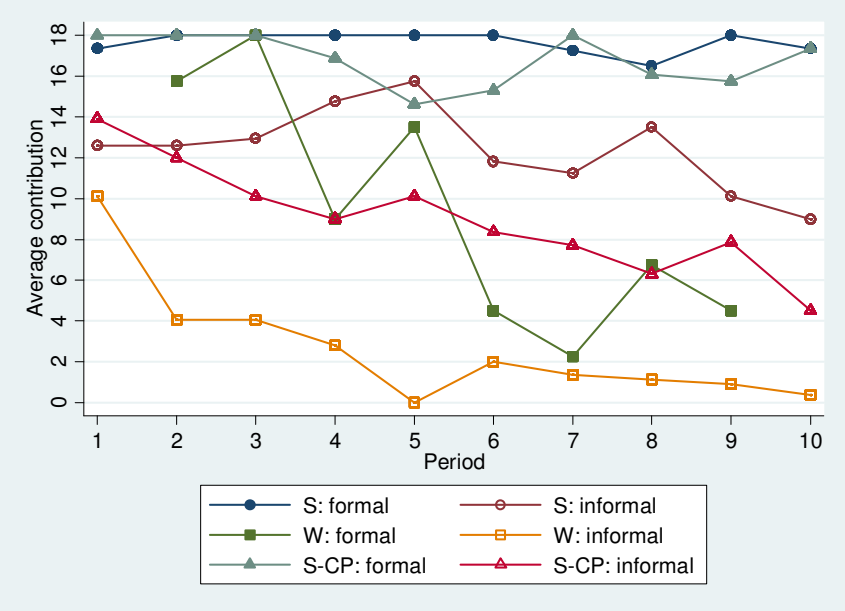


Figure 1D.2: Percentage of exerted sanctions in informal institution over time across treatments

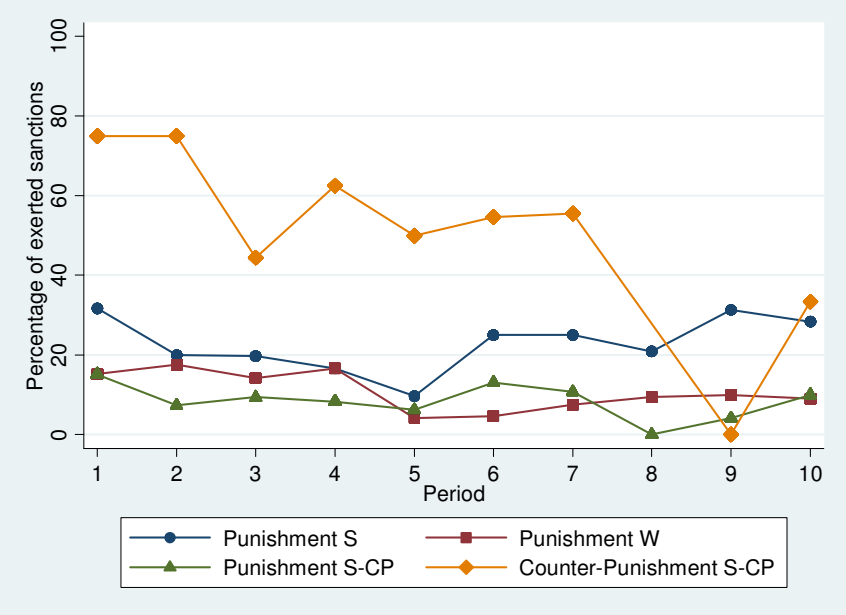
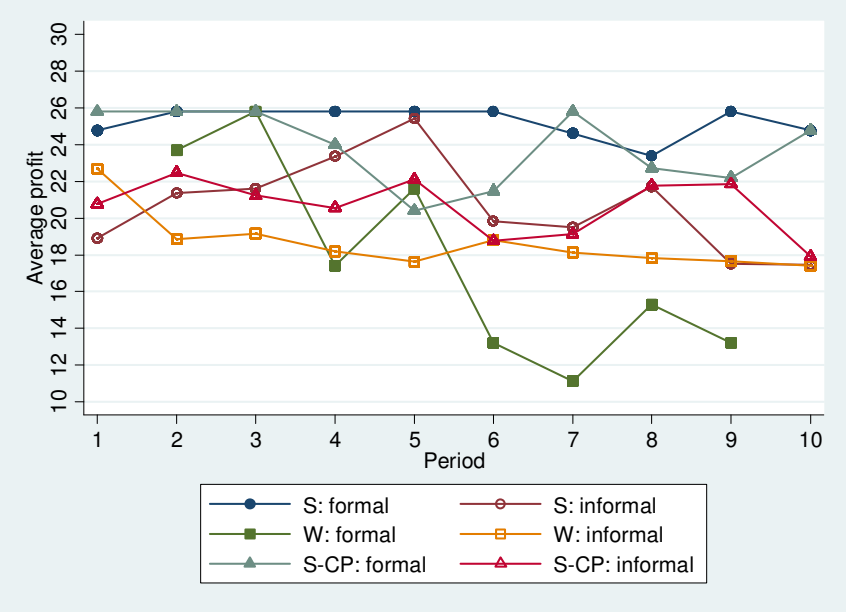
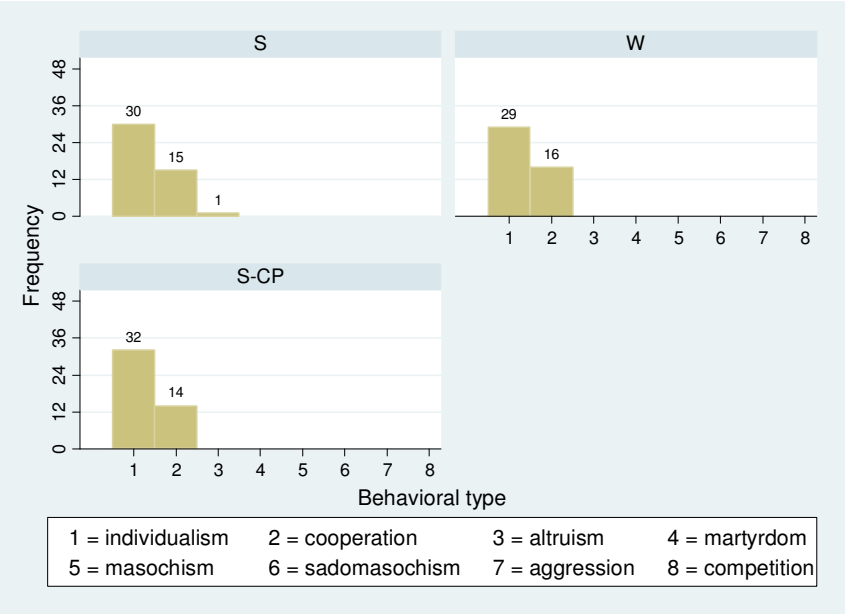


Figure 1D.3: Average profits over time across institutions and treatments



Results from the social value orientation questionnaire (ring test)

Figure 1D.4: Frequency of behavioral types by treatment (consistency ratio ≥ 2/3 required)



Chapter 2

Preferences over Punishment and Reward Mechanisms in Social Dilemmas[51]

2.1 Introduction

We investigate preferences over informal punishment and reward institutions in social dilemmas. A social dilemma such as the private provision of a public good is characterized by a collectively inefficient equilibrium if all group members are selfish and rational. Sanctioning institutions can, however, sustain cooperation and, thus, enforce the collectively optimal outcome. In order to study preferences over such enforcement mechanisms, we combine an experimental public goods game with an endogenous choice of the enforcement mechanism within the provision group. Precisely, group members decide to implement either a standard voluntary contribution mechanism (henceforth, standard VCM), a VCM with informal punishment opportunities, or a VCM with informal reward opportunities. Furthermore, they choose the effectiveness of the sanctioning instrument in case of punishment or reward, i.e. the strength of the enforcement mechanism. We study a multi-period game with repeated mechanism choice both theoretically and experimentally and vary group composition exogenously (i.e. use a partner and stranger design as treatment variable) in order to control for the impact of reputation on mechanism choice.

The choice of the appropriate enforcement mechanism is crucial in many real-life social dilemmas as groups often have the power to establish themselves their own guidelines. These self-imposed rules can be either explicit or implicit. Explicit guidelines exist, for example, on the international level where agreements like the EU treaties or the directives of the United Nations Security Council govern the legitimacy of informal sanctions. Implicit guidelines, for instance, are incorporated in social norms that develop in small work or sports teams to suggest a particular reaction to others' degrees of cooperation. Note that both repeated and one-shot interactions are observed in reality, and it is thus important to consider different time

[51] This chapter is joint work with Martin Kocher.

horizons of group members' collaboration by experimentally implementing partner and stranger settings.

There are only a handful of existing papers that study mechanism choice in public goods games.[52] Among them are Botelho et al. (2005), Gürerk et al. (2006), and Sutter et al. (2010) who implement different versions of institutional choice in social dilemmas. Following Sutter et al. (2010), we are interested in institutional choice in fixed groups and not in a voting-by-feet mechanism as used by Gürerk et al. (2006). Of course, both approaches are relevant for different applications. In comparison to the existing literature, our study features two important innovations: First, we allow group members to agree on either a standard VCM, or a VCM with punishment or a VCM with reward in every period. This enables us to study the dynamics of institutional choice over time. Especially, such a setting facilitates institutional learning, and later-period decisions are clearly more informed. Second, we are the first to study preferences over the level of effectiveness that an enforcement device brings about. Evidently, the strength of an enforcement mechanism is a very important feature. One might, for instance, prefer reward with a high effectiveness, but punishment with a low level of effectiveness if reward is not available. In general, the level of effectiveness of an institution seems to be the main determinant that drives institutional choice both in the real world and in the experimental laboratory. As a consequence, our study is able to provide empirical evidence for a relevant aspect of institutional choice in social dilemmas that has always been implemented exogenously so far and, thus, preferences over different strengths could hitherto not be analyzed.

We compare our empirical findings with predictions based on standard theory and two famous other-regarding preference models: the inequity aversion model of Fehr and Schmidt (1999) and the social welfare model of Charness and Rabin (2002). While inequity aversion can explain a subject's preference for the VCM with punishment, the Charness and Rabin (2002) model provides an underpinning for a preference for the reward institution. Hence, we are able to assess our results in the light of two important but conflicting behavioral motivations.

As a robustness check, we implement both a partner and a stranger design in our experiment. While groups remain unchanged in the partner matching, we randomly re-match experimental subjects in each period of the stranger design sessions. The two treatments allow us to verify to what extent results hinge on the repetition of group members' interaction.

[52] There is a large literature on the effects of exogenously implemented reward and punishment mechanisms in social dilemmas. Among the seminal papers are Fehr and Gächter (2000, 2002), Masclet et al. (2003), Andreoni et al. (2003), or Sefton et al. (2007).

Another important issue in endogenous choice experiments is self-selection of cooperative types into specific institutions (Dal Bó et al., 2010; Sutter et al., 2010). Suppose that certain types of decision makers have a preference for certain mechanisms. If a group, by chance, consists of such types, it is impossible to distinguish whether the level of cooperativeness is a consequence of their type or of the chosen institution. We therefore employ an independent, individual, and fully-incentivized measure of the level of cooperativeness in our experiment, the so-called social value orientation questionnaire or ring test. It allows us to disentangle potential self-selection effects from the impact of the chosen institution.

While a vote on the institution and the enforcement mechanism in the way we implement it in the laboratory is unlikely to be taken in the real world, the main features that we incorporate in our experimental design capture real-world decision making problems in face of social dilemmas. We believe that, by giving up a bit of reality in terms of the actual bargaining process involved in establishing rules and norm-enforcement mechanisms, we gain a lot of analytical strength from our experimental design.

Our main empirical results, first, clearly indicate that the VCM with reward is the most favored and the most efficient institution, although the punishment institution leads to higher contributions. Indeed, we observe an astonishingly stable pattern of approximately 50% of votes for the VCM with reward and 25% each for the two other institutions. Second, subjects prefer extreme values of the strength of the sanctioning technologies, and contributions increase in the level of strength. Third, there is no significant difference in the voting pattern between partner and stranger matching treatments and no significant self-selection of behavioral types into specific mechanisms, but women vote significantly less frequently for the VCM with punishment and more frequently for the standard VCM than men.

The remainder of the chapter has the following structure: Section 2.2 presents a short overview of the related literature. In Section 2.3 we explain our basic public goods setup, and Section 2.4 adds the details of the experimental procedure. Section 2.5 deals with theoretical predictions. In Section 2.6 we provide the results of our experiment and Section 2.7, finally, concludes the chapter.

2.2 A brief review of related literature

The standard VCM is widely examined in the literature. For an up-to-date overview of scholarly research we refer to Chaudhuri (2011).[53] One of the central insights is that contributions in small groups typically decay over time without enforcement mechanisms. This has raised researchers' interest to search for informal (i.e. contract-free) mechanisms that enhance the average level of cooperation. Among those are communication (Isaac and Walker, 1988; Bochet et al., 2006) and leadership (e.g. Güth et al., 2007). The two most prominent mechanisms, however, are a costly punishment option and a costly reward option. Both social sanctions are exerted individually after the observation of each other's contribution levels.

The literature on the punishment mechanism is huge and shows strong positive effects on contributions (e.g. Fehr and Gächter, 2000, 2002). Recent research exogenously varies the cost-effectiveness ratio of the punishment instrument revealing contribution levels to increase in the strength of the enforcement mechanism (Egas and Riedl, 2008; Nikiforakis and Normann, 2008).[54] However, the impact on efficiency levels caused by the punishment institution is ambiguous as the act of punishing reduces both the punisher's and the punished person's payoff (positive effects were found for the very long-run by Gächter et al., 2008). Regarding reward, Andreoni et al. (2003) show in a proposer-responder game that the instrument is frequently applied but less successful in sustaining cooperation than punishment. In combination with a public goods game, Sefton et al. (2007) confirm the weaker effect of a VCM with reward.

In contrast to the literature above that focuses solely on *exogenously* determined mechanisms, there are also a few papers that address voting decisions in the context of a sanctioning institution. For example, Decker et al. (2003) allow participants to bid for the right to select a punishment institution. Guillen et al. (2006) let subjects decide whether they want to abolish a costly centralized punishment mechanism. Kroll et al. (2007) show that a non-binding vote on minimum contributions increases cooperation only when connected with a punishment mechanism.

Among the papers that are most closely related to ours is Botelho et al. (2005). They let subjects play ten periods of a standard VCM followed by ten periods of a punishment institution (based on the Fehr and Gächter, 2000, 2002 design). Afterwards, subjects can vote

[53] For an older survey see Ledyard (1995). Zelmer (2003) provides a meta-analysis.

[54] Variations in the cost-effectiveness ratio are also considered by Anderson and Putterman (2006), Carpenter (2007a) or Masclet and Villeval (2008). See also the discussion in Casari (2005). Papers focusing on non-monetary punishment are for example Masclet et al. (2003) or Noussair and Tucker (2005).

for one of the two institutions to be implemented for a single last period. Interestingly, Botelho et al. (2005) report a predominant preference for the sanction-free environment. Ertan et al. (2009) allow for more voting decisions but confirm the inclination not to implement punishment institutions in the beginning. However, over the course of their experiment, an option to punish low contributors gains more support, while the possibility to punish high contributors is constantly ruled out.

Gürerk et al. (2006) use a completely different approach as they do not impose a fixed group structure. In their setting subjects can rather *move* between an open group with a standard VCM and an open group with a sanctioning institution that provides both punishment and reward options. Analyzing the results from this voting-by-feet design they find that over time more and more subjects self-select themselves into the sanctioning environment which turns out to be the more efficient institution. Rockenbach and Milinski (2006) extend this line of research by reporting increasing support for a punishment institution on top of an option for indirect reciprocity.

In the study of Sutter et al. (2010) subjects cast their vote either for a standard VCM, a VCM with punishment or a VCM with reward. Hence, they implement the same institutional choice as we do. However, in contrast to our design, subjects in their experiment decide only once in the beginning and are then forced to live under the chosen institution for the entire course of the repeated interaction. Moreover, Sutter et al. (2010) use unanimity voting requirement that includes the possibility of multiple voting rounds and/or only few subjects deciding for the group because of voting costs in their experiment. Furthermore, Sutter et al. (2010) only exogenously vary the effectiveness of punishment and reward. They report evidence that the reward institution is the most attractive institution in case of high levels of effectiveness, whereas there is a higher demand for the standard VCM when effectiveness levels are low. The punishment institution is never popular. Moreover, Sutter et al. (2010) compare their results with corresponding exogenous treatments and reveal an endogeneity premium (confirmed in Dal Bó et al., 2010), i.e. having the democratic choice between different mechanisms within a small group boosts cooperation regardless of the implemented mechanism.

Finally, note that, as far as we are aware of, there is no study comparing voting behavior in social dilemmas under both a partner and a stranger matching protocol. There is of course a large literature on group composition effects in standard public goods games. Andreoni (1988) initiates this research by reporting the counter-intuitive result that reshuffled groups show higher levels of cooperation. However, this finding is controversial. While Burlando

and Hey (1997) and Brandts and Schram (2001), for instance, find no robust effect in either direction, evidence for higher cooperation under a partner design is provided by Keser and van Winden (2000) and in the meta-analysis of Zelmer (2003).[55] In the presence of punishment instruments, Fehr and Gächter (2000) and Masclet and Villeval (2008) find positive effects of a fixed group structure, while Nikiforakis (2008) does not provide such evidence. However, in the latter paper there is weak support for more severe punishment activities with a partner design.

2.3 Our public goods setup

The basic game that we employ can be described as follows. Subjects play T periods. In each period $t \in \{1, 2, \dots, T\}$ an individual i is matched with $n-1$ other subjects to a group $I = \{1, 2, \dots, n\}$. She receives an endowment E and decides on how to split the endowment between a public and her private account. The voluntary contribution to the public account is denoted by $c_{i,t}$ and has to satisfy $0 \leq c_{i,t} \leq E$. We denote C_t as the sum of group members' contributions, i.e. $C_t = \sum_{j=1}^{n} c_{j,t}$, and γ as the marginal per capita return (MPCR) from the investment into the public account. For the latter we require $0 < \gamma < 1 < n\gamma$ to ensure that individuals face a social dilemma. Individual i's period payoff is then determined by

$$\pi_{i,t} = E - c_{i,t} + \gamma C_t. \tag{2.1}$$

So far, this game describes the standard VCM and decisions are made simultaneously. In case of the VCM with punishment or the VCM with reward, a second stage is added after the contribution decision. In this stage, subjects can sanction each other individually at own costs. Precisely, subjects receive information on the contribution level of each group member in their group and decide for each other group member separately whether they want to punish (reward) this person or not. Sanctions are a *binary* decision, i.e. subjects cannot assign more than one sanctioning point to a specific person. In case of punishment, each point costs the punisher one unit and *reduces* the punished person's payoff by $|L|$ units. In case of reward, each point costs the assigning person one unit but *increases* the rewarded person's payoff by L units. The letter L stands for "leverage" and captures the effectiveness (or relative costs) of the sanctioning institution. Note that the leverage level is determined endogenously by the

[55] See Andreoni and Croson (2008) for a review on partner versus stranger settings. Botelho et al. (2009) find that a perfect stranger design (with a zero re-match probability) lowers contribution levels compared to a standard stranger setting.

group members (for details regarding the voting mechanism, see the next section). Hence, in addition, a voting stage (stage zero) becomes part of the game before contributions and possible sanctions take place.

We incorporate the sanctioning possibility (punishment or reward) into equation (2.1) by defining $p_{ji,t} = 1$ if group member j is punished (rewarded) by member i in period t and zero otherwise. This yields the following period payoff for an individual i in the presence of a sanctioning institution:

$$\pi_{i,t} = E - c_{i,t} + \gamma C_t + L \sum_{\substack{j=1 \\ j \neq i}}^{n} p_{ij,t} - \sum_{\substack{j=1 \\ j \neq i}}^{n} p_{ji,t} \tag{2.2}$$

Note that the sign of L determines the nature of the sanctioning institution, i.e. VCM with punishment ($L < 0$) or reward ($L > 0$), while the absolute value indicates the strength of the respective sanction. The higher the absolute value, the larger are the monetary consequences for the punished (rewarded) subject per sanctioning incident (or, equivalently, the lower are the relative costs for the punishing or rewarding subject).[56] We interpret $L = 0$ as the standard VCM although it represents, strictly applying equation (2.2), the case of non-monetary punishment (reward). However, as we do not focus on non-monetary sanctions in our experiment, the simplification is innocuous. Hence, we refer to $L = 0$ as the case in which both last terms of equation (2.2) are dropped, i.e. we are left with the payoff formulation of equation (2.1).

Finally, we want to remark that our simple binary sanctioning mechanism is in line with the one used in Sutter et al. (2010) but relatively weak compared to institutions in the spirit of Fehr and Gächter (2000), in which subjects could assign multiple points to a specific group member yielding much more severe monetary consequences. Hence, any significant effects we find should constitute a lower boundary for the size of the effect.

2.4 Experimental design and procedure

We implemented the public goods game described in Section 2.3 and applied the following parameter realizations: $T = 10$ periods, endowment per period $E = 20$ tokens (experimental currency unit), group size $n = 4$ and MPCR $\gamma = 0.4$. At the beginning of each period groups

[56] Note that, by nature, punishment activities reduce efficiency. On the contrary, reward can be efficiency-reducing (if $L < 1$), efficiency-neutral ($L = 1$) or efficiency-improving ($L > 1$).

had to choose the institution *together with* the level of effectiveness in case of punishment or reward by selecting a leverage level from the interval $[-5, 2]$ by majority rule (only integers were accepted).

The reason for choosing an asymmetric interval is twofold. First, in reality there are natural limits to the leverage of reward, while the leverage of punishment can be almost infinite. Therefore, an asymmetric interval is more realistic than a symmetric one. Second, a very high leverage for reward would have been too attractive to allow for much variation in the preferences over the institutions of our subjects.

The voting process was implemented as follows: First, each group member had to state her preferred value out of the interval above. The four proposals were grouped into the three institutional categories $L < 0$, $L = 0$ and $L > 0$, and the category with the relative majority was then implemented as the relevant institution for the group in the respective period. If $L = 0$ got the relative majority, the standard VCM was implemented, and the voting stage ended. However, if there was a relative majority for the punishment or reward institution, subjects entered a second voting round in which the level of effectiveness was determined within the already chosen institution. For example, in case there was a relative majority for the punishment institution in the first voting round, the VCM with punishment was implemented and in voting round two the actual leverage level had to be determined out of the reduced interval $[-5, -1]$. All group members (also those who had voted for another category in voting round one) had to state their preferred value again, and the mean of their second proposals was then implemented as the relevant leverage level. Obviously, there was a chance that some participants had to change their decision from voting round one to voting round two because their preferred values were no longer available, while others had the possibility to restate their first preference. However, everyone explicitly had to make a second decision. In case of a tie in the first voting round, one of the four proposals within a group was chosen randomly and directly implemented. There was no second voting round in such a case. It is important to note that the chosen parameter L held only for the respective period. In each period the same voting procedure was implemented, and a group could agree on a different leverage level and even a different institution.

There are at least three advantages of our voting mechanism compared to the decision mechanisms used in related papers. First, the mechanism is incentive compatible in the sense that participants have an incentive to state their true preferences. Note that in voting round one an individual's proposal is implemented with positive probability (in case of a tie). Hence, everybody has an incentive to state their first preference. In the second voting round,

if applicable, the mean of the four proposals determines the strength of the chosen mechanism. Since all votes influence the mean, there are no votes that are unimportant. Obviously, the optimal proposal of a player now depends on her belief regarding the preferences of the other group members. Hence, stating one's first preference might potentially not be optimal in voting round two. Nonetheless, we have true preferences from voting round one for our empirical analysis. Second, our two-round voting procedure is easy to understand and easy to implement. Third, it gives us more structure on the preferences than many other voting procedures, because, even if the second voting round does not necessarily elicit true preferences, the information from the voting entails some interesting insights in the order of players' preferences.

We consider two treatments in our experiment: In the *partner* treatment, group composition remained constant over all periods, whereas in the *stranger* treatment groups were reassembled after each period. In both treatments ten rounds of the standard VCM were played as a separate part before our main treatments, the endogenous choice games, took place to ensure that subjects really understood how the social dilemma works.[57] In the partner treatment, group composition was the same in both parts and subjects got to know this before starting with the endogenous choice part. In the stranger treatment, group composition varied every period in both parts and this was also known by the participants.[58]

At the beginning of both treatments we included a social value orientation questionnaire, the so-called "ring test".[59] The test matches subjects into pairs of two and each group member completes the same set of 24 decision tasks. Each task contains a choice between two own-other payoff allocations. These allocations lie equally spaced on a circle around the origin of a two-dimensional coordinate system (that's why it is called ring test), and subjects always choose between two adjacent allocations. By summing up a subject's decisions for herself and the other group member we obtain a motivational vector and can sort the subject into a specific behavioral category. The most prevalent categories are *individualism* and *cooperation*. Moreover, the test provides us with a check for choice consistency and, thus, allows us to exclude subjects that just pick the allocations randomly. With the ring test (for more details regarding, for instance, incentivization of the test consider Appendix 1A) we

[57] Such a procedure is quite common and applied for example by Fehr and Gächter (2000) or Sefton et al. (2007).

[58] We did not impose a *perfect* stranger design, i.e. subjects could meet a certain group member more than once during and across specific parts of the experiment. However, the likelihood of meeting one group member again in the next period was low and, due to anonymity, subjects never knew whether they were matched with a person they already played with before. Hence, reputation effects should not matter in such a setting.

[59] This procedure stems originally from psychology (see van Lange et al., 1997 for a review). In the meanwhile it is quite common in economics and used for example by Offerman et al. (1996), Park (2000), Brosig (2002), van Dijk et al. (2002) or Sutter et al. (2010).

receive an independent measure of an individual's inclination to cooperate, allowing us to separate type effects from institutional effects in the public goods game. Moreover, we can study whether or not cooperative subjects vote differently than individualistic (i.e. selfish) subjects. One could for example argue that the punishment institution is much more attractive for cooperative subjects as it includes a mechanism to punish free-riders. On the other hand, even individualistic subjects might be willing to vote for the punishment institution if they presume that free-riding in the other institutions leads to lower payoffs due to lower contributions.

To sum up, our experimental design consists of three separate parts: first, the ring test; second, ten periods of the standard public goods game (either in a partner or a stranger design); and third, ten periods of the endogenous choice game (again, either in a partner or a stranger design). Subjects, at the start of the experiment, knew that the experiment is going to consist of three parts, but the details of each part were only revealed at the start of the respective part. Instructions of all parts (see Appendix 2A) were written neutrally, read aloud, and participants got enough time and were encouraged to ask questions. All questions were answered privately.[60]

In both public goods parts, and independently of the selected institution, subjects received feedback on the contributions of all group members in the respective period. In the sanctioning stages of Part III, subjects thereafter decided separately for each other group member whether to punish or to reward the respective co-player. At the end of the period, subjects were informed on how many punishment or reward points each group member had assigned and received as well as on the resulting period payoffs. Period payoffs were also explicitly revealed in case of the standard VCM. It was, however, not possible for a participant to identify which group member had punished or rewarded whom. Moreover, group members' IDs changed every period and made it impossible to track contribution and sanctioning behavior of a certain group member over time. Hence, retaliation motives across periods can be excluded (person i punishing person j in period t as a response to being punished by j in an earlier period).[61]

Regarding feedback in the voting stage, subjects did neither learn the exact voting result of the first voting round nor of the second voting round, but were only informed about the implemented institution and the leverage level. Moreover, they were informed about the fact

[60] An exchange rate of 1 token equalling 3 euro-cent was used for both Part II and Part III. Only for the ring test (Part I) we used a slightly different conversion rate, as 1 token corresponded to 15 euro-cent.
[61] Retaliation behavior, also known as "counter-punishment", is studied, for instance, in Denant-Boemont et al. (2007) and Nikiforakis (2008).

of a tie. This information could not be concealed for both the punishment and the reward institution as subjects of course knew whether they had voted once or twice. We therefore decided to also reveal the information in case a tie led to $L = 0$.

Table 2.1 shows for both treatments the number of independent observations. In the partner treatment, we had 120 participants in groups of four, yielding 30 statistically independent observations. In the stranger treatment we had 92 participants in matching groups of 12 (respectively 8) subjects resulting in 8 statistically independent observations.[62]

Table 2.1: Treatments and number of independent observations

Treatment	# Independent observations
Partner	30 (based on 120 participants)
Stranger	8 (based on 92 participants)

In sum, 212 participants, almost all undergraduate students studying various disciplines, took part in the computerized experiment which was conducted at the experimental laboratory MELESSA of the University of Munich in February, March and June 2009. We used the experimental software z-Tree by Fischbacher (2007) and the organizational software Orsee by Greiner (2004). Participants were randomly drawn into treatments and participated only in one session. We did not allow for any communication and decisions were taken under complete anonymity. Each session lasted for about two hours, and subjects earned 22 € on average. Payments were made in private and in cash at the end of each session.

2.5 Theoretical predictions

In this section we present theoretical predictions based on the standard *homo oeconomicus* approach (assuming selfish individuals) as well as on two prominent models of outcome-based other-regarding preferences: the inequity aversion model by Fehr and Schmidt (1999) and the social welfare model by Charness and Rabin (2002).[63] The latter two models are especially appropriate for dealing with the endogenous choice of our enforcement mechanism as they allow for deriving clear behavioral predictions. Note, however, that it is not our aim to

[62] Precisely, we had 7 matching groups of 12 subjects and one matching group of 8 subjects, the latter being caused by the non-show-up of experimental participants. Note that subjects were not informed about the formation of matching groups within a session. Sessions consisted of 24 (20) subjects.

[63] In their general model, Charness and Rabin (2002) also include intentions. However, to keep our theoretical part tractable we focus here only on their outcome-based version. See Sutter et al. (2010) for an application of this version and of the Fehr and Schmidt (1999) model in a similar endogenous setup.

provide a literal test of these theories. Rather, we try to organize our theoretical expectations under the consideration of two important behavioral motivations.

The inequity aversion model by Fehr and Schmidt (1999) states that a subject i benefits from her own material payoff π_i but suffers from payoff inequities towards the $n-1$ other members of her reference group. Precisely, the utility function looks as follows:

$$U_i(\pi) = \pi_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{\pi_j - \pi_i, 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{\pi_i - \pi_j, 0\} \qquad (2.3)$$

with $\pi = (\pi_1, \dots, \pi_n)$ denoting the vector of monetary payoffs and α_i and β_i being individual weights capturing subject's concern about disadvantageous (α_i) and advantageous inequity (β_i), respectively. The weights are restricted to $0 \leq \beta_i < 1$ and $\beta_i \leq \alpha_i$ in order to ensure that (i) subjects weakly dislike inequity, (ii) disadvantageous inequity is equally or more harmful than advantageous inequity, and (iii) subjects are not willing to burn money in order to reduce advantageous inequity. A subject's willingness to reduce payoff inequities can be a basic motivation to use sanctioning instruments. This is already shown in Fehr and Schmidt (1999) with regard to punishment. Sutter et al. (2010) extend this finding to the reward instrument and analyze subjects' endogenous choice between both sanctioning instruments. Hence, inequity aversion is an interesting motivation to study in our setup. Note that for $\alpha_i = \beta_i = 0$ we are back in the world of standard preferences.

The social welfare model by Charness and Rabin (2002) states that a subject i cares about her own monetary payoff π_i and about social welfare. Two aspects of social welfare are considered: the minimum payoff in the group and the sum of all group members' payoffs. This leads to the following utility function (cf. their Appendix 1):

$$U_i(\pi) = (1-\lambda_i)\pi_i + \lambda_i[\delta_i \min(\pi_1, \dots, \pi_n) + (1-\delta_i)(\pi_1 + \pi_2 + \cdots + \pi_n)] \qquad (2.4)$$

where $\pi = (\pi_1, \dots, \pi_n)$ indicates the vector of monetary payoffs of the n group members, and λ_i and δ_i are individual weights satisfying $\lambda_i, \delta_i \in [0, 1]$. The parameter λ_i captures a subject's preference for social welfare relative to her own payoff, while δ_i weights the importance of the "maximin" aspect ("Rawlsian" preferences) relative to the preference for group efficiency. The model includes the case of standard preferences for $\lambda_i = 0$. Since both the "maximin"-aspect and the general efficiency concern might influence subjects facing a social dilemma, these preferences provide an important complementary theoretical foundation to the inequity aversion model of Fehr and Schmidt (1999), when deriving theoretical

predictions. We will show that the social welfare concern predicts a completely different voting behavior than inequity aversion.

In the following, we describe the main predictions from our three models obtained by backward induction. Considerations are restricted to the one-shot game (implemented in our stranger design) but we will shortly discuss repeated game effects afterwards. We do not derive the formal conditions for the two models of other-regarding preferences in this section but provide the main intuition for the results. The detailed analysis can instead be found in Appendix 2B. All models assume that it is common knowledge that subjects are rational and risk neutral.

2.5.1 Predictions with standard preferences (homo oeconomicus model)

In stage two of the game, a selfish subject refrains from punishing or rewarding another person as both sanctions are costly. Thus, assuming common knowledge of selfishness, contributions in the first stage are not influenced by the implemented leverage level. As we have $\gamma < 1$, free-riding is the dominant strategy and all leverage levels yield zero contributions. Anticipating this, a subject in stage zero, i.e. in voting rounds one and two, is indifferent between the available leverage levels. Hence, all possible vote outcomes could be part of a subgame perfect equilibrium.

However, if a subject believes with a small positive probability that her group members may make mistakes, then it becomes optimal to vote for reward with the highest possible leverage level (i.e. $L = 2$). Thus, an equilibrium refinement concept like trembling-hand perfection (see Selten, 1975) rules out equilibria in which $L < 2$ is played but maintains the zero contribution and zero sanction prediction.

Predictions (Standard preferences). *Subjects are indifferent between leverage levels and any vote outcome could be part of a subgame perfect equilibrium. Moreover, the theory predicts complete free-riding ($c_i = 0\ \forall i$) and zero punishment and reward. Applying trembling-hand perfection as an equilibrium refinement, subjects are expected to vote for the reward institution and its maximum leverage of $L = 2$.*

2.5.2 Predictions with Fehr and Schmidt (1999) preferences

In contrast to the *homo oeconomicus* approach, inequity aversion is able to explain positive contributions to the public account. For the standard VCM ($L = 0$), this is due to the fact that subjects who are sufficiently averse to *advantageous* inequity will abstain from free-riding,

once their group members contribute to the public account. Hence, cooperation is possible if *all* group members care sufficiently about advantageous inequity. In Appendix 2B, it is illustrated that this is true if $\beta_i \geq 0.6\ \forall i$, i.e. if we have four so-called "conditional cooperators". However, such an event happens, taking parameter estimates from Fehr and Schmidt (1999), only in 2.56% of cases. On the contrary, if at least one member satisfies $\beta_i < 0.6$, complete free-riding is the unique equilibrium outcome for our parameter constellation and $L = 0$. In case of $L < 0$, cooperation possibilities are greatly improved as the punishment threat caused by subjects who are willing to sacrifice money to reduce *disadvantageous* inequity might credibly deter group members from free-riding. In fact, we show that one conditional cooperator can already be sufficient to enforce positive contributions from her group members. As a stronger punishment mechanism (i.e. a higher $|L|$) has more power in reducing free-riders' payoffs and, hence, inequity, higher leverage levels both raise the contribution level enforceable in equilibrium and relax the necessary restrictions on conditional cooperators' inequity aversion parameters. In Appendix 2B, taking again the parameters from Fehr and Schmidt (1999), we show that for $L = -5$ the probability of having at least one sufficiently inequity averse conditional cooperator is around 75% and thus very high. In case of $L > 0$ the expectation of receiving rewards might motivate subjects to contribute. However, for our setup with $L \leq 2$ conditional cooperators' incentives to assign rewards are too weak to make reward credible if there is at least one other subject in the group. Hence, reward can only be part of the equilibrium if $\beta_i \geq 0.6\ \forall i$ (becoming more easily sustainable and more beneficial with a higher L) and cooperation possibilities in equilibrium equal those of $L = 0$.

Putting all this together and recognizing that cooperation is hard to achieve for $L \geq 0$, subjects are expected to prefer the punishment institution and $L = -5$ as the latter (in most of the cases) yields the best cooperation and payoff possibilities. Note that voting for $L = -5$ weakly dominates the other punishment leverages already in voting round one if we assume that ties happen with a small positive probability.

Predictions (Fehr and Schmidt (1999) preferences). *If $L \geq 0$, positive contributions can only be part of an equilibrium if $\beta_i \geq 0.6\ \forall i$ (i.e., all group members are conditional cooperators) as reward by subgroups is not credible. On the other hand, one conditional cooperator can already be enough to enforce positive contributions if $L < 0$. Equilibria with positive contributions can more easily be sustained when $|L|$ increases and can support*

higher contribution levels. Hence, subjects are expected to vote for the punishment institution and its maximum leverage of $L = -5$.

2.5.3 Predictions with Charness and Rabin (2002) preferences

Charness and Rabin (2002) claim that subjects care about two aspects of social welfare: the minimum payoff in the group and the sum of payoffs. Hence, even in the absence of punishment and reward ($L = 0$) subjects have an incentive to contribute to the public account if they are sufficiently total-surplus oriented. We show in Appendix 2B that the relevant individual requirements are $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$. Subjects that fulfill these conditions, called "cooperators", contribute their entire endowment to the public account, whereas subjects with $\lambda_i < 0.5$ free-ride. The same prediction holds for the VCM with punishment ($L < 0$) as the punishment threat is not credible. The act of punishing would reduce a punisher's own payoff, the group's total payoff and perhaps even the minimum payoff in the group. Hence, no subject punishes in equilibrium, and there is no possibility to motivate relatively selfish subjects to contribute. On the contrary, the VCM with reward ($L > 0$) contains an instrument for cooperators to enforce positive contributions from the other group members. The reward instrument is credible if it sufficiently increases group's total payoff. We derive in Appendix 2B that the enforcement can entail higher contribution levels and becomes more easily achievable the higher L is. Indeed, each cooperator can credibly reward selfish subjects for $L = 2$. Hence, subjects have an incentive to vote for the maximum leverage level if the reward institution is implemented.

Combining our results, it is obvious that the reward institution can lead to equilibria with higher contribution levels and reward being given and thus to higher payoffs. Subjects are therefore expected to vote for the VCM with reward and $L = 2$. Note again that voting for the maximum leverage of $L = 2$ already weakly dominates $L = 1$ in the first voting round if ties happen with a small positive probability.

Predictions (Charness and Rabin (2002) preferences). *If* $L \leq 0$*, cooperators who satisfy* $\lambda_i \geq 0.5$ *and* $\delta_i \leq 1 - 1/2\lambda_i$ *contribute their entire endowment, while subjects with* $\lambda_i < 0.5$ *do not contribute. In case of* $L > 0$*, cooperators can enforce positive contributions from other subjects if each cooperator cares sufficiently about efficiency. The enforcement becomes more easily achievable and can entail higher contribution levels the higher* L *is. Hence, subjects are expected to vote for the reward institution and its maximum leverage of* $L = 2$.

2.5.4 Summary of our predictions

The two models based on social preferences yield opposite theoretical predictions. The model by Fehr and Schmidt (1999) reveals a tendency of subjects to vote for the punishment institution, whereas the model by Charness and Rabin (2002) provides arguments to vote for the reward institution in equilibrium. Note that with a slight refinement, standard preferences also yield that subjects should vote for the reward institution and not be indifferent between the three institutions. All considered models would then expect the highest feasible leverage to be the outcome in equilibrium. Finally, punishment is never part of an equilibrium strategy, although it might be very important as a threat that sustains cooperation in equilibrium, whereas reward is part of an equilibrium strategy with preferences à la Charness and Rabin (2002), but not according to standard preferences with trembling hands.

Before we proceed to the results of our experiment, let us briefly discuss potential differences between our stranger design and our partner design. So far, we were agnostic about potential effects of repeated play in fixed groups. Since we interpret our treatment variation primarily as an empirical robustness check, we do not present an explicit theory about the role of reputation here. Given that we face multiple equilibria with social preferences already in the one-shot game, there is a multitude of equilibria in the repeated game. For our purpose, a more detailed classification is not necessary. One can, nonetheless, speculate about behavioral effects of repeated interaction. Cooperation without enforcement could become easier to achieve. Furthermore, due to the additional presence of strategic motivations in the partner treatment, there might be a higher credibility for using punishment and reward instruments (because they serve as a signal how future behavior will be treated), especially in the first periods. As a consequence, subjects could be relatively more willing to vote either for the punishment or for the reward institution than for the standard VCM.

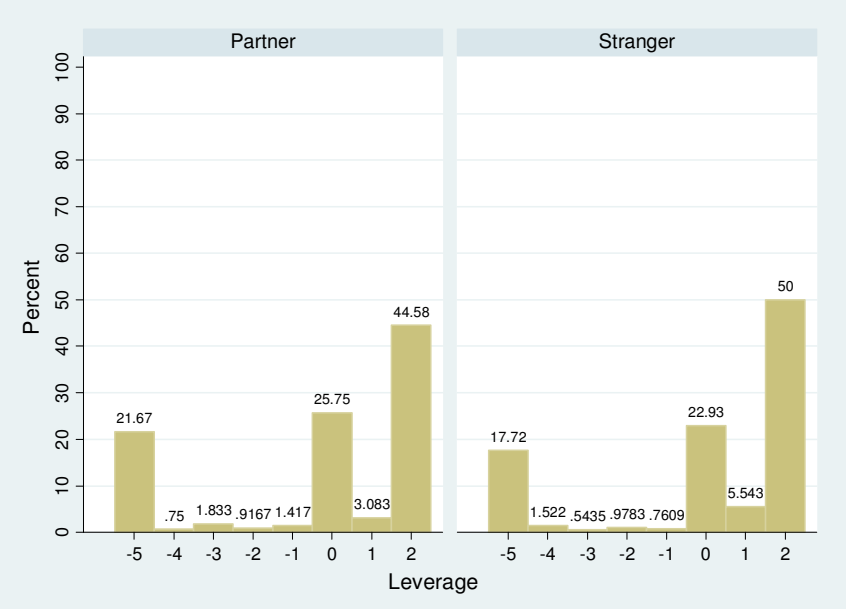
2.6 Experimental results

This section is divided into four parts. In Section 2.6.1, we present the results on individual preferences and endogenous mechanism choice within groups. In Section 2.6.2 we focus on the level of cooperation that arises out of the chosen mechanism and consider the efficiency of institutions. In Section 2.6.3, we study the sanctioning behavior of individuals. Finally, in Section 2.6.4 we take a closer look at gender effects, as they turn out to have a significant impact on voting behavior.

2.6.1 Individual preferences and endogenous mechanism choice

Remember, in the first voting round of the voting procedure participants indicate their preferred leverage level out of the interval $[-5, 2]$. Figure 2.1 reports the resulting distribution of the L parameter for both treatments. Three peaks are clearly visible: the values -5, 0 and 2 together are chosen in both treatments in more than 90% of cases and therefore much more frequently than the other values, implicating that weak enforcement mechanisms are rarely preferred, irrespective of the treatment. Hence, subjects do not shy away from entrusting their group members with strong sanctioning power if they prefer the respective sanctioning instrument. Considering our theoretical predictions the preference for $L = 2$ and $L = -5$ is not surprising at all as the former can be explained by standard preferences (if we apply trembling-hand perfection) or by Charness and Rabin (2002) preferences, while the latter is in line with the Fehr and Schmidt (1999) model. Somewhat more surprising is the large number of votes for $L = 0$. There is obviously a group of subjects that have a preference against both sanctioning mechanisms, although they have experienced in Part II of the experiment that standard VCMs perform poorly (for details on the results from Part II, see Section 2.6.2). Moreover, we will see in Section 2.6.4 that the demand for $L = 0$ is much stronger among women than among men.

Figure 2.1: Distribution of preferred leverage levels for partner and stranger treatment



As Figure 2.1 shows, the highest possible reward level ($L = 2$) is clearly the single most preferred mechanism in both treatments. Aggregating preferences at the institutional level

reveals that the VCM with reward is also the most favored institution, and this holds irrespective of the treatment (see Table 2.2). Actually, the treatment has no significant effect on subjects' voting decisions. This is confirmed by a multinomial logistic regression which uses robust standard errors clustered on the (matching) group level: Irrespective of the chosen base outcome, the treatment dummy as the only explanatory variable is never significant (each $p > 0.20$).[64] In particular, there is no evidence that the standard VCM is less attractive in the partner treatment, in contrast to the intuition formulated in Section 2.5.4. Very roughly, we can summarize our results for the institutional preferences in both treatments as follows: The VCM with reward is favored in 50% of cases, while both other institutions have around 25% support each. The predominant preference for the reward institution is in line with the social welfare model of Charness and Rabin (2002).

Table 2.2: Percentages of preferred institutions by treatment

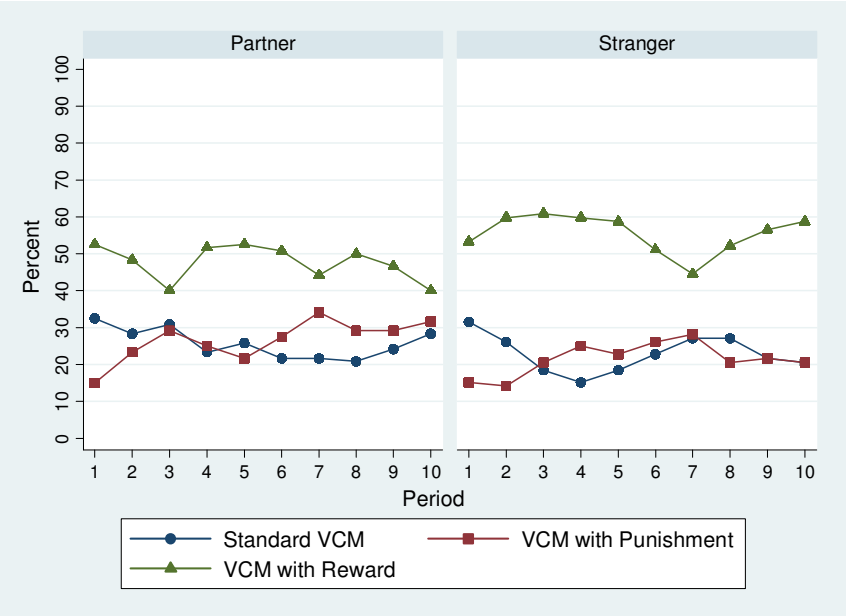
	Standard VCM	VCM with Punishment	VCM with Reward
Partner	25.75%	26.59%	47.66%
Stranger	22.93%	21.53%	55.54%

Result 2.1. $L = 2$, $L = 0$ *and* $L = -5$ *(in this order) are the three most preferred leverage levels, and the reward institution is by far the most attractive institution. There is no significant difference between treatments.*

If we have a closer look at the preferred institutions over time (see Figure 2.2), we find a similar and quite stable pattern in both treatments. The VCM with reward is the most attractive mechanism in all periods. The standard VCM is more attractive than the VCM with punishment in the first two periods, whereas from period 3 on, there is no clear difference between the two institutions anymore. If anything, the difference in preference between the VCM with reward and the other institutions is slightly smaller in the partner treatment, and we observe a bit of convergence in the last two periods. Importantly, in both treatments it is not the case that one or two institutions are eliminated over time. This means in particular that the substantial preference for $L = 0$ depicted in Figure 2.1 does not vanish over time. Even in period 10, 20 – 30% of subjects reject a sanctioning institution.

[64] Note that by using multinomial logistic regressions (see also Table 2.3), we do not have to impose any institutional ordering. In particular, we do not assume that the standard VCM ($L = 0$) serves as an "intermediate" category. In fact, individual's switching behavior across periods (cf. Figures 2C.2 and 2C.3 in Appendix 2C) shows relatively few direct switches between the standard VCM and the punishment institution, indicating that the reward institution is the much more natural "intermediate" institution.

Figure 2.2: Preferred institutions over time



If we, however, consider preferences for extreme values of the sanctioning instruments relative to the attractiveness of the respective sanctioning institution, we find a clear development. As Figure 2.3 shows, $L = -5$ and $L = 2$ are chosen much more often in later periods than weaker institutions. Starting from percentages partially far below 90% they climb in both treatments to almost 100%. Our explanation is that group members experience that stronger enforcement mechanisms have more power in increasing group cooperation, a fact which we will confirm empirically below. The decline of votes for weak sanctioning instruments is in line with our behavioral theories which suggest that subjects prefer extreme values of the sanctioning technologies to increase prospects for cooperation. However, the development in the reward institution could also simply be explained by standard preferences as voting for $L = 2$ becomes optimal, once subjects believe that group members may tremble in the reward stage.

Result 2.2. *Aggregated institutional preferences are quite stable over time. Especially, the reward institution stays the most favored institution in all periods of both treatments. However, subjects increasingly vote for extreme values of the strength of the enforcement mechanism when they prefer the respective sanctioning institution.*

Figure 2.3: Relative development of extreme preferences over time

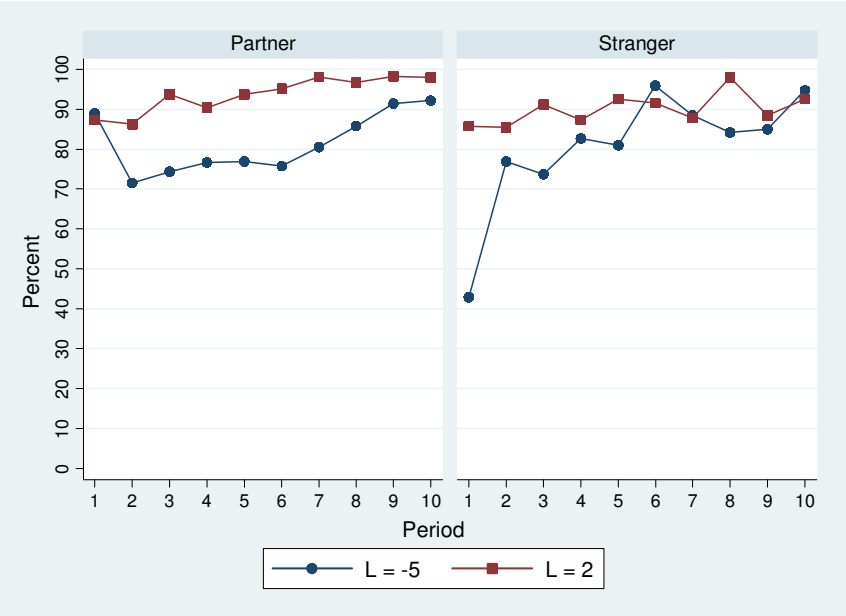


Table 2.3 reports additional evidence from multinomial logistic regressions on subject's institutional choice, again using robust standard errors clustered on the (matching) group level. We report regressions for the partner (columns 1-3) and the stranger treatment (columns 4-6) as well as a combined version (columns 7-9). While the first two columns under each heading have the standard VCM as the base category (denoted by *S*) and report results for the punishment (*P*) and the reward institution (*R*), respectively, the third column presents the punishment institution in comparison to the VCM with reward. As covariates we include a time trend (*Period*), a gender dummy (*Woman*), a dummy controlling for subject's general willingness to cooperate (*Cooperative*), a variable capturing earnings in Part II of the experiment (*Profit_Part2*), and, in the combined version, additionally a treatment dummy (*Partner*). Note that *Profit_Part2* and *Cooperative* are variables which we obtain from earlier parts of the experiment. The former covariate controls for subject's experience with the standard VCM in Part II. The latter covariate stems from the social value orientation questionnaire in Part I and accounts for self-selection effects which might occur due to different individual preferences for cooperation.[65]

[65] Results from the social value orientation questionnaire are presented in Appendix 2C (see Figure 2C.1). 22 participants (10.38%) are excluded from the analysis as they have a consistency index below 20. From the resulting sample (190 participants), roughly speaking, 2/3 can be classified as individualistic and 1/3 as cooperative, while all other motivations are almost non-existent (only five subjects). Figure 2C.1 further shows that the distribution of types does not differ across treatments (confirmed by a χ^2-test on the first two categories: $p = 0.70$). For our regressions in Table 2.3, we introduce a type dummy allowing *only* for individualistic and cooperative subjects. Hence, we exclude the five subjects from the analysis who show a different motivation.

Results of Table 2.3 confirm that the time trend plays a minor role. While there is no significant development in the stranger treatment, the punishment institution gains significant support in the partner treatment but only compared to the standard VCM. In fact, if one takes a second look at Figure 2.2, one can see a slight upward trend for the punishment institution, ranging from 15.00% acceptance in period 1 to 31.67% in period 10. However, there is no further increase from period 7 onwards; thus, the effect would not become stronger if we allowed for a longer time horizon. Finally, the combined model also confirms that the punishment institution gains support. Hence, the VCM with punishment becomes a bit more prevalent over time compared to the two other mechanisms.

A much clearer effect than for the time trend can be found for the gender dummy. We will discuss the issue in detail in Section 2.6.4. Subjects' personal motivations as elicited by the ring test do not matter. This means that cooperative subjects do not show a different voting behavior than individualistic subjects. This result is robust to specifications using other consistency thresholds or the exact motivational degrees of the ring test. There is no self-selection of behavioral types into specific institutions in our data.

In contrast, subjects' earnings in Part II matter. Results show that a lower total profit in Part II can explain why subjects prefer a sanctioning institution in the course of Part III. In the partner treatment, subjects with a worse experience vote significantly more often for both the punishment and the reward institution (compared to the standard VCM), while subjects in the stranger treatment do not trust the reward mechanism. This is evidence that frustration about the (poor) outcome in social dilemmas in the past is one major reason for subjects to try to implement a sanctioning institution.

Finally, the combined model in columns 7-9 confirms a result already mentioned above: the treatment dummy is not significant and, hence, there is no evidence that repeated interaction versus one-shot interaction has a significant influence on the voting behavior.

Result 2.3. *Regression results reveal that the punishment institution, at least in the partner treatment, gains slightly increasing support over time. Moreover, a subject's past experience with social dilemmas influences the voting decision and makes the implementation of a sanctioning institution more likely. Self-selection of behavioral types into specific mechanisms is not observed.*

Note that we use a common consistency requirement and that none of the reported results would change by demanding a higher or allowing for a lower threshold.

Table 2.3: Explaining institutional choice (multinomial logistic regressions)

	Dependent variable: Preferred institution								
	Partner			Stranger			Combined		
	Base: S		Base: R	Base: S		Base: R	Base: S		Base: R
	P	R	P	P	R	P	P	R	P
Period	0.093 ***	0.028	0.066	0.038	0.014	0.024	0.070 ***	0.021	0.049 *
	(0.035)	(0.047)	(0.043)	(0.026)	(0.030)	(0.028)	(0.023)	(0.027)	(0.027)
Woman (= 1)	-1.874 ***	-0.928 **	-0.945 **	-1.513 ***	-0.757 ***	-0.756	-1.689 ***	-0.770 ***	-0.919 ***
	(0.407)	(0.384)	(0.387)	(0.456)	(0.235)	(0.587)	(0.295)	(0.244)	(0.319)
Cooperative (= 1)	-0.043	-0.010	-0.033	0.211	0.455	-0.244	0.028	0.120	-0.091
	(0.473)	(0.368)	(0.348)	(0.507)	(0.605)	(0.482)	(0.341)	(0.306)	(0.251)
Profit_ Part2	-0.024 ***	-0.017 ***	-0.008	-0.019 ***	0.002	-0.021 ***	-0.022 ***	-0.011 **	-0.011 ***
	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.008)	(0.004)	(0.005)	(0.004)
Partner (= 1)	-	-	-	-	-	-	0.282	-0.106	0.388
							(0.350)	(0.294)	(0.288)
Constant	6.466 ***	5.244 ***	1.222	4.768 ***	0.689	4.079 **	5.640 ***	3.827 ***	1.814 *
	(1.448)	(1.341)	(1.294)	(1.731)	(1.183)	(1.800)	(1.189)	(1.050)	(1.033)
# Obs.	1010	1010	1010	840	840	840	1850	1850	1850
# Ind. Obs.	30	30	30	8	8	8	38	38	38
Pseudo R^2		0.079			0.057			0.065	

Notes: Robust standard errors in parentheses (clustered on (matching) group level). *** Significant at 1% level; ** significant at 5% level; * significant at 10% level. 27 subjects excluded according to the ring test (19 in Partner and 8 in Stranger). *S* denotes standard VCM, *P* VCM with punishment and *R* VCM with reward.

Coming to the individual level, one might ask how many subjects change their institutional choice over the course of the experiment and whether they switch between all three available institutions or only between two of them. This is especially interesting as we do not observe much of a development on the aggregate level. Note that we provide detailed information about the voting behavior of each single subject in Appendix 2C. Table 2.4 summarizes this information by forming three groups of subjects: subjects that never change their institutional preference (row 1), subjects that switch between two institutions (row 2) and subjects that vote for all three institutions over the course of the experiment (row 3). The letter *S* stands for subjects that prefer the standard VCM in at least one period and, respectively, *P* denotes the VCM with punishment and *R* the VCM with reward. For example, a subject classified in *SR* never votes for the VCM with punishment but prefers the standard VCM and the VCM with reward each in at least one period. Table 2.4 provides the frequencies for the different voting patterns. It is apparent that in both treatments there is

roughly one quarter of subjects that vote for the same institution all the time. We can interpret them as having a strict institutional preference. Second, about half of the subjects switch between two institutions. However, switches predominantly occur either between the punishment and the reward institution or between the reward institution and the standard VCM. While we can interpret the former group of subjects as persons with a clear preference for a sanctioning mechanism, the latter group systematically rejects a punishment mechanism. Interestingly, almost nobody switches solely between the punishment institution and the standard VCM. Finally, one quarter of subjects alternates between all three institutions. A χ^2-test confirms that the distribution of voting patterns is not significantly different between our two treatments ($p = 0.37$).

Table 2.4: Frequency of voting patterns

	Partner	Stranger
S, P, R	7, 4, 12 (= 23)	7, 3, 15 (= 25)
SP, PR, SR	7, 33, 26 (= 66)	2, 18, 24 (= 44)
SPR	31	23
Sum	*120*	*92*

Notes: *S* denotes standard VCM, *P* VCM with punishment and *R* VCM with reward.

Overall, we observe a quite substantial average institutional switching rate of 2.43 in the partner treatment and 2.21 in the stranger treatment showing subjects' general willingness to try different institutional settings. Again, the treatment difference is not statistically significant ($p = 0.28$; Mann-Whitney U-test, $N = 38$). We also checked for a time trend in the number of institutional switches (see Figure 2C.4 in Appendix 2C) but do not find any tendency. In both treatments between 20-35% of subjects change their institutional choice from one period to the next and this rate does not decrease towards the end.

Result 2.4. *In both treatments we roughly observe 25% of subjects never changing their institutional choice, 50% switching between two institutions and 25% alternating between all three available options. Institutional switching rates are quite substantial and do not decrease over time.*

Finally, considering voting behavior in the second voting round, we can study how subjects change their decision in case there is a relative majority for the VCM with punishment or the VCM with reward within their group. Note, however, that we have only selected data on this issue as not all subjects enter the second voting round in a given period

and their choice set depends on the group decision in voting round one. Hence, the following descriptive results provide only a very rough but nevertheless insightful overview. Two cases can be distinguished:

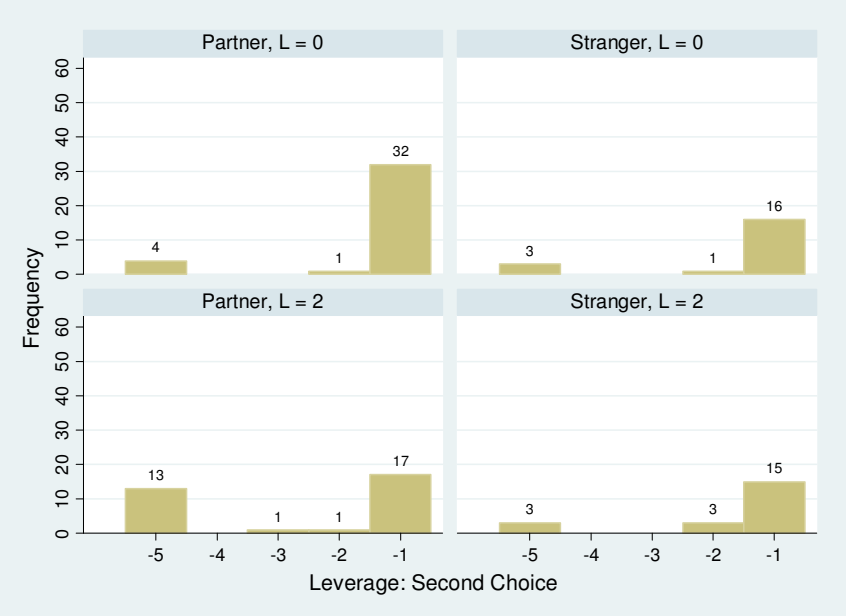
First, if the respective subject is part of the relative majority, then there is no need for this person to change her decision as her first choice is still available within the reduced interval. Nevertheless, each subject has to vote again, and it is interesting to see whether subjects stay with their preference or change it. We find that in 468 out of 503 cases (93.04%) in the partner treatment and in 401 out of 435 cases (92.18%) in the stranger treatment subjects do not change their vote. While there is obviously no difference between the two treatments, changes are slightly more often observed if the group opted for the punishment institution (30/182: 16.48%; pooled data) than for the reward institution (39/756: 5.16%; pooled data). One explanation for this phenomenon is that in the punishment case there is simply a much larger set of alternative options. Overall, however, this result is good news. It suggests that subjects take their first voting round decisions seriously and that there is no large fraction of subjects who do not state their true preferences in voting round two but vote strategically.

Second, if the respective subject is not part of the relative majority, then her first choice is no longer available in the reduced interval of voting round two, and the subject has to choose her preferred leverage level out of a different mechanism. In case the VCM with punishment was chosen in the first voting round (see Figure 2.4), subjects preferring the standard VCM or the VCM with reward have to pick a value out of the negative sub-interval of L. Note that we do not have too many data points for this case but the results are quite striking. Focusing on cases in which subjects preferred the standard VCM ($L = 0$), 16 out of 20 (80.00%) cases in the stranger treatment and 32 out of 37 (86.49%) in the partner treatment show a preference for the lowest possible leverage level. Thus, a high effectiveness level of the punishment instrument is very unattractive for these subjects. While we observe a similar result for subjects preferring a high level of reward ($L = 2$) in the stranger treatment, the preference is somewhat less obvious for $L = 2$-loving subjects in the partner treatment because in this case we observe 13 out of 32 votes (40.63%) for the strongest possible sanctioning mechanism.[66] The unattractiveness of high effectiveness levels of the mechanism for overruled subjects can be explained both by Charness and Rabin (2002) and by standard preferences. Remember that these models predict a preference for $L = 2$ in the first voting round (for standard preferences, only if we apply trembling-hand perfection). Charness and Rabin (2002) preferences then

[66] We ignore subjects who have chosen $L = 1$ in the first voting round because we have too few observations. For the same reason we neglect subjects voting for $-4 \leq L \leq -1$ in Figure 2.5.

postulate indifference in the negative sub-interval of L in voting round two due to the incredibility of using the punishment instrument (see Appendix 2B). The same result is obtained when using standard preferences. However, if we apply trembling-hand perfection both models would predict $L = -1$ as this minimizes the impact of potential punishment.

Figure 2.4: Choice of negative L by subjects not preferring VCM with punishment



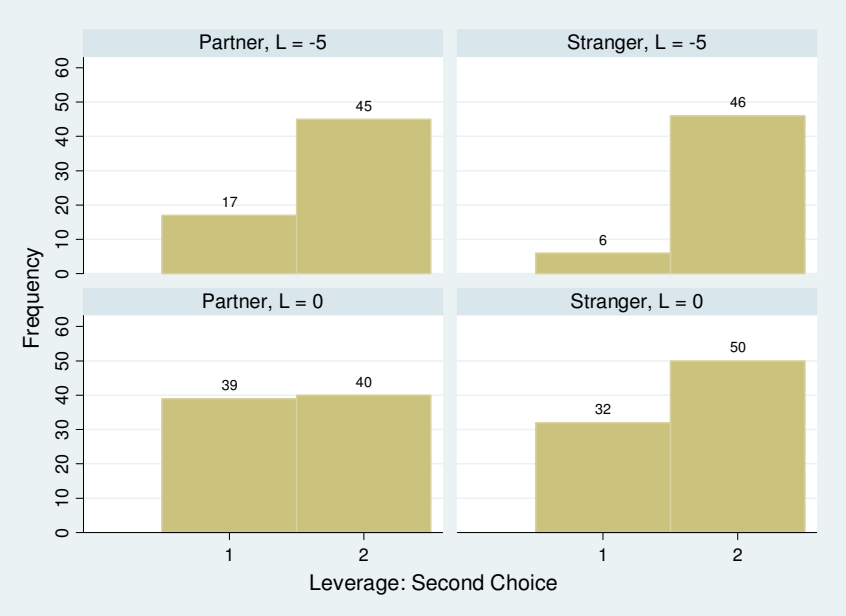
In case the VCM with reward was chosen in the first voting round (see Figure 2.5), subjects who voted for the standard VCM or the VCM with punishment have to switch to a positive leverage level in the second voting round. We find that a high level of effectiveness ($L = 2$) is very attractive for subjects who preferred $L = -5$. Indeed, we observe 45 out of 62 (72.58%) votes for $L = 2$ in the partner treatment and 46 out of 52 (88.46%) in the stranger treatment. The same tendency holds for $L = 0$-loving subjects in the stranger treatment (50/82: 60.98%), while for subjects who have wished to play the standard VCM in the partner treatment, $L = 2$ is not more attractive than $L = 1$. Can we explain the observed preferences of overruled subjects by our theory? Fehr and Schmidt (1999) preferences suggest that subjects vote for $L = -5$ in voting round one in order to maximize contribution incentives. However, if forced to play the VCM with reward, conditional cooperators might want to switch to $L = 2$ in order to maximize the likelihood and impact of mutual reward. As mutual reward cannot be part of the equilibrium if there is at least one other group member, subjects with $\beta_i < 0.6$ are completely indifferent. If we apply trembling-hand perfection there might

also be a reason for inequity averse subjects to vote for $L = 1$ as this minimizes payoff differences due to an accidental use of the sanctioning instrument.

To sum up, the analysis of the second choice of overruled subjects reveals an important asymmetry: a high level of effectiveness of the punishment (reward) institution is unattractive (attractive) for subjects originally preferring another institution. Moreover, subjects with an original preference for the standard VCM tend to prefer lower sanctioning levels more often than subjects with an original preference for the converse sanctioning institution.

Result 2.5. *The analysis of the second voting round indicates that subjects rarely change their choice if there is no need to do so. For overruled subjects a high level of effectiveness of the punishment (reward) institution is unattractive (attractive).*

Figure 2.5: Choice of positive L by subjects not preferring VCM with reward



2.6.2 Cooperation and the efficiency of institutions

Table 2.5 shows average contributions and profits in the three different institutions of Part III, separately for the partner and the stranger treatment. We always add the number of observations in parentheses. For reasons of comparison, results of the standard VCM in Part II are also included. We find that subjects contribute in this part, on average, roughly 25% of their endowment. Hence, our subjects make the typical experience that the level of

cooperation in a standard linear public goods game is rather low. Not surprisingly, average contributions in the standard VCM of Part III are significantly lower than in Part II.[67]

Concentrating only on Part III, we find that the chosen institution has a large influence on contribution levels. Average contributions in both treatments are highest in the VCM with punishment, intermediate in the VCM with reward and lowest in the standard VCM. Each pairwise comparison is significant, at least at the 10% level. Thus, the possibility to reward or punish increases the level of cooperation in comparison to the standard VCM, and the VCM with punishment leads to more cooperation than the reward institution. This is in line with the findings of Sutter et al. (2010). Positive contribution effects of the sanctioning institutions can partially be explained either by Fehr and Schmidt (1999) preferences (for the VCM with punishment) or by outcome-based Charness and Rabin (2002) preferences (VCM with reward). Both theories, though, fail to predict the entire data pattern. Furthermore, we find that all institutions have slightly higher averages in the partner than in the stranger treatment but none of the pairwise differences between treatments is significant. Hence, repeated interaction does not influence cooperation levels in a significant way.

Figures 2C.5 and 2C.6 in Appendix 2C show the average contribution pattern over time for the partner treatment and the stranger treatment, respectively. They confirm that the observed differences in contributions between the three institutions of Part III hold in almost all periods. Furthermore, all institutions (in both parts) show a decline in contributions over time, indicating that even our sanctioning mechanisms are, on average, too weak to sustain high levels of cooperation in the long run.[68] This might explain why we observe quite a few preference switches regarding the institution on the individual level.

By looking at average profits we observe that higher contributions in the standard VCM of Part II compared to Part III translate into higher profits (which is true by construction). For the institutions of Part III we obtain an interesting picture. Although contributions are clearly highest in the VCM with punishment, profits are lower than in the two other institutions, and they are even below the initial endowment. This is true for both treatments. Thus, from an efficiency point of view, choosing the VCM with punishment is not a good idea. The highest average profits, instead, are found for the VCM with reward. Comparing our two treatments we observe that profits are always slightly higher in the partner treatment, but pairwise comparisons never yield significant differences.

[67] Note that all significances shown in Tables 2.4 and 2.5 are computed by OLS regressions that use robust standard errors clustered on the (matching) group level and that have the respective dummy variable as the only regressor. Random effects and tobit regressions yield similar results.

[68] There is some more noise in the stranger treatment. However, we find no reason to believe in any difference between treatments.

Result 2.6. *Contributions and profits depend significantly on the implemented institution. While the punishment institution leads to the highest levels of cooperation, choosing the reward institution maximizes average profits. This is true for both treatments, and there is no significant difference between them.*

Table 2.5: Average contributions and profits in the different institutions by treatment

		Standard VCM – Part II	Standard VCM – Part III	VCM with Punishment	VCM with Reward
Contributions	Partner	$5.73^{\#}$ (N = 1200)	$2.42^{\#*\dagger}$ (N = 300)	8.29^{*+} (N = 236)	$5.28^{\dagger+}$ (N = 664)
	Stranger	$4.53^{\#}$ (N = 920)	$2.31^{\#*\wedge}$ (N = 156)	$6.83^{*\ddagger}$ (N = 160)	$3.66^{\wedge\ddagger}$ (N = 604)
Profits	Partner	$23.44^{\#}$ (N = 1200)	$21.45^{\#\circ\S}$ (N = 300)	$19.69^{\circ\ddagger}$ (N = 236)	$23.91^{\S\ddagger}$ (N = 664)
	Stranger	$22.72^{\#}$ (N = 920)	$21.39^{\#\circ\S}$ (N = 156)	$18.82^{\circ\ddagger}$ (N = 160)	$22.67^{\S\ddagger}$ (N = 604)

Notes: Significant difference between Standard VCM - Part II and Standard VCM - Part III: $^{\#}$ $p < 0.01$; between Standard VCM – Part III and VCM with Punishment: * $p < 0.01$ and ° $p < 0.05$; between Standard VCM – Part III and VCM with Reward: § $p < 0.01$, † $p < 0.05$ and ^ $p < 0.10$; between VCM with Punishment and VCM with Reward: ‡ $p < 0.01$ and $^{+}$ $p < 0.05$.

In Figure 2.6 we take into account that the implemented leverage level differs within the punishment and the reward institution and that these differences may influence contribution levels. Note that this is suggested by our other-regarding preferences models (see Section 2.5). Moreover, such technology effects were found in previous studies varying leverage levels exogenously. For example, for the punishment institution, Nikiforakis and Normann (2008) in a partner matching and Egas and Riedl (2008) in a perfect stranger design showed that average contributions increase in the effectiveness level. The same result was observed for the reward institution; see e.g. Sutter et al. (2010) for a partner design. Hence, we expect to observe a leverage effect also in our endogenous setting. And, indeed, Figure 2.6 reveals that the leverage level has a strong impact on contributions. Contributions clearly increase in the leverage level of both institutions, irrespective of the treatment. The ups and downs depicted in Figure 2.6 are due to the fact that we have few observations for some parameter values. One must therefore be careful when interpreting single dots in the figure, and a regression analysis is required to get a better impression of the data.

Figure 2.6: Average contributions depending on implemented leverage level

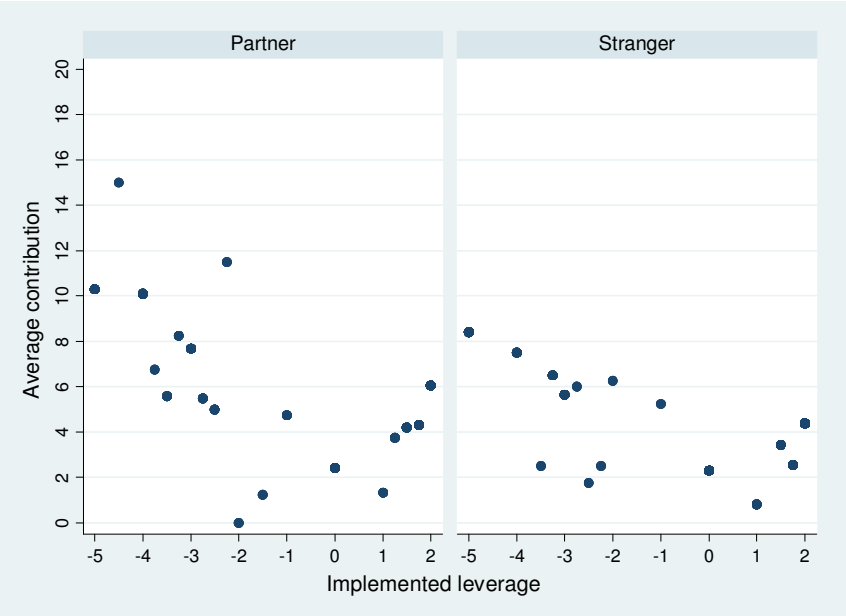


Table 2.6 presents three OLS regressions explaining contributions in Part III. We use again robust standard errors clustered on the (matching) group level. The regression model in the first column is based on data from the partner treatment. We include the time trend *Period* which appears to be negative and significant, implying that contributions are significantly lower in later periods. To control for our different institutions and leverage levels, we define two further variables. |*Leverage_neg*| captures the implemented negative leverage levels (in absolute values) and *Leverage_pos* the positive ones. Both variables are set to zero if the respective institution is not implemented. Furthermore, a constant is added to the model that captures contribution levels in the standard VCM. We choose this model, since Figure 2.6 suggests that both larger negative and larger positive deviations from $L = 0$ lead to increasing contribution levels (see Fehr and Fischbacher, 2004 for the use of such an approach in a different setting). Column 1 in Table 2.6 shows that the constant is highly significant, indicating that contribution levels in the standard VCM are, at least in the beginning, significantly different from zero. Moreover, both leverage variables are highly significant and the signs confirm the main message from Figure 2.6 that contributions increase with stronger sanctioning mechanisms.

Column 2 in Table 2.6 provides the results for the same regression for the stranger treatment, yielding basically the same results. Again, the two leverage variables, the constant and the time trend are highly significant and have the same sign as in the partner treatment. A comparison of the coefficients in the two treatments further shows that the effect of increasing

leverage levels is somewhat smaller in the stranger treatment. To find out whether the difference in coefficients is significant, we conduct a third regression (column 3 in Table 2.6) where we combine both data sets. We add a treatment dummy (*Partner*) and two interaction terms (|*Leverage_neg*| * *Partner*, *Leverage_pos* * *Partner*) that connect the treatment dummy with our leverage variables. Results from column 3 reveal that the treatment dummy is insignificant, i.e. there is no difference in contributions between the partner and the stranger treatment for $L = 0$. While *Leverage_pos* * *Partner* is not significant either, we find a weakly significant effect for the punishment institution (p = 0.09). Hence, there is at least some evidence that an increase in the punishment leverage has a larger effect on contributions in the partner treatment. To conclude, a regression analysis confirms that contribution levels depend crucially on the combined realization of the institution *and* the effectiveness level.

Result 2.7. *Contribution levels significantly increase in the leverage levels of both sanctioning institutions, but they decrease over time. There are no clear differences between our two treatments.*

Table 2.6: Contributions to the public account (OLS regressions)

	Dependent variable: Contributions		
	Partner	Stranger	Combined
Period	-0.587***	-0.472***	-0.537***
	(0.099)	(0.057)	(0.061)
|Leverage_neg|	1.850***	1.289***	1.294***
	(0.271)	(0.184)	(0.172)
Leverage_pos	1.827***	0.849***	0.840***
	(0.645)	(0.245)	(0.233)
Partner (= 1)	-	-	-0.080
			(0.822)
|Leverage_neg| *	-	-	0.546*
Partner			(0.318)
Leverage_pos *	-	-	0.973
Partner			(0.685)
Constant	5.217***	4.677***	5.042***
	(0.785)	(0.863)	(0.771)
# Observations	1200	920	2120
# Ind. observations	30	8	38
R^2	0.142	0.117	0.139

Notes: Robust standard errors in parentheses (clustered on (matching) group level). *** Significant at 1% level; ** significant at 5% level; * significant at 10% level.

2.6.3 The sanctioning behavior of individuals

In Table 2.7, we present information on how often the sanctioning instruments are in fact used if they are available. We find that, on average, in 38.56% (36.25%) of cases a subject punishes another group member in the partner (stranger) treatment. The difference between the two treatments is not significant ($p = 0.53$).[69] In contrast, we find a significant treatment effect for the use of reward ($p = 0.05$), which is more often observed in the partner (27.26%) than in the stranger treatment (17.94%). Hence, there is only partial evidence for the prediction stated in Section 2.5.4 that sanctioning instruments could be more frequently applied in a fixed group structure.

Overall, in both treatments the punishment instrument is used significantly more often than the reward instrument ($p < 0.05$ each). The relatively high percentages of punishing explain why profits are lower in the VCM with punishment than in the other institutions, although, on average, contribution levels are higher. On the other hand, the relatively low reward levels provide a possible explanation for why some subjects prefer to cast their vote for the standard VCM instead of the reward institution: they could want to avoid feelings of disappointment due to not being rewarded. Note that with the other-regarding preferences models presented in Section 2.5 we cannot explain the use of the punishment instrument. While outcome-based Charness and Rabin (2002) preferences suggest that punishment is never used, Fehr and Schmidt (1999) preferences also predict zero punishment in equilibrium. On the contrary, the Charness and Rabin (2002) model is able to support equilibria in which subjects reward their group members.

Table 2.7: Use of punishment and reward instruments by treatment

	VCM with Punishment	VCM with Reward
Partner	38.56%$^{+}$ (N = 708)	27.26%$^{\$+}$ (N = 1992)
Stranger	36.25%‡ (N = 480)	17.94%$^{\$\ddagger}$ (N = 1812)

Notes: Significant difference between Partner and Stranger: $^{\$}$ $p < 0.10$; between VCM with Punishment and VCM with Reward: ‡ $p < 0.01$ and $^{+}$ $p < 0.05$.

To get a more detailed view on subject's sanctioning behavior, we run probit regressions with more controls separately for the punishment and the reward decisions. Table 2.8 reports our results by presenting marginal effects along with robust standard errors that are, once again, clustered on the (matching) group level. Columns 1-3 focus on the decision to punish

[69] Note that all significances shown in Table 2.7 are computed by probit regressions that use robust standard errors clustered on the (matching) group level and that have the respective dummy variable as the only regressor.

in the partner treatment (column 1) and in the stranger treatment (column 2) as well as in a combined model (column 3), while columns 4-6 present the respective regression models for the reward decision. As covariates we include the time trend *Period*, the average contribution level within the group (*Average group contribution*), the individual other-own difference in contributions (*Contribution difference (other – own)*), the two leverage variables introduced above (|*Leverage_neg*|, *Leverage_pos*), and for the combined models the treatment dummy *Partner*.

Our results do not reveal a significant time trend for the punishment decision, i.e. punishment activities are relatively stable over time. On the contrary, we find evidence that subjects reward less in later periods. While the effect is insignificant in the partner treatment, it is weakly significant in the stranger treatment and even more so in the combined model. In line with Sutter et al. (2010), we find that an increase in the average contribution level tends to decrease punishment activities (although not significantly), but it increases rewarding behavior. Moreover, our data confirm Sutter et al.'s (2010) finding that the use of both sanctioning instruments critically hinges on the individual other-own difference in contributions. The more a group member j contributes compared to i, the lower is the probability that i punishes j and the more likely will she reward. Surprisingly, both of our leverage variables do not explain the sanctioning behavior in the respective institution. In contrast to the existing literature reporting linear or even quadratic effects of exogenous price and effectiveness variations on sanctioning decisions (e.g. Anderson and Putterman, 2006; Carpenter, 2007a; Nikiforakis and Normann, 2008), we fail to replicate any relationship in our endogenous setup. However, this could also be due to the fact that we use a binary punishment decision and do not allow for the assignment of multiple punishment points.

Finally, the treatment dummy is insignificant concerning both sanctioning decisions, pointing out that we do not observe more sanctioning activities in the partner treatment. Remember that in the absence of further controls we observed a significant difference for the reward decision (see Table 2.7). We can show that this effect is caused by small differences in the average contribution levels between treatments (cf. Table 2.5) and that it vanishes when *Average group contribution* is included in the regression. Hence, the strategic motivation behind the use of sanctioning instruments is not important enough in our context to create significant differences. From an ex post perspective this explains why we do not observe a significant difference in the contribution decisions and in the voting behavior between our two treatments.

Table 2.8: Sanctioning behavior (probit regressions)

	Dependent variable: Decision to…					
	punish			reward		
	Partner	Stranger	Combined	Partner	Stranger	Combined
Period	-0.009	0.003	-0.006	-0.012	-0.009*	-0.010**
	(0.013)	(0.016)	(0.010)	(0.008)	(0.005)	(0.005)
Average group contribution	-0.007	-0.003	-0.006	0.027***	0.024***	0.025***
	(0.007)	(0.013)	(0.006)	(0.006)	(0.003)	(0.004)
Contribution difference	-0.023***	-0.028***	-0.025***	0.010***	0.008***	0.009***
(other – own)	(0.005)	(0.004)	(0.004)	(0.001)	(0.002)	(0.001)
|Leverage_neg|	0.028	0.007	0.020	-	-	-
	(0.027)	(0.041)	(0.022)			
Leverage_pos	-	-	-	0.145	-0.055	0.038
				(0.093)	(0.049)	(0.061)
Partner (= 1)	-	-	0.051	-	-	0.050
			(0.046)			(0.033)
# Observations	708	480	1188	1992	1812	3804
# Ind. observations	23	7	30	28	8	36
Pseudo R^2	0.139	0.200	0.161	0.168	0.116	0.154

Notes: Marginal effects shown. Robust standard errors in parentheses (clustered on (matching) group level). *** Significant at 1% level; ** significant at 5% level; * significant at 10% level.

Result 2.8. *Punishment activities are more frequent than reward decisions. The use of both sanctioning instruments crucially depends on individual contribution differences within the group. Furthermore, we find a significant impact of both the average contribution level and the time trend on the decisions to reward. The treatment itself has no significant effect.*

2.6.4 Gender effects

Interestingly, there are strong gender effects on a subject's voting decision in the first voting round as Table 2.9 indicates. Women prefer the standard VCM much more often than men do. Men, on the contrary, prefer the punishment instrument more than women. Both facts hold irrespective of the treatment. However, the percentage of subjects preferring the reward institution differs only slightly between both sexes and is around 50% in both treatments. Note further that both sexes share the preference for the three leverage peaks observed in Figure 2.1 (we omit the corresponding figure split up according to sex).

If we consider the multinomial logistic regressions for the institutional choice shown in Table 2.3, the regression results confirm a highly significant impact of the gender dummy. In both treatments women choose sanctioning institutions (compared to the standard VCM) significantly less often than men. Moreover, both in the partner treatment and in the combined model we find evidence that women vote relatively less often for the punishment institution

when compared to the reward institution. All these effects are significant at the 5% or even at the 1% level.

Table 2.9: Percentages of preferred institutions by gender and treatment

		Standard VCM	VCM with Punishment	VCM with Reward
Partner	Man	17.42%	36.72%	45.86%
	Woman	33.55%	17.10%	49.35%
Stranger	Man	13.94%	33.94%	52.12%
	Woman	27.97%	14.57%	57.46%

Result 2.9. *In both treatments women vote significantly less for the punishment institution but show a much larger preference for the standard VCM than men. Both sexes have in common that one half of their members prefer the reward institution.*

Finally, we consider gender effects in subjects' contribution and sanctioning behavior. The results can be shortly summarized as follows: We typically observe that women tend to contribute more than men in the standard VCM, less than men in the punishment institution, and that there is almost no difference for the VCM with reward. This suggests that preferences at least partly take over into actual behavior in the respective institution. However, significant contribution differences can only be shown for the punishment institution by either running simple regressions with a gender dummy analogous to those conducted in Table 2.5 (is only significant for the partner treatment; $p < 0.01$) or by including gender variables in the regressions of Table 2.6, additionally controlling for the leverage level (provides a significantly weaker increase for |*Leverage_neg*| in both treatments; $p < 0.05$ each).

Regarding sanctioning behavior, we find that including gender in the regressions of Table 2.8 yields insignificant results for the punishment institution. However, if we replace *Contribution difference (other – own)* by the gender dummy, we find women to punish significantly less often than men both in the partner treatment and in the combined model (both $p < 0.02$). Hence, women punish less because they contribute, on average, lower amounts. On the contrary, women reward significantly more than men in the partner treatment ($p < 0.05$), irrespective of whether we control for contribution differences or not. No significant effect can be found for the stranger treatment and the combined model.

2.7 Conclusion

This chapter deals with the repeated endogenous choice of enforcement mechanisms in a social dilemma. In particular, we let subjects cast their vote between a standard voluntary contribution mechanism (standard VCM), a VCM with informal punishment opportunities and a VCM with informal reward opportunities. For the punishment and reward institutions, subjects additionally have to state their preferred strength of the enforcement mechanism out of a given set of "leverage" levels. We exogenously vary the time horizon of the interaction (partner versus stranger design) to control for effects of reputation formation on endogenous choice.

Our setup considerably extends a recent study of Sutter et al. (2010) by the following important features. First, the repeated vote on the institution allows controlling for learning effects over the course of the experiment. Second, we are the first who combine a vote on institutions with a choice on the strength of the enforcement mechanism. The latter design feature captures the important fact that subjects in real life social dilemmas are usually able to influence the strength of sanctions, as it is the case, for example, in the evolution of social norms. Additionally, we are the first to provide a robustness check concerning the effects of repeated versus one-shot interaction by implementing a partner and a stranger design in our experiment. A priori it is conceivable that the choice of institution differs across the two treatments. To sum up, Sutter et al. (2010) have a strong focus on the comparison between exogenously implemented institutions and endogenously selected ones, whereas the design of the current chapter is much more suitable to study the details of individual preferences over the three institutions and their effectiveness levels.

Our results show that one half of the population prefers the VCM with reward while one quarter each votes for the punishment institution and the standard VCM. This aggregate pattern is astonishingly stable over time - although we observe substantial switching behavior on the individual level - and it differs only slightly across treatments. The higher preference for the reward institution is justified as it is the most efficient institution. This is true although our reward leverage is restricted to lower levels than the punishment instrument's leverage and although the VCM with punishment leads to significantly higher contributions. In other words, we have made the reward mechanism less attractive, but it is still clearly the preferred mechanism and the most successful one in terms of efficiency. Moreover, we find subjects to very strongly prefer extreme values of both sanctioning technologies, and contributions to increase in the levels of effectiveness. There is no significant self-selection effect into a

specific mechanism caused by subjects' basic inclinations to cooperate. However, our results reveal that women prefer the standard VCM relatively more often, while men show a relatively stronger preference for the punishment instrument. Both sexes nevertheless have in common that a majority of their members vote for the VCM with reward.

The predominant preference for the VCM with reward is in line with the predictions from the social welfare model of Charness and Rabin (2002), whereas the Fehr and Schmidt (1999) model of inequity aversion would rather predict a more common preference for the punishment institution. Since the latter abstracts from efficiency aspects, this result is not too surprising. Both models, however, correspond with the empirical finding that subjects preferably vote for high (absolute) values of the leverage for the sanctioning instruments.

In general, our results confirm existing evidence that for a fixed group composition, people are not really willing to implement an informal option to punish. They rather prefer an environment with rewarding opportunities, which works better in terms of efficiency. Since we often have formal, centralized institutions in mind, when we think of sanctioning institutions, it seems at first sight that punishment is relatively more prevalent in the real world than in our experiment. When focusing more on informal mechanisms, for instance in work and sports teams, casual observation tells us that reward is indeed important and widespread. Perhaps, our view on the potentials of formal rewarding institutions is somewhat biased. Future research will have to show whether our results for informal institutions carry over to formal institutions.

Appendix

2A Experimental instructions (originally in German)[70]

Welcome to an experiment on decision making.

Thank you for participating!

During the experiment you and all other participants will be asked to make decisions. Your decisions as well as the decisions of the other participants determine your earnings from the experiment according to the following rules.

Please stop talking to other participants from now on. If you have any questions after going through the instructions or while the experiment is taking place, please raise your hand, and one of the experimenters will come to you and answer your questions privately.

The whole experiment is computerized and will last up to two hours. All your decisions and answers remain anonymous. We evaluate data from the experiment only on the aggregate level and never link names to data from the experiment. At the end of the experiment, you will be asked to sign a receipt for your earnings. This has accounting purposes only. The other participants will not find out how much you have earned.

The experiment consists of three parts. At the beginning of each part, you will receive the corresponding instructions for this part. The instructions will be read aloud and you will get time to ask questions. Please, do not hesitate to ask if anything is unclear to you. The decisions in different parts of the experiment are completely independent from each other.

While taking your decisions at the PC, there will be a clock counting down in the right upper corner of the screen. The clock serves as a guide for how much time you should need. You may nevertheless exceed the time. Only the information screens on which no decision is required to be taken will be turned off when time has run out.

In the interest of clarity, we will only use male terms in the instructions. These should be interpreted as being gender-neutral.

During the experiment your earnings will be calculated in “tokens”. At the end of the experiment, the “tokens” get converted into euro at the exchange rate announced in the respective part. In addition, you receive 4 euro for your arrival on time. Your total earnings from the experiment will be paid out to you privately and in cash at the end of the experiment.

For means of help, you will find a pen on your table, which you, please, leave behind on the table after the experiment.

[70] Baseline instructions describe the partner treatment. Differences in the stranger treatment are indicated by [**STRANGER**].

Part I

In Part I of the experiment all participants are randomly assigned into groups of two. Nobody will find out with whom he forms a group - not during the experiment and not after the experiment either. You take **24 decisions** in this part of the experiment. In each decision you can choose between 2 options, A and B. Each option allocates a positive or negative payoff (earning) in tokens to you and the other person in your group. The other person answers exactly the same questions. Your total payoff from Part I depends on your decisions *and* on the decisions taken by the other person in your group.

A decision example:

	Option A	Option B
Your payoff	10.00	7.00
Other's payoff	-5.00	4.00

- If you choose Option A, you receive 10 tokens, and the other person loses 5 tokens. If the other person also chooses Option A, he, too, receives 10 tokens and you lose 5 tokens. In total, you therefore earn 5 tokens (10 tokens from your choice minus 5 tokens from the other person's choice).
- In case you choose Option B and the other person chooses Option A, you earn 2 tokens (7 tokens from your choice minus 5 tokens from the other person's choice). The other person earns 14 tokens (10 tokens + 4 tokens).

Overall, you take 24 decisions like the one described above. Your total payoff is computed as follows: The 24 values for "your payoff" are summed up over your decisions. The 24 values for "other's payoff" are summed up over the other person's decisions. The sum of these two sums determines your total payoff from this part and is converted into euro as follows: **20 tokens = 3 euro** (1 token = 15 cent). This exchange rate is valid only for Part I of the experiment.

Note that you are not receiving information on each single decision taken by the other person in your group but find out only about your total payoff from this part at the end of Part I.

If you have any questions, please raise your hand now. We will come to you and answer your questions.

Part II

At the end of the experiment, the tokens that you earn in Part II will be converted into euro at the exchange rate of **20 tokens = 0.6 euro** (1 token = 3 cent).

In Part II all participants are randomly assigned into groups of four. Nobody will find out with whom he forms a group – not during the experiment and not after the experiment either. This part of the experiment consists of **10 identical periods**. Group composition remains constant over all periods [**STRANGER**: varies randomly from one period to the next]; this means you are connected to the same [**STRANGER**: different] persons in every period.

Endowment and alternatives in each period

Each participant receives an initial endowment of **20 tokens** at the beginning of each period. These 20 tokens can be allocated to two alternatives, **X** and **Y**:

1. You can put 0 to 20 tokens into **pot X**. The sum of all contributions within your group to pot X will be multiplied by 1.6 and equally distributed among the group members afterwards. This means that for any token in pot X you receive 0.4 (=1.6/4) tokens. For example, if the sum of tokens in pot X in your group is 60, each group member receives 60 • 0.4 = 24 tokens out of pot X. If all group members together contribute 10 tokens to pot X, you and all other group members receive 10 • 0.4 = 4 tokens from pot X.
2. You can put 0 to 20 tokens into **pot Y**. The tokens in pot Y enter your profit one-to-one. For example, this means that if you contribute 6 tokens to pot Y, you receive exactly 6 tokens from pot Y.

Your profit per period is the sum of the earnings from pot X and from pot Y.
Mathematically:

Result (for group member i) = (20 - x) + (S*1.6)/4

x = contribution of member i to pot X

S = sum of contributions of *all* group members to pot X

On the screen, you will be asked how many tokens you want to contribute to pot X. The rest of the tokens will automatically be allocated to pot Y. Saving tokens for a later period is therefore not possible. You can only choose integer numbers between and including 0 and 20 tokens.

After each period you receive information on the contributions to pot X and Y of all group members as well as what each group member has earned in this period. However, you are of course not able to link the information to specific persons in this room because all decisions (as mentioned above) will remain anonymous. Moreover, the participant-IDs within your group will change every period so that it is impossible for you to track the behavior of other group members over periods. After receiving feedback, the next period starts. After 10 periods, this part of the experiment ends. The profits from all periods will be added and converted into euro.

If you have any questions, please raise your hand now. We will come to you and answer your questions.

Part III

At the end of the experiment, the tokens that you earn in Part III will be converted into euro at the exchange rate of **20 tokens = 0.6 euro** (1 token = 3 cent).

Again, this part of the experiment consists of **10 identical periods** in which you interact in groups of four. Nobody will find out with whom he forms a group – not during the experiment and not after the experiment either. The group composition is the same as in Part II and remains constant during the entire Part III, too [**STRANGER**: The group composition, as in Part II, varies randomly from one period to the next]. This means that you are connected to the same [**STRANGER**: different] persons in every period.

Endowment and alternatives in each period

Each participant receives an initial endowment of **20 tokens** at the beginning of each period. However, each period now consists of three stages:

Stage 0:

For reasons of comprehensibility, the details of stage 0 will be described below.

Stage 1 (**same as Part II**):

In stage 1 you can again allocate your 20 tokens to two alternatives, **X** and **Y**:

1. You can put 0 to 20 tokens into **pot X**. The sum of all contributions within your group to pot X will be multiplied by 1.6 and equally distributed among the group members afterwards. This means that for any token in pot X you receive 0.4 (=1.6/4) tokens. For example, if the sum of tokens in pot X in your group is 60, each group member receives 60 • 0.4 = 24 tokens out of pot X. If all group members together contribute 10 tokens to pot X, you and all other group members receive 10 • 0.4 = 4 tokens from pot X.
2. You can put 0 to 20 tokens into **pot Y**. The tokens in pot Y enter your profit one-to-one. For example, this means that if you contribute 6 tokens to pot Y, you receive exactly 6 tokens from pot Y.

Your profit per period is the sum of the earnings from pot X and from pot Y.

Mathematically:

Result (for group member i) = (20 - x) + (S*1.6)/4

x = contribution of member i to pot X

S = sum of contributions of *all* group members to pot X

On the screen, you will be asked how many tokens you want to contribute to pot X. The rest of the tokens will automatically be allocated to pot Y. Saving tokens for a later period is therefore not possible. You can only choose integer numbers between and including 0 and 20 tokens.

Stage 2:

You receive information on the contributions to pot X and Y of all group members and you may be able, depending on the decisions in stage 0, to change the result of your group members. There are three alternatives:

1. You, as a group, could have decided in stage 0 that it is possible to **subtract L tokens** from the result of another group member in stage 2 at **own costs of 1**. (L < 0)
2. You, as a group, could have decided in stage 0 that it is possible to **add L tokens** to the result of another group member in stage 2 at **own costs of 1**. (L > 0)
3. You, as a group, could have decided in stage 0 that the result of stage 1 remains **unchanged**. (L = 0)

In case alternative 1 or 2 is chosen, each group member is able to change another group member's result in stage 2 by assigning a (subtraction or addition) point to this person. Assume, for example, your group agreed on L = -2. Then, in stage 2 each group member can decide for each other group member individually whether or not he wants to assign a subtraction point to this person. If he assigns a subtraction point to exactly one other group member, then his payoff will be reduced by one token and the payoff of the respective group member will be reduced by two tokens. If he assigns a subtraction point to two other group members, then his payoff will be reduced by 2*1 token and the payoff of both other group members will be reduced by 2 tokens each, etc.
At the end of stage 2 you will receive information, if applicable, on how many (subtraction or addition) points each group member assigned and received, as well as what each group member has earned in this period. Afterwards, the next period starts. However, you will not find out on an individual level who assigned a point to whom and, as in Part II, you are of course not able to link information to specific persons in this room because all decisions remain anonymous. The participant-IDs within your group will change every period so that it is again impossible for you to track the behavior of other group members over periods.
During Part III you will take decisions in 10 identical periods which correspond to the description above. Each period consists of the three stages mentioned.

Now coming to the exact description of stage 0:

As already mentioned, you can choose between three alternatives that matter for stage 2:

1. **Subtraction possibility:** possibility to **subtract L tokens** from the result of other group members in stage 2 (L < 0)
2. **Addition possibility:** possibility to **add L tokens** to the result of other group members in stage 2 (L > 0)
3. **unchanged result** (L = 0)

For the subtraction and addition possibility you also have to determine the exact size of L within your group.

How does the selection process work within your group?

1. First, each group member states his preferred L-value out of the following interval: $L \in \{-5, -4, -3, -2, -1, 0, +1, +2\}$. Note that values L < 0 correspond to the subtraction possibility (**alternative 1**), L = 0 corresponds to **alternative 3** and values L > 0 correspond to the addition possibility (**alternative 2**).

2. The four L-values proposed by the group members will then be assigned to the three alternatives (i) $L < 0$, (ii) $L = 0$, and (iii) $L > 0$.
3. The alternative with the relative majority within the group will be implemented.
4. If $L = 0$ (no subtraction or addition possibility) gets the relative majority, stage 0 ends and is immediately followed by stage 1.
5. If $L < 0$ or $L > 0$ gets the relative majority, there will be a second voting round in which all four group members vote on the exact size of L. In doing so, group members are not tied to their previous proposals.

   If $L < 0$ gets the relative majority, the now available values are: $L \in \{-5, -4, -3, -2, -1\}$.

   If $L > 0$ gets the relative majority, the now available values are: $L \in \{+1, +2\}$.

   The mean of the second proposals of all four group members will finally be implemented for the whole group.
6. If there is a tie in the first voting round (this means that there is no relative majority for one alternative), one of the four first-round proposals of the group members will randomly be determined and directly implemented. Stage 1 will then start without having any further voting round.

Please note that the chosen L-value is valid only for the respective period. In the following period, the same election process starts again and the group can agree on a different value of L.

Description of profits

In the following we summarize period profits depending on the alternative chosen in stage 0.

(a) Unchanged result (L = 0):

Result (for group member i) = $(20 - x) + (S*1.6)/4$

x = contribution of member i to pot X

S = sum of contributions of *all* group members to pot X

(b) Subtraction possibility (L < 0):

Result (for group member i) =

$$(20 - x) + (S*1.6)/4 + \underbrace{L*(\text{sum of received subtraction points})}_{\text{this expression is negative}} - (\text{sum of assigned subtraction points})$$

(c) Addition possibility (L > 0):

Result (for group member i) =

$$(20 - x) + (S*1.6)/4 + \underbrace{L*(\text{sum of received addition points})}_{\text{this expression is positive}} - (\text{sum of assigned addition points})$$

The End

After the 10 periods, the whole experiment ends. Profits from all periods of Part III will be added and converted into euro. After filling out a short post-experimental questionnaire, you will receive your total earnings from the experiment privately and in cash.

If you have any questions, please raise your hand now. We will come to you and answer your questions.

2B Fehr and Schmidt (1999) and Charness and Rabin (2002) preferences[71]

2B.1 Theoretical predictions with Fehr and Schmidt (1999) preferences

2B.1.1 Standard VCM ($L = 0$)

For $L = 0$, we can apply Proposition 4 of Fehr and Schmidt (1999, see p. 839). Following 4(a), each group member with $\gamma + \beta_i < 1$, i.e. $\beta_i < 0.6$, contributes $c_i = 0$ irrespective of the choices of her group members. Moreover, in line with 4(b) there is no equilibrium with positive contributions in the standard VCM if the number of group members with $\beta_i < 0.6$ is larger than $(n-1) \cdot \gamma/2 = 0.6$. Hence, one person with $\beta_i < 0.6$ is already enough to completely destroy any cooperation within the group. Equilibria with positive contributions are only possible if *all* group members satisfy $\beta_i \geq 0.6$, i.e. we have four so-called "conditional cooperators" that are sufficiently averse to advantageous inequity. Note that this event is rather unlikely and occurs, following the parameter distribution given in Fehr and Schmidt (1999, p. 844), only in $0.4^4 = 2.56\%$ of cases. If it occurs, then there are according to 4(c) multiple equilibria and each group member contributes $c_i = c \in [0, E]$.

Proposition 2B.1. *In the standard VCM ($L = 0$), complete free-riding ($c_i = 0 \; \forall i$) is the equilibrium outcome if at least one group member satisfies $\beta_i < 0.6$. Only if all group members fulfill $\beta_i \geq 0.6$, i.e. if they are sufficiently averse to advantageous inequity, there exist equilibria with positive contributions $c_i = c \in [0, E]$.*

2B.1.2 VCM with Punishment ($L < 0$)

For $L < 0$, we apply Proposition 5 of Fehr and Schmidt (1999, p. 841) to account for the punishment possibilities. Assume that there exists a group of $n' \leq n$ "conditional cooperators" who satisfy $\gamma + \beta_i \geq 1$, i.e. $\beta_i \geq 0.6$, whereas all other group members do not care at all about inequity, i.e. for them $\alpha_i = \beta_i = 0$. Further, denote k as the costs that arise when a subject punishes another group member. Note that we have $k = 1$ in the experiment. Consider then the following strategies derived from Proposition 5: All group members contribute $c_i = c \in [0, E]$. If each group member does so, no punishment occurs. If one of the selfish subjects deviates by contributing $c_i < c$, all conditional cooperators punish the deviator while the remaining subjects do not punish. If one of the conditional cooperators chooses $c_i < c$ or if some subject contributes $c_i > c$ or if more than one subject deviates from

[71] The presentation in Appendix 2B follows the procedure described in the appendix of Sutter et al. (2010). We align and extend their approach appropriately to deal with our setup.

c, then one Nash equilibrium of the punishment game is played. These strategies form a subgame perfect equilibrium with contribution level c and zero punishment if the following three conditions are satisfied: (*i*) no conditional cooperator benefits from contributing less than c, (*ii*) no selfish subject benefits from contributing less than c given the punishment of the n' conditional cooperators and (*iii*) each conditional cooperator has an incentive to punish selfish subjects who contribute $c_i < c$ thereby generating a credible punishment threat.[72]

Condition (*i*) is satisfied by construction as all subjects with $\beta_i \geq 0.6$ have an intrinsic motivation to contribute $c_i = c$ if all other group members contribute c (that is why we call them *conditionally* cooperative). Hence, conditional cooperators will never deviate from a symmetric equilibrium. Two things follow immediately: if $n' = 4$ there exists a multiplicity of equilibria in which each group member contributes $c_i = c \in [0, E]$ and if $n' = 0$, i.e. all subjects are selfish, there is no equilibrium with positive contributions. These findings are equivalent to the case of $L = 0$.

Regarding condition (*ii*), notice the following: If a selfish subject deviates by contributing $c_i < c$, she obtains a monetary gain of $(c - c_i)(1 - \gamma)$ which she maximizes by choosing $c_i = 0$. However, if she gets punished by the n' conditional cooperators she additionally suffers a monetary loss of $n'|L|$. Hence, deviating from contributing c is never profitable as long as

$$c \leq \frac{n'|L|}{(1-\gamma)} \equiv \bar{c}. \tag{2B.1}$$

The parameter $\bar{c}$ denotes the maximum contribution level that can be sustained in equilibrium given the marginal per capita return (MPCR) γ and n' conditional cooperators who punish through a leverage of $|L|$. Condition (2B.2) shows that $|L|$ increases in $\bar{c}$. Hence, the requirements on the strength of the enforcement mechanism rise, holding the MPCR and n' constant, in the contribution level that one wants to sustain in equilibrium.

$$\frac{\partial |L|}{\partial \bar{c}} = \frac{(1-\gamma)}{n'} > 0 \tag{2B.2}$$

It is noteworthy, that for our parameter constellation full cooperation (i.e. $c_i = 20\ \forall i$) is only possible for $n' = 3$ and $|L| \geq 4$ (and, of course, for $n' = 4$).

[72] Note that deviating by contributing $c_i > c$ is never profitable as it both reduces the monetary payoff and increases inequity. Furthermore, there is no reason to punish if all group members contribute c due to the lack of inequity.

Finally, condition (*iii*) requires that the punishment threat is credible. Hence, we compare a conditional cooperator i's utility change in the punishment stage if she punishes the deviator with her utility change if she does not punish maintaining the assumption that the $n' - 1$ other conditional cooperators punish the deviator. The former utility change is not worse than the latter if the following inequality holds:[73]

$$-k - \frac{\alpha_i}{n-1}(n - n' - 1)k - \frac{\alpha_i}{n-1}(k - n'|L|) \geq -\frac{\alpha_i}{n-1}(-(n'-1)|L|) - \frac{\beta_i}{n-1}(n'-1)k \quad \text{(2B.3)}$$

The left-hand side of condition (2B.3) describes the case that person i punishes the deviator and consists of three terms. The first term captures the monetary costs of punishing. The second term contains the disadvantageous inequity towards the $n - n' - 1$ selfish subjects who contribute but do not punish and the third term comprises the disadvantageous inequity towards the deviator who receives punishment by all n' conditional cooperators. On the right-hand side we have the case that person i does not punish. Then, there is disadvantageous inequity towards the deviator based on the punishment of the $n' - 1$ other conditional cooperators and advantageous inequity towards the $n' - 1$ punishing conditional cooperators. If we rearrange and simplify condition (2B.3) we arrive at

$$\frac{|L|}{k} \geq (n - n') + \frac{1}{\alpha_i}[(n-1) - \beta_i(n'-1)]. \quad \text{(2B.4)}$$

Note that this inequality condition can be fulfilled even if there is only one conditional cooperator. Hence, punishment can be a credible threat. Inserting $k = 1$ and $n = 4$ the critical value of L is given by

$$|L| = (4 - n') + \frac{1}{\alpha_i}[3 - \beta_i(n'-1)]. \quad \text{(2B.5)}$$

Thus, we can compute how the critical value of $|L|$ changes with regard to the inequity aversion parameters α_i and β_i:

$$\frac{\partial|L|}{\partial\alpha_i} = -\frac{1}{\alpha_i^2}[3 - \beta_i(n'-1)] < 0 \quad \text{for } n' \geq 1$$

$$\frac{\partial|L|}{\partial\beta_i} = -\frac{n'-1}{\alpha_i} \begin{cases} = 0 & \text{for } n' = 1 \\ < 0 & \text{for } n' > 1 \end{cases} \quad \text{(2B.6a,b)}$$

[73] Note that we assume in condition (2B.3) that the free-rider's payoff after punishment is not below those of the conditional cooperators. This is satisfied for $c \geq n'L - k$ which is in our experiment for sure the case if $c \geq 14$.

Both derivatives are negative (except for $n' = 1$ in (2B.6b) when β_i does not play any role for the determination of $|L|$). This means that the requirement on $|L|$ which is necessary to make the punishment threat credible (holding n' constant) decreases in both inequity aversion parameters. Or, put it differently, the less inequity averse the conditional cooperators are, the stronger the punishment mechanism must be to induce punishment of deviators. Hence, a higher value of $|L|$ is more likely to make punishment of free-riders credible. To give some numerical examples, first consider $n' = 3$. If we assume the lowest bounds on α_i and β_i each conditional cooperator has to satisfy per definition, i.e. $\alpha_i = \beta_i = 0.6$, the critical value is $|L| = 4$. On the contrary, if we for example assume that each conditional cooperator satisfies $\beta_i = 1$ and $\alpha_i = 4$, all $|L| \geq 1.25$ make punishment credible. Similarly, for $n' = 2$ ($n' = 1$), the corresponding values for $|L|$ are 6 (8) and 2.5 (3.75).[74]

To sum up, Fehr and Schmidt preferences provide two reasons to vote for a strong punishment instrument in voting round two. First, the higher $|L|$ the less restrictive are the requirements on the conditional cooperators' inequity aversion parameters α_i and β_i and, hence, the more likely is punishment a credible threat and cooperation possible in a subgame perfect equilibrium. Second, the higher $|L|$ the higher contribution levels can be sustained in equilibrium for a given number of punishing conditional cooperators. Note that efficiency strictly increases in equilibria with higher contribution levels which makes the implementation of a strong punishment instrument profitable. As the mean value of proposed L parameters is implemented within the group, each group member affects the result and has therefore a preference for $L = -5$.

Proposition 2B.2. *In the VCM with punishment ($L < 0$), a group of $n' \leq n$ conditional cooperators with $\beta_i \geq 0.6$ can enforce positive contributions $c \leq n'|L|/(1-\gamma) \equiv \bar{c}\ \forall i$ if each conditional cooperator cares sufficiently about disadvantageous inequity. Equilibria with positive contributions (and zero punishment) can more easily be sustained when $|L|$ increases and can support higher contribution levels. Hence, subjects in voting round two have an incentive to vote for the maximum punishment level $L = -5$. If all group members satisfy $\beta_i \geq 0.6$, equilibria with $c_i = c \in [0, E]$ can be sustained irrespective of L.*

2B.1.3 VCM with Reward ($L > 0$)

Analogous to the case of punishment, we assume that there exists a group of $n' \leq n$ "conditional cooperators" who satisfy $\beta_i \geq 0.6$, whereas all other group members have

[74] Note that in our experiment the punishment leverage is restricted by $|L| \leq 5$.

$\alpha_i = \beta_i = 0$. Moreover, the costs of rewarding another subject are defined by $k = 1$ in the experiment. Consider the following strategies: All group members contribute $c_i = c \in [0, E]$. If each group member does so, each of the n' conditional cooperators rewards all her group members while the other subjects do not reward. If one group member deviates by contributing $c_i < c$, no group member rewards the deviator. If some subject contributes $c_i \geq c$ or if more than one subject deviates from c, then one Nash equilibrium of the reward game is played.

Let us check whether the reward incentive is credible given that all group members contribute c. Note that selfish subjects never reward as they do not care at all about inequity. For a conditional cooperator i we have to determine whether the utility change of rewarding is beneficial or not. We assume in condition (2B.7) that she either rewards all her group members (left-hand side) or none of them (right-hand side).[75]

$$-(n-1)k - \frac{\alpha_i}{n-1}(n-n')[L+(n-1)k] \geq -\frac{\beta_i}{n-1}(n'-1)[L+(n-1)k] \qquad (2B.7)$$

The first term on the left-hand side captures the monetary costs of rewarding each of the other $n-1$ group members while the second term describes the disadvantageous inequity towards the selfish subjects who do not reward. The term on the right-hand side denotes the advantageous inequity towards the $n'-1$ conditional cooperators who reward. If we rearrange condition (2B.7) this leads to the following expression:

$$(n-1)k + \frac{1}{n-1}[L+(n-1)k][\alpha_i(n-n') - \beta_i(n'-1)] \leq 0 \qquad (2B.8)$$

Inserting $k = 1$ and $n = 4$ and further rearranging yields

$$\alpha_i(4-n') - \beta_i(n'-1) \leq -\frac{9}{L+3}. \qquad (2B.9)$$

Note that in our reward institution $1 \leq L \leq 2$ holds and, hence, the right-hand side cannot be larger than -9/5. To fulfill condition (2B.9), i.e. to make reward credible, the left-hand side has to be smaller than this value. However, this is never satisfied for $n' < n$. To see this, consider the most critical case $n' = 3$, in which condition (2B.9) demands at least $\alpha_i - 2\beta_i \leq -9/5$. Inserting the lowest possible value of α_i, $\alpha_i = \beta_i$, would imply $\beta_i \geq 9/5$ which is impossible. Thus, irrespective of L there is no equilibrium in which reward is used only by a

[75] This is done because there is no possibility for subject i to distinguish between selfish and non-selfish subjects.

subgroup of individuals. For $n' = n$, however, mutual reward can be part of the equilibrium. In this case condition (2B.9) shortens to $\beta_i \geq 3/(L+3)$ or $L \geq 3/\beta_i - 3$. Hence, the critical value of L decreases in β_i (see condition (2B.10)) and the minimum requirements on β_i lie between $\beta_i \geq 0.75$ for $L = 1$ and $\beta_i \geq 0.6$ for $L = 2$.

$$\frac{\partial L}{\partial \beta_i} = -\frac{3}{\beta_i^2} < 0 \tag{2B.10}$$

To sum up, if all group members satisfy $\beta_i \geq 0.6$ each contribution level $c_i = c \in [0, E]$ can be sustained in a subgame perfect equilibrium like in the absence of reward ($L = 0$). Moreover, mutual reward can be part of such an equilibrium if $L = 2$. For lower levels of L, mutual reward is only possible if we have correspondingly stronger assumptions on β_i. Thus, subjects have an incentive to vote for $L = 2$ in the second voting round to increase prospects for mutual reward and to make reward more beneficial. As the mean value of proposed L parameters is implemented, this incentive holds for each group member.

Nevertheless, such equilibria are rather unlikely as $\beta_i \geq 0.6\ \forall i$ appears only in about 2.56% of cases (see Section 2B.1.1). If one group member satisfies $\beta_i < 0.6$, this group member has no incentive to reward and, hence, there is no reward at all within the group. It follows from Section 2B.1.1 that we observe complete free-riding in equilibrium, irrespective of L, when there is at least one group member with $\beta_i < 0.6$. Hence, subjects are indifferent between both positive leverage levels in this case.[76]

Proposition 2B.3. *In the VCM with reward ($L > 0$), complete free-riding ($c_i = 0\ \forall i$) and zero reward arise in equilibrium if at least one group member satisfies $\beta_i < 0.6$. In this case, subjects are completely indifferent between leverage levels. Only if all group members fulfill $\beta_i \geq 0.6$, i.e. if they are sufficiently averse to advantageous inequity, there exist equilibria with positive contributions $c_i = c \in [0, E]$. Moreover, in equilibrium subjects are willing to reward each other if all of them satisfy $\beta_i \geq 3/(L+3)$. As mutual reward can more easily be sustained and is more beneficial when L increases, subjects have an incentive to vote for $L = 2$ in the second voting round.*

[76] If we apply trembling-hand perfection, inequity-averse subjects might have an incentive to vote for $L = 1$ as this leverage level minimizes the monetary consequences of an accidental use of the reward instrument.

2B.1.4 Institutional voting (First voting round)

We have seen that for the case of $\beta_i \geq 0.6\ \forall i$ equilibria with positive contributions $c_i = c \in [0, E]$ can be sustained irrespective of the chosen leverage level. Moreover, if $L = 2$ mutual reward can for sure be part of the equilibrium strategy. Hence, subjects have an incentive to vote for $L = 2$ if $\beta_i \geq 0.6\ \forall i$ holds in order to implement the most profitable equilibrium. Note that voting for $L = 2$ weakly dominates the alternative reward option $L = 1$ in the first voting round if we assume that ties occur with some positive probability (e.g. because subjects make mistakes). Otherwise, subjects would be indifferent between $L = 1$ and $L = 2$ as the exact level of reward is only determined in the second voting round.

In contrast, if there is at least one group member satisfying $\beta_i < 0.6$, positive contributions cannot be sustained in equilibrium in both the standard VCM and the reward institution. Taking into account the distribution of β presented in Fehr and Schmidt (1999), this is the large majority of cases (97.44%). The situation differs in the punishment institution. Here, a single conditional cooperator can already be enough to enforce positive contributions if she is sufficiently averse to disadvantageous inequity. The likelihood of this event and the enforceable contribution level increase in the implemented leverage level and, thus, subjects should choose $L = -5$ in the second voting round if punishment was selected. Choosing $L = -5$ also weakly dominates the other punishment levels in voting round one when subjects believe that ties happen with some positive probability. For the case of $L = -5$, one conditional cooperator with $\beta_i \geq 0.6$ and $\alpha_i \geq 1.5$ is sufficient to sustain equilibria with positive contributions (see condition (2B.5)). Following Fehr and Schmidt (1999), 40% of subjects satisfy $\beta_i \geq 0.6$ and $\alpha_i \geq 1$. If we assume that 30% also fulfill $\alpha_i \geq 1.5$[77], the probability of having at least one such inequity averse conditional cooperator in a four-person group is $1 - 0.7^4 = 75.99\%$. Hence, cooperation chances are greatly improved under the punishment institution compared to the standard VCM or the reward institution and it is therefore optimal (in most of the cases) to vote for $L = -5$.

Proposition 2B.4. *Given that subjects can vote for the standard VCM, the VCM with punishment or the VCM with reward by choosing a leverage level out of* $[-5, 2]$*, they prefer the punishment institution and choose* $L = -5$ *as this maximizes cooperation possibilities and payoffs.*

[77] This percentage cannot be directly inferred from Fehr and Schmidt (1999) as they only state that 40% of subjects fulfill $\alpha_i \geq 1$ and 10% even satisfy $\alpha_i \geq 4$. By employing 30% we therefore carefully assume a decreasing probability mass in the interval [1, 4].

2B.2 Theoretical predictions with Charness and Rabin (2002) preferences

2B.2.1 Standard VCM ($L = 0$)

For $L = 0$, a selfish subject with $\lambda_i = 0$ has obviously no incentive to contribute to the public account. A subject that cares about social welfare (i.e. $\lambda_i > 0$) has to consider that contributing one unit to the public account reduces her monetary payoff by $1 - \gamma$, increases the sum of group members' payoffs by $n\gamma - 1$, and decreases the minimum payoff in the group by $1 - \gamma$.[78] Weighting these aspects according to the utility function yields the following inequality:

$$-(1 - \lambda_i)(1 - \gamma) + \lambda_i(1 - \delta_i)(n\gamma - 1) - \lambda_i\delta_i(1 - \gamma) \geq 0 \qquad \text{(2B.11)}$$

If we rearrange condition (2B.11) and insert $\gamma = 0.4$ and $n = 4$ we arrive at

$$\delta_i \leq 1 - \frac{1}{2\lambda_i}. \qquad \text{(2B.12)}$$

Condition (2B.12) shows that subject i contributes to the public account if $\lambda_i \geq 0.5$ (necessary to get non-negative values for δ_i) and if δ_i is sufficiently low, depending on the exact value of λ_i (for sure ≤ 0.5).[79] In this case, i.e. if she cares enough about group efficiency, she contributes her entire endowment E. On the contrary, subjects with $\lambda_i < 0.5$ do not contribute. Note that condition (2B.12) contains only the individual's own social welfare parameters. Hence, if (2B.12) is satisfied she contributes independent of the number of free-riders within the group. Such subjects are called "cooperators".[80]

Proposition 2B.5. *In the standard VCM ($L = 0$), each group member with $\lambda_i < 0.5$ contributes $c_i = 0$. On the contrary, group members who fulfill $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$, i.e. who are sufficiently total-surplus oriented, contribute their entire endowment: $c_i = E$.*

[78] Note that the decrease in the minimum payoff is strictly true only for the case in which no other subject contributes. If there are already positive contributions, a subject could on the contrary increase the minimum payoff by contributing positive amounts. In this case condition (2B.12) changes into $\delta_i \leq 6 - 3/\lambda_i$ which also yields $\lambda_i \geq 0.5$ but contains a less restrictive requirement for δ_i. For the ease of analysis we neglect this aspect in the following. Note that condition (2B.12) generates contribution incentives *irrespective* of group members' decisions.

[79] Strictly speaking, subject i is indifferent if (2B.12) holds with equality. We mostly ignore such indifferences in the following for simplicity.

[80] Note that these subjects are *un*conditionally cooperative in contrast to the Fehr and Schmidt (1999) model.

2B.2.2 VCM with Punishment ($L < 0$)

Assume that there exists a group of $n'' \leq n$ "cooperators" who satisfy $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$ whereas all other group members do not care at all about social welfare (i.e. $\lambda_i = 0$). Note that cooperators can only motivate selfish subjects to contribute $c_i > 0$ if the latter's monetary gain from contributing is higher than from free-riding. This is fulfilled if condition (2B.1) from above holds, i.e. $c \leq n''|L|/(1-\gamma) \equiv \bar{c}$. However, the punishment threat is not credible with Charness and Rabin (2002) preferences. The act of punishing would reduce (i) a cooperator's own payoff, (ii) the sum of group members' payoffs and perhaps even (iii) the minimum payoff in the group. Hence, punishment of free-riders does never occur and no subject can be motivated to contribute through the threat of punishment. Subjects are therefore indifferent between leverage levels and equilibrium outcomes equal those of $L = 0$.[81]

Proposition 2B.6. *In the VCM with punishment ($L < 0$), punishment is not credible and subjects are indifferent between leverage levels. Like for $L = 0$, each group member with $\lambda_i < 0.5$ contributes $c_i = 0$ while group members who fulfill $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$ contribute their entire endowment: $c_i = E$.*

2B.2.3 VCM with Reward ($L > 0$)

Again, define a group of $n'' \leq n$ cooperators who satisfy $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$ whereas all other group members do not care at all about social welfare (i.e. $\lambda_i = 0$). Consider the following strategies: Each cooperator contributes $c_i = E$ while all other group members contribute $c_i = c \in [0, E]$. If each group member does so, each of the n'' cooperators rewards all her group members while selfish subjects do not reward. If one selfish subject deviates by contributing $c_i < c$, no group member rewards the deviator. If one of the selfish subjects chooses $c_i > c$ or if one of the cooperators contributes $c_i < E$ or if more than one subject deviates, then one Nash equilibrium of the reward game is played.

We have to check our three conditions: First, we need cooperators that have an intrinsic motivation to contribute E, i.e. satisfy $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$. This is fulfilled by construction. It follows immediately that if $n'' = 0$ there is no equilibrium with positive contributions (and reward) and if $n'' = 4$ there exists an equilibrium in which each group member contributes $c_i = E$.

[81] The concept of trembling-hand perfection suggests that subjects vote for $L = -1$ as this minimizes the impact of an accidental use of the punishment instrument.

Second, a selfish subject does not benefit from contributing $c_i < c$. Deviating leads to a monetary gain of $(c - c_i)(1 - \gamma)$ but generates a monetary loss of $n''L$ because it destroys reward of the n'' cooperators. This results in the analogous condition as for the Fehr and Schmidt preferences (see condition (2B.1) from above):

$$c \leq \frac{n''L}{(1-\gamma)} \equiv \bar{c} \tag{2B.13}$$

If condition (2B.13) is satisfied, a selfish subject cooperates. Note that the leverage level L increases in the maximum contribution level $\bar{c}$ as shown in condition (2B.2). Hence, higher contribution levels of selfish subjects can only be enforced, holding the MPCR and n'' constant, if the reward mechanism becomes stronger. It is, however, noteworthy that for our reward institution with $L \leq 2$, full cooperation (i.e. $c_i = 20\ \forall i$) is never possible if there is at least one selfish group member. From this member only contributions up to a level of $c = 10$ can be enforced (the latter for the case of three cooperators and $L = 2$).

Third, it must be beneficial for a cooperator i to reward selfish subjects. Hence, we compare the utility change in the reward stage when cooperator i rewards all her group members with the utility change if she only rewards the $n'' - 1$ other cooperators given that they stick to their reward strategy.[82] This yields the following inequality:

$$\begin{aligned}
&(1-\lambda_i)[(n''-1)L-(n-1)k]+\lambda_i\big[\delta_i\big((n''-1)L-(n-1)k\big)\\
&+(1-\delta_i)\{n''[(n''-1)L-(n-1)k]+(n-n'')(n''L)\}]\\
&\geq (1-\lambda_i)[(n''-1)L-(n''-1)k]+\lambda_i[\delta_i\big((n''-1)L-(n''-1)(6-1.5n'')k\big)\\
&\quad+(1-\delta_i)\{(n''-1)[(n''-1)L-(n-1)k]\\
&\qquad+(n-n'')(n''-1)L+1[(n''-1)L-(n''-1)k]\}]
\end{aligned} \tag{2B.14}$$

Rearranging and simplifying condition (2B.14) leads to

$$\frac{L}{k} \geq \frac{(1-\lambda_i)(n-n'')+\lambda_i\delta_i\big((n-1)-(6-1.5n'')(n''-1)\big)+\lambda_i(1-\delta_i)(n-n'')}{\lambda_i(1-\delta_i)(n-n'')}. \tag{2B.15}$$

Note that the right-hand side of condition (2B.15) is larger or equal to 1 as the last term of the numerator equals the denominator and both prior terms are non-negative. From this it follows that for $L = k$ the reward strategy can only be credible through indifference if $\lambda_i = 1$ (first

[82] Note that "reward only cooperators" instead of "reward nobody" is the relevant alternative strategy as it can reduce monetary costs without decreasing the minimum payoff in the group. Further, due to the linearity of preferences it can never be profitable to discriminate within the group of cooperators and/or the group of selfish subjects. For $n'' = 4$, there is no selfish subject and the following calculations do not hold. We will cover this special case below.

term vanishes) and in case of $n'' = 1$ additionally $\delta_i = 0$ (second term vanishes). Hence, if there is no efficiency gain in rewarding, reward can only be part of the equilibrium if cooperators do not care at all about their costs ($\lambda_i = 1$) and if reward either does not influence the minimum payoff ($n'' > 1$) or subjects do not care about the "Rawlsian" criterion ($\delta_i = 0$). Further, note that by inserting $n = 4$ condition (2B.15) reduces to

$$\frac{L}{k} \geq \frac{1}{\lambda_i(1-\delta_i)} \quad \text{for } n'' = 1 \qquad \text{and} \qquad \frac{L}{k} \geq \frac{1-\lambda_i\delta_i}{\lambda_i(1-\delta_i)} \quad \text{for } n'' > 1. \tag{2B.16a,b}$$

Inserting $k = 1$ and $n = 4$ in condition (2B.15) yields the following critical value of L:

$$L = \frac{(1-\lambda_i)(4-n'') + \lambda_i\delta_i\big(3-(6-1.5n'')(n''-1)\big) + \lambda_i(1-\delta_i)(4-n'')}{\lambda_i(1-\delta_i)(4-n'')} \tag{2B.17}$$

The critical value of L changes with regard to the social welfare parameters λ_i and δ_i as follows:

$$\frac{\partial L}{\partial \lambda_i} = -\frac{1}{\lambda_i^2(1-\delta_i)} < 0$$

$$\frac{\partial L}{\partial \delta_i} = \frac{1}{(1-\delta_i)^2}\left(\frac{1}{\lambda_i} + \frac{5-6.5n''+1.5n''^2}{4-n''}\right) \begin{cases} > 0 & \text{for } n'' = 1 \\ \geq 0 & \text{for } n'' > 1 \end{cases} \tag{2B.18a,b}$$

Hence, the requirement on L which is necessary to make reward credible decreases in the social welfare parameter λ_i and increases in the maximin parameter δ_i. Regarding λ_i, this means that the more social welfare oriented subjects are the less efficient can the reward mechanism be to sustain reward in equilibrium. Or, reversing the interpretation, a higher value of L puts less restriction on subjects' social welfare parameters and is therefore more likely to make reward credible. Regarding δ_i, note that a higher weight on the maximin aspect of social welfare vice versa decreases the importance of the efficiency concern. Hence, we need a higher value of L to compensate for the reduced weight on efficiency as efficiency is the only reason for cooperators to reward selfish subjects. The weight δ_i is irrelevant (see the zero derivative in condition (2B.18b)) for the special case of $n'' > 1$ and $\lambda_i = 1$ because in this situation there is no trade-off between the different aspects of the utility function as subjects do not care about their own monetary costs and the minimum payoff cannot be affected by the reward choice. To give some numerical examples, first consider $L = 1$. As already mentioned above, in this case reward can only be part of an equilibrium strategy in the extreme case of $\lambda_i = 1$. Moreover, for $n'' = 1$ we additionally need $\delta_i = 0$. For $L = 2$ (and $k = 1$), conditions (2B.16a) and (2B.16b) imply $\delta_i \leq (2\lambda_i - 1)/2\lambda_i$ for $n'' = 1$ and

$\delta_i \leq (2\lambda_i - 1)/\lambda_i$ for $n'' > 1$. Note that the first condition exactly equals the cooperator's constraint from above (see condition (2B.12)) while the second condition is less demanding than the first one. Hence, each cooperator can credibly reward selfish subjects for $L = 2$. In comparison to the case of $L = 1$ reward possibilities are greatly improved. Therefore, subjects have an incentive to vote for $L = 2$ in the second voting round to improve conditions for the reward of selfish subjects (and to raise the enforceable contribution level). Note that this incentive holds for all group members as the mean value of the proposed parameters is implemented.[83]

Finally, let us briefly consider $n'' = 4$. In this case there is no selfish subject and positive contributions need not to be enforced as each subject has an intrinsic motivation to contribute $c_i = E$. However, mutual reward is not necessarily part of the equilibrium because a cooperator in the reward stage can have an incentive to reward none of her group members. Therefore, we have to check the following inequality:[84]

$$\begin{aligned}
&(1-\lambda_i)[(n''-1)L-(n-1)k]+\lambda_i\big[\delta_i\big((n''-1)L-(n-1)k\big)\\
&+(1-\delta_i)\{n''[(n''-1)L-(n-1)k]+(n-n'')(n''L)\}]\\
&\geq (1-\lambda_i)(n''-1)L+\lambda_i[\delta_i\big((n''-1)L-L-(n-1)k\big)\\
&\quad+(1-\delta_i)\{(n''-1)[(n''-1)L-L-(n-1)k]\\
&\qquad+(n-n''+1)(n''-1)L\}]
\end{aligned} \tag{2B.19}$$

Rearranging and simplifying leads to condition (2B.20) which holds irrespective of n'':

$$\frac{L}{k} \geq \frac{(n-1)(1-\lambda_i\delta_i)}{\lambda_i(2\delta_i - 1 + (1-\delta_i)n)} \tag{2B.20}$$

Inserting $k = 1$ and $n = 4$ yields the following critical value of L:

$$L = \frac{3 - 3\lambda_i\delta_i}{\lambda_i(3 - 2\delta_i)} \tag{2B.21}$$

The critical value of L changes with regard to the social welfare parameters λ_i and δ_i as follows:

[83] Selfish subjects are indifferent if the maximum contribution level $\bar{c}$ is enforced. However, if they believe that there is a possibility to participate in the cooperators' gains from choosing $L = 2$ or if we use a reasonable equilibrium refinement argument like pareto optimality this leads to the conclusion that those subjects always vote for $L = 2$.

[84] Note that this inequality holds for all $1 < n'' \leq 4$ but that it does not capture the most relevant deviation for $n'' < 4$. For $n'' = 1$ the correct formula is given by condition (2B.14).

$$\frac{\partial L}{\partial \lambda_i} = \frac{-3}{\lambda_i^2(3-2\delta_i)} < 0$$

$$\frac{\partial L}{\partial \delta_i} = \frac{6-9\lambda_i}{\lambda_i(3-2\delta_i)^2} \quad \begin{cases} > 0 & \text{for } \lambda_i < 2/3 \\ = 0 & \text{for } \lambda_i = 2/3 \\ < 0 & \text{for } \lambda_i > 2/3 \end{cases} \qquad \text{(2B.22a,b)}$$

Hence, the requirement on L which is necessary to make reward credible in the case of $n'' = 4$ decreases in the social welfare parameter λ_i while the relative influence of the maximin aspect of social welfare δ_i can be positive or negative (depending on λ_i). To give some numerical examples, first consider $L = 1$. In this case condition (2B.20) requires $3\lambda_i \geq 3 - \lambda_i\delta_i$. If we combine this requirement with the cooperator's constraints $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$, we find that for credible reward λ_i has to lie in the interval $[0.875, 1]$ with δ_i appropriately. Note that compared to $n'' < 4$ it is more likely that a cooperator rewards all her group members because reward increases the minimum payoff in the group. For $L = 2$, again, each cooperator with $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$ (the cooperator's constraints) can credibly reward as condition (2B.20) shortens to $\delta_i \leq 6 - 3/\lambda_i$ whose right-hand side is greater than $1 - 1/2\lambda_i$ for all $\lambda_i \geq 0.5$. To sum up, in the case of $n'' = 4$ subjects have an incentive to vote for $L = 2$ in the second voting round to improve conditions for mutual reward (and its impact). Again, this incentive holds for all group members as the mean value of the proposed parameters is implemented.

Proposition 2B.7. *In the VCM with reward ($L > 0$), a group of $n'' \leq n$ cooperators with $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$ can enforce positive contributions $c \leq n''L/(1-\gamma) \equiv \bar{c}$ from the selfish subjects if each cooperator cares sufficiently about efficiency to reward those group members. The enforcement becomes easier and can entail higher contribution levels the higher L is. Hence, subjects have an incentive to vote for $L = 2$ in the second voting round. The latter holds also if there is no selfish subject as a higher leverage level improves the conditions for mutual reward and its impact.*

2B.2.4 Institutional voting (First voting round)

As we have seen, each group member with $\lambda_i \geq 0.5$ and $\delta_i \leq 1 - 1/2\lambda_i$ has an intrinsic motivation to contribute $c_i = E$ irrespective of the chosen institution. On the contrary, selfish subjects never contribute positive amounts both in the standard VCM ($L = 0$) and in the punishment institution ($L < 0$). Only in the reward institution they can be motivated to contribute positive amounts if there are cooperators who care sufficiently about efficiency.

Hence, subjects have an incentive to vote for the reward institution in the first voting round to implement an equilibrium with higher contribution levels and reward.[85] As the first voting round only determines the institution, subjects are indifferent between $L = 1$ and $L = 2$. However, assuming a small positive probability of observing ties leads to the result that choosing $L = 2$ weakly dominates $L = 1$.

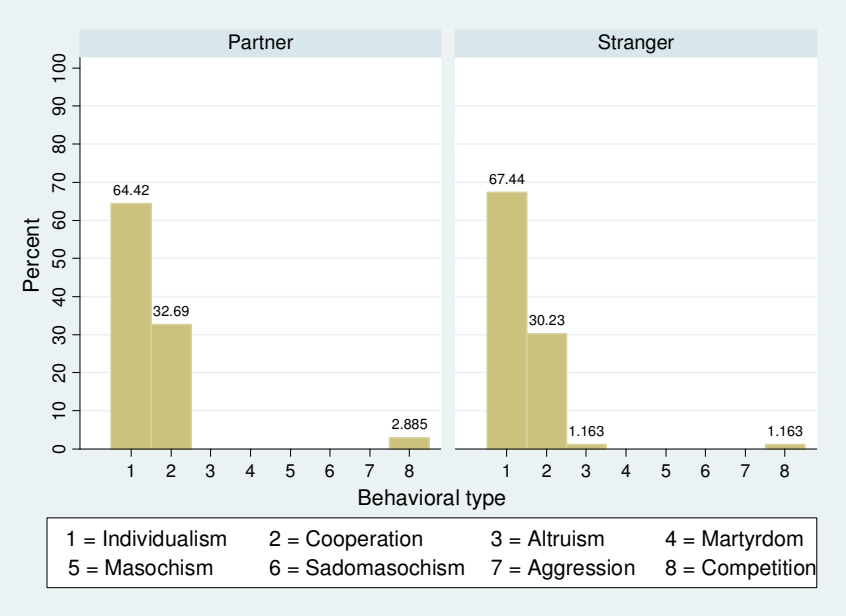
Proposition 2B.8. *Given that subjects can vote for the standard VCM, the VCM with punishment or the VCM with reward by choosing a leverage level out of* $[-5, 2]$*, they prefer the reward institution and choose* $L = 2$ *as this maximizes cooperation possibilities and payoffs.*

[85] Note that selfish subjects are indifferent if the reward institution is implemented with the maximum contribution level $\bar{c}$. However, as stated in footnote 83, there are reasonable arguments why selfish subjects vote for the VCM with reward even in this case.

2C Further results

Results from the social value orientation questionnaire (ring test)

Figure 2C.1: Distribution of behavioral types by treatment (consistency index ≥ 20 required)



Individual's voting behavior in the first voting rounds

Figures 2C.2 and 2C.3 report the detailed voting behavior of each of our participants in the partner and stranger treatment, respectively. Subjects with a number between 101 and 124 took part in the first session, subjects with 201-224 in session two, etc. Moreover, subjects are sorted that way that persons 101-104 (105-108, etc.) formed a group in the partner treatment, while each block of 12 subjects formed a matching group in the stranger treatment (e.g. 601-612, 613-624). Letters show which institutions subjects preferred over the course of the experiment. *S* stands for the standard VCM, *P* for VCM with punishment and *R* for VCM with reward. For example, a subject classified by *SR* never votes for VCM with punishment but prefers in at least one period each of the other two institutions.

Overall, we observe 291 (26.94%) institutional switches between periods in the partner and 203 (24.52%) switches in the stranger treatment. Whereas switches away from the reward institution go approximately one half into the standard VCM and one half into the punishment institution (48.85% into standard VCM in Partner vs. 43.04% in Stranger), switches away from the standard VCM and the punishment institution go in large majority into the reward institution (71.59% and 73.61% in Partner vs. 76.47% and 57.14% in Stranger).

Figure 2C.2: Preferences for *L* over time in partner treatment (for each person separately)

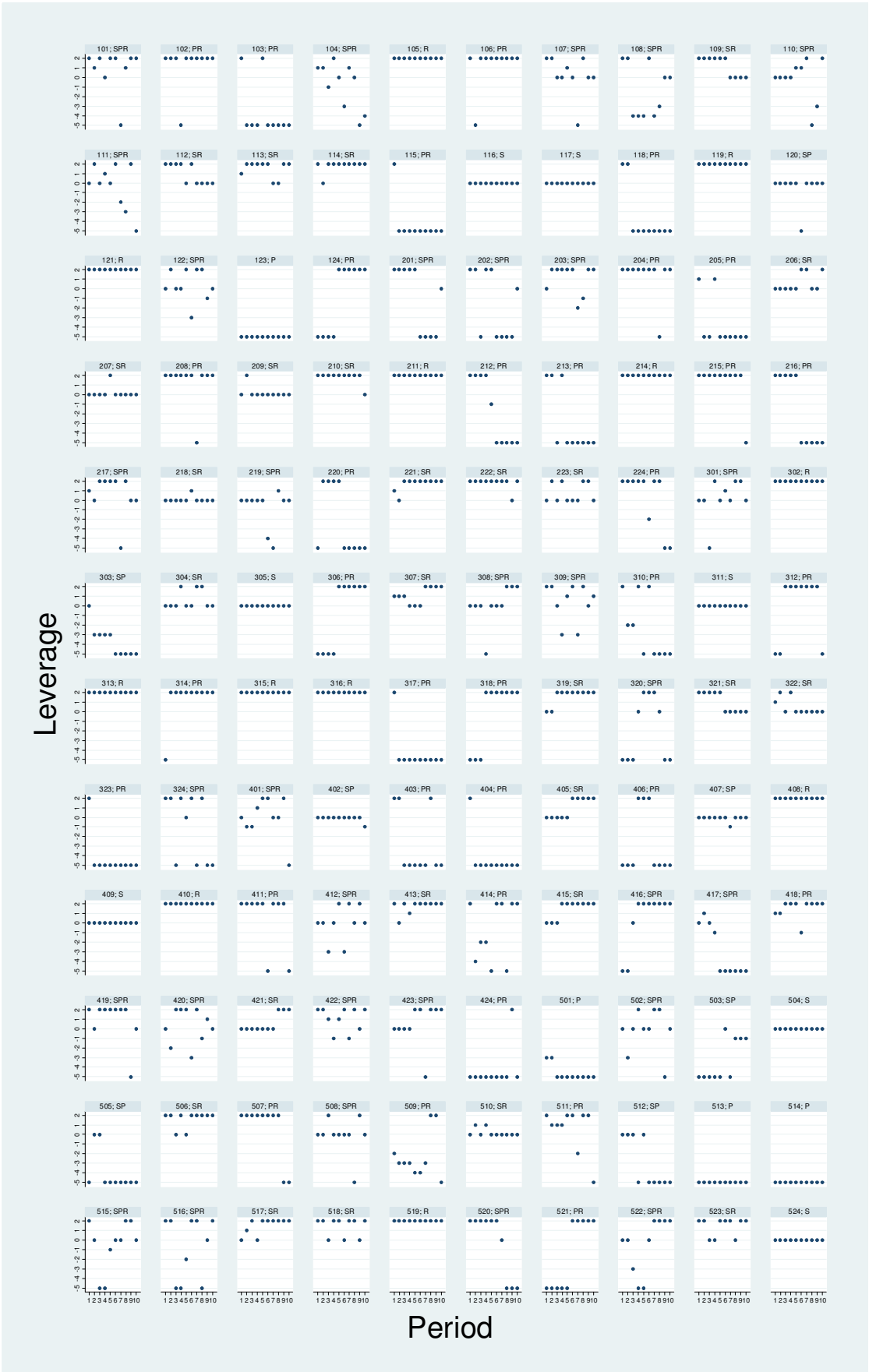
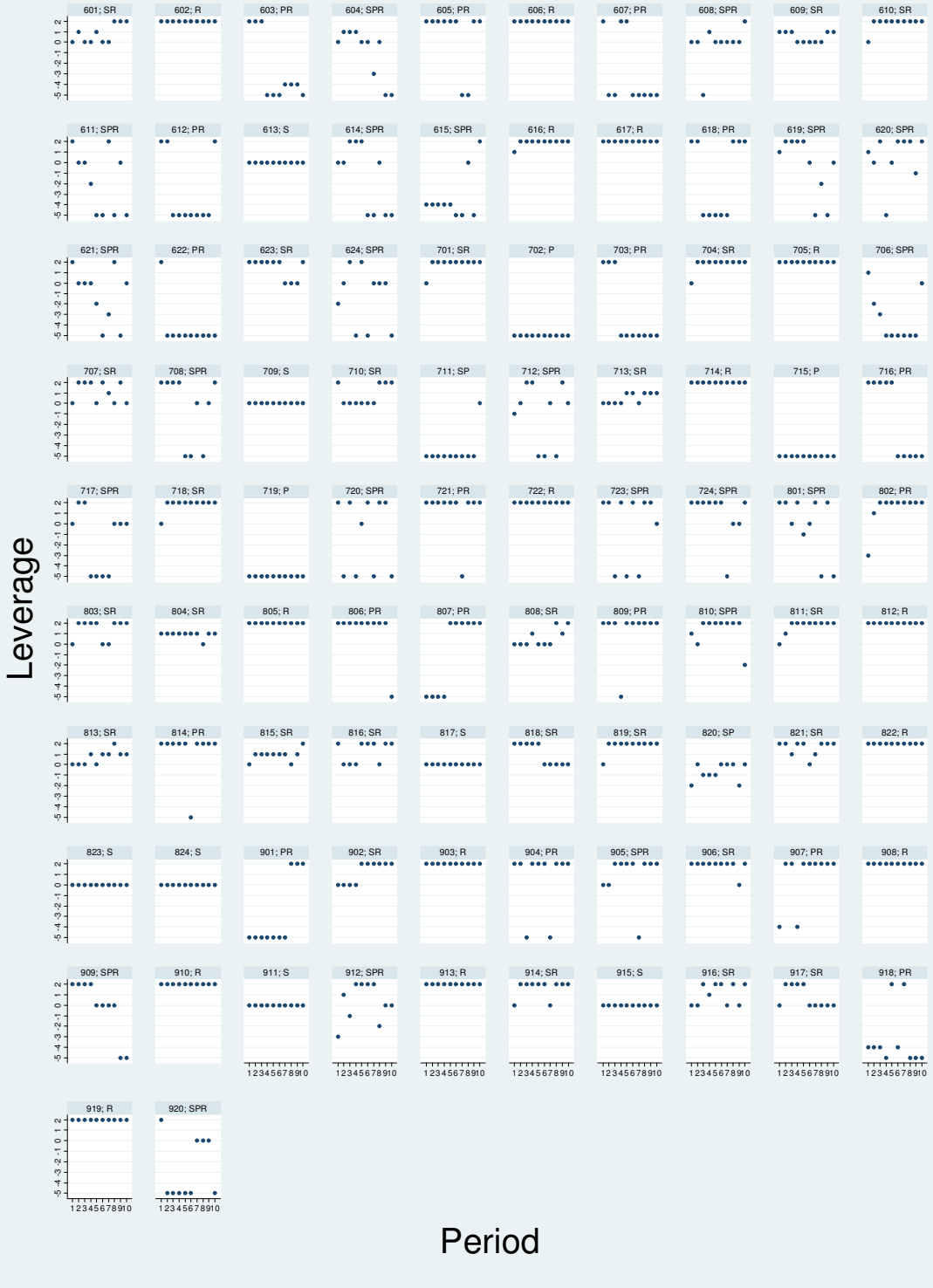
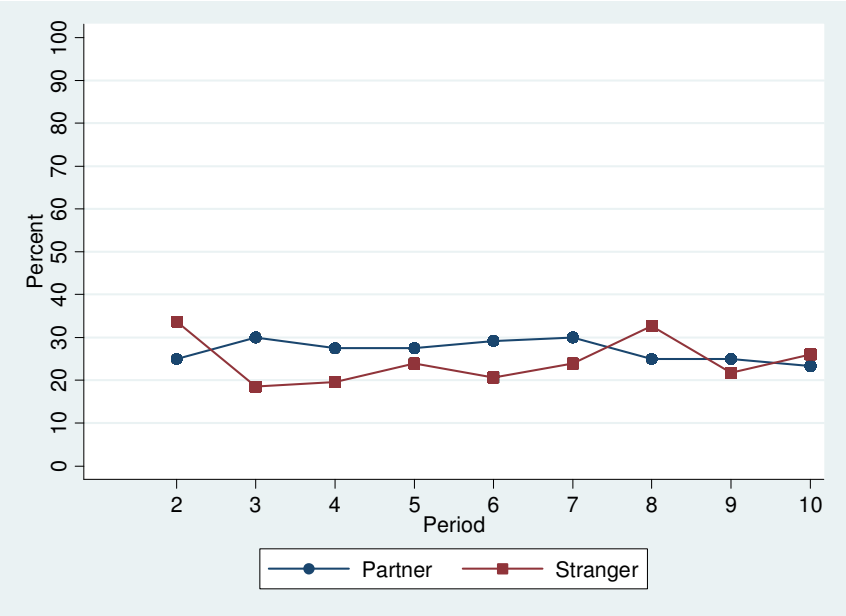


Figure 2C.3: Preferences for *L* over time in stranger treatment (for each person separately)



Institutional switches over time

Figure 2C.4: Percentage of institutional switches compared to previous period by treatment



Evolution of contributions

Figure 2C.5: Average contributions over time in partner treatment

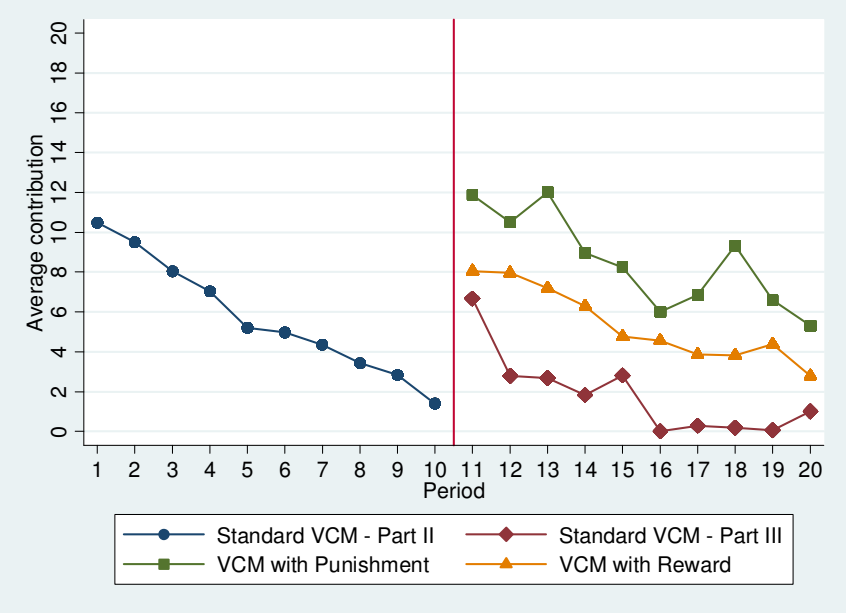
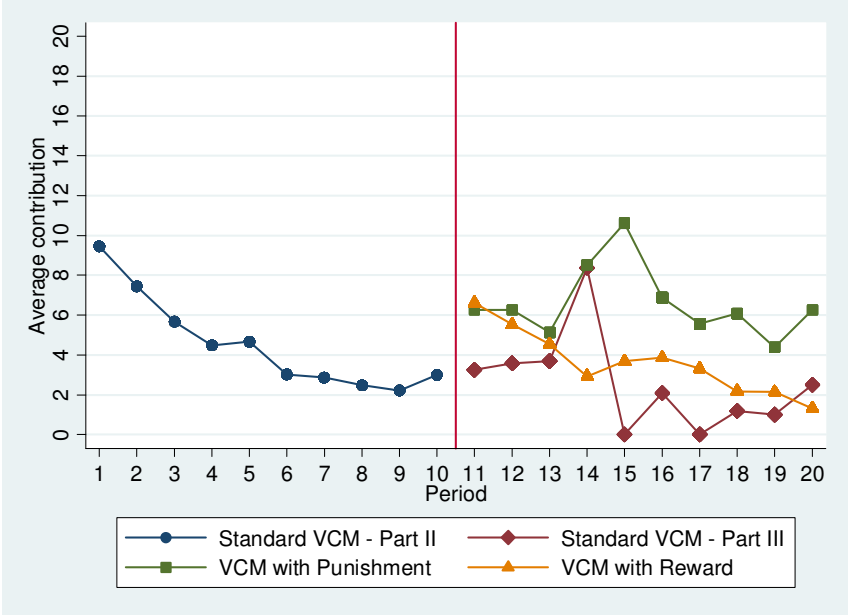


Figure 2C.6: Average contributions over time in stranger treatment



Chapter 3

The Team Allocator Game: Allocation Power in Public Goods Games[86]

3.1 Introduction

Consider a team work setting with a straightforward hierarchy. There is one team member that, although contributing to the team effort, has some discretionary power to allocate gains from team production to all team members. Individual effort levels in the team are observable, but not verifiable. Hence, contracts on effort are infeasible, and the team faces a modified social dilemma: it is individually rational for a selfish ordinary team member to withhold effort, even though it would be socially optimal to provide effort. The incentives for the team member with allocation power, henceforth denoted team allocator, could be different; we shall return to this issue momentarily. In real life, we often find teams with a natural or exogenously imposed hierarchical structure that gives one team member (a team leader or manager) property rights over team output: small work teams, for instance, in consulting companies, sports teams, music bands, military units, or families readily come to mind. However, also political parties with a designated leader have a similar incentive structure when effort contributions to an election campaign are considered to have a social dilemma component.[87]

This study is the first that provides a rigorous empirical test of the (behavioral) incentive effects of such a team structure. We analyze a modified public goods game theoretically and implement it experimentally in the laboratory. From the several conceivable implementations of allocation power of a team member, we chose the most parsimonious one. In our *team allocator game (TAG)*, each team member – regardless of whether the member is an *ordinary team member* or the *team allocator* – can contribute to a public account. The sum of contributions is multiplied by an efficiency factor larger than one, but – in contrast to the

[86] This chapter is joint work with Martin Kocher and Gerhard Riewe.
[87] All these examples have in common that the team output is not a pure public good but unequally dividable among the team members. A pure public good, of course, entails no allocation power.

standard public goods game – not distributed equally among all team members. Rather, the team allocator receives the entire amount in the public account and has full discretionary power over the allocation of the revenues from the account within the team. More precisely, she can implement any distribution of the benefits from the public account over the ordinary team members and herself.

It is straightforward to show that in such a setting, ordinary team members with standard preferences have no incentive whatsoever to contribute to the public account, whereas the team allocator will contribute her entire endowment if the public goods mechanism is linear. Our experimental results, however, interestingly show that the level of contributions in the team allocator game is significantly higher than in an appropriate control treatment in which there is no team allocator, i.e. there is an equal split of the public account, but one team member is forced to contribute her entire endowment. In other words, giving one team member dictator or allocation power leads to much higher levels of cooperation than expected. We show that team allocators' distribution behavior influences, together with the time horizon of the team interaction, the pattern of contributions. Although there is some heterogeneity in the behavior of team allocators, we find that they predominantly use the reward channel in case of high contributions, i.e. they allocate large shares of the public account to cooperating ordinary team members, but they also tend to punish non-contributors by excluding them from the benefits from the public account. Overall, the returned amounts to contributors are astonishingly high and generate strong contribution incentives for ordinary team members. We provide theoretical evidence that such a generous behavior of team allocators could be caused by other-regarding preferences such as inequity aversion (Fehr and Schmidt, 1999) or maximin-orientation (Charness and Rabin, 2002), but the repeated-game aspect plays a role as well.

In general, it seems that teams with a straightforward hierarchy – with a team member that dominates the allocation process – have an advantage over teams with members of equal standing and an automatic equal distribution of the team benefits. They are more likely to overcome the social dilemma inherent to public good provision, i.e. team effort provision. Our result is the more remarkable since the described mechanism can be implemented without any monetary costs. Other mechanisms to sustain cooperation that have been studied extensively in the literature – the most prominent one is an informal sanctioning mechanism (starting with the seminal paper by Fehr and Gächter, 2000) – bring about substantial costs. While, for instance, an informal punishment option increases contributions dramatically, its efficiency balance depends very much on the length of the interaction among the team

members (Gächter et al., 2008). On the contrary, an informal reward option can be more efficient but is typically less effective in sustaining high contribution levels. Thus, from an efficiency perspective, the implementation of hierarchy in teams seems especially positive when compared to other mechanisms.

Our study is related to the huge literature of institutional provisions in social dilemmas. This literature has, for instance, studied the effects of punishment (e.g., Yamagishi, 1986; Ostrom et al., 1992; Fehr and Gächter, 2000; Masclet et al., 2003; Casari, 2005; Noussair and Tucker, 2005; Anderson and Putterman, 2006; Carpenter, 2007a; Denant-Boemont, et al., 2007; Sefton et al., 2007; Egas and Riedl, 2008; Gächter et al., 2008; Herrmann et al., 2008; Masclet and Villeval, 2008; Nikiforakis, 2008; Nikiforakis and Normann, 2008; Casari and Luini, 2009; Ule et al., 2009; Nikiforakis, 2010; Gächter and Herrmann, 2011), the effects of reward (e.g., Andreoni et al., 2003; Sefton et al., 2007), the effects of communication before the contribution decision (e.g., Isaac and Walker, 1988; Ostrom et al., 1994; Cason and Khan, 1999; Brosig et al., 2003; Bochet et al., 2006; Bochet and Putterman, 2009), the effects of an expulsion option from the benefits of the public good (e.g., Cinyabuguma et al., 2005), or the effects of voluntary association (e.g., Page et al., 2005; Charness and Yang, 2008) on contribution levels in and the resulting efficiency of social dilemmas. There is also a literature on the formal implementation of institutions in social dilemmas, usually by a voting mechanism (e.g., Kroll et al., 2007; Kosfeld et al., 2009).

All related papers on institutional provisions mentioned above share the feature that team members are equal in their personal endowments and options to implement and use the institutional mechanism. In other words, there is no hierarchy within the team. Examples of papers that study cooperation-fostering institutions with unequal team members are rare.[88] One exception that we are aware of is Reuben and Riedl (2009). The paper is, however, only loosely related to ours, because they analyze the effects of endowment differences in a public goods game on norm enforcement. Another exception, Cárdenas et al. (2009), is related more closely. They analyze a specific problem in collective water management that is modeled as a public good with asymmetric access. More precisely, in their setup there is sequential access of the team members to the benefits from the public good. The idea of sequential access is intended to capture the situation of a collective water supply with the natural feature that upstream users (farmers) can appropriate benefits from the public good before downstream users. Their main finding in terms of cooperation is that asymmetric appropriation leads to

[88] There is a large literature on asymmetries in standard public goods games in the absence of norm-enforcement devices. For reasons of succinctness, we do not discuss this literature here.

lower levels of cooperation than the usual symmetric appropriation in the standard linear public goods game (voluntary contribution mechanism).[89]

Another way of looking at our mechanism is in relation to the seminal trust game (Berg et al., 1995). Our mechanism can be viewed as a collective trust game in which the amount that can be returned by the trustee (the team allocator) depends on the collective level of trust by the trustors (the ordinary team members). Trust games with more than one trustor are for example studied in Cassar and Rigdon (2011). However, their trustees are more restricted in their allocation power as they cannot allocate benefits from one trustor's investments to another trustor.[90]

In reality, the allocation power of the team allocator is sometimes limited by law or contract, or allocation power is shared by several team members. Analytically, it makes nevertheless sense to start with a comparison of the two extremes: full allocation power by one team member versus an automatic equal distribution of the team benefits (with one team member being forced to contribute the entire endowment to keep incentives constant across the two conditions). Our study thus intends to provide a benchmark for a hierarchical social dilemma setting with allocation power. Future studies should address different contractual limitations associated with allocation power.

The remainder of the chapter proceeds as follows: In Section 3.2 we introduce our experimental design and describe the procedures of the experiment. Section 3.3 derives theoretical predictions. Section 3.4 reports the experimental results, and Section 3.5 discusses our findings and concludes the chapter.

3.2 Experimental design and procedures

In the following we describe the basic experimental setup (Section 3.2.1) and the details of the experimental procedure (Section 3.2.2).

3.2.1 Basic setup of the team allocator game

Let $I = \{1, 2, \dots, n\}$ denote a team of n subjects who interact in T periods with subject 1 being called the *team allocator (TA)* and subjects $2, \dots, n$ called the *ordinary team members*

[89] Finally, our research is also related, at least loosely, to recent contributions on leadership in public goods games. See, for instance, Güth et al. (2007), Levati et al. (2007), Rivas and Sutter (2011), or Gächter et al. (2010).
[90] For a recent meta-analysis of trust games, see Johnson and Mislin (2011).

(OTMs). Each period $t \in \{1, 2, \dots, T\}$ consists of two stages. In stage 1, each individual $i \in I$ receives an endowment E which can be allocated either to her private account or to a public account. The contribution of individual i to the public account in period t, denoted $c_{i,t}$, must satisfy $0 \leq c_{i,t} \leq E$. Let C_t be the sum of all team members' contributions in period t (i.e. $C_t = \sum_{j=1}^{n} c_{j,t}$). In order to retain the public goods nature C_t is multiplied by a factor γ, which satisfies $1 < \gamma < n$.[91]

In the second stage, the TA can freely distribute the amount γC_t among the team members (the OTMs and herself), following only two restrictions for the returned amount. Every team member has to get a non-negative amount that cannot be greater than γC_t, and the sum of all returned amounts has to be equal to γC_t. Formally, the returned amount is denoted by $d_{i,t}$, with i being the receiving team member:

$$0 \leq d_{i,t} \leq \gamma C_t \; \forall i, \quad \sum_{j=1}^{n} d_{j,t} = \gamma C_t \tag{3.1}$$

Individual team member i's payoff from the TAG in period t is then given by

$$\pi_{i,t} = E - c_{i,t} + d_{i,t}. \tag{3.2}$$

3.2.2 Experimental procedures

The experiment implements two treatments: (i) treatment *TAG* and (ii) treatment *VCM+*. TAG is a treatment according to the setup laid out in Section 3.2.1 with the following parameters: team size $n = 4$, endowment per period $E = 20$ points (the experimental currency unit)[92], $\gamma = 1.6$, and the number of periods $T = 10$. Returned amounts $d_{i,t}$ can have up to one decimal.[93] Experimental participants are matched randomly in teams at the beginning of the experiment and one randomly chosen team member is assigned the role of TA. We use a partner design with fixed subject IDs that allows building reputation because a one-shot interaction in a team with hierarchy seems less realistic. Roles are kept over the course of the experiment. Since we are not interested in studying irrational behavior or potentially complicated signaling through contributions by the TA that are below the full

[91] Indeed, γ could also be smaller than 1 or larger than n in the TAG without changing the incentives for OTMs. In contrast to the standard public goods game, there is no individual incentive to contribute to the public account, no matter how high γ is. The condition is just imposed to keep the setup comparable to the standard voluntary contribution mechanism. A $\gamma < 1$ changes, however, the incentives for the TA.

[92] At the end of the experiment earned points from all periods are summed up and converted into euro using the following exchange rate: 1 point = 4 euro-cent.

[93] Note that we allow for one decimal place to ensure that the entire amount of γC_t can be distributed to the team members. This also gives TAs the ability to return exactly 1.6 times the invested amount to each OTM.

endowment, we automatically enforce full contributions for a TA, i.e., $c_{1,t} = E$. This is innocuous because both a completely selfish and an other-regarding TA would want to contribute the highest possible amount.[94] All details of the setup and all parameters are common knowledge.

VCM+ is the appropriate control treatment for TAG. It is a standard voluntary contribution mechanism with identical parameters and provisions as in TAG. In order to align incentives, however, we need to randomly select one team member that is forced to contribute her entire endowment to the public account – hence, the plus in the notation.[95] The team member who is forced to contribute the entire endowment is the same in every period of the experiment and will be denoted, analogous to the denomination of the TA in the TAG, as subject 1. Note that the benefit from the public account in VCM+, γC_t, is distributed equally among all team members, just as in a standard VCM.[96]

Information conditions are as follows: In the TAG treatment, the TA receives the full vector of individual contributions within her team before deciding about the returned amounts to each team member in stage two of the game. In both treatments at the end of each period, all team members are informed about the vector of contributions within their teams, the resulting benefit from the public account, the distribution of this benefit among the team members (either equally in the VCM+ or according to the allocation decision of the TA in the TAG), and the final individual profits from this period. Hence, information conditions are identical across the two treatments.

An experimental session consisted of two parts (instructions can be found in Appendix 3A) in which the second part was either the TAG or the VCM+ treatment. In the first part we used a social value orientation questionnaire (henceforth referred to as ring test) to obtain an independent measure of an individual's social motivation (i.e. her generalized other-regarding preferences).[97] The measure from the ring test helps us to assess one of our main research

[94] An anti-social TA would not necessarily want to contribute the highest possible amount. However, we will show later on that we do not have any anti-social TA in our experiment.

[95] Partial coercion does not change contribution incentives for unforced contributors compared to a standard VCM. This is shown in a recent study by Cettolin and Riedl (2011). They implement two coercion treatments (low and high) in which they force one randomly selected group member to contribute at least a minimum amount (approximately 25% and 75% of the endowment, respectively). The authors show that partial coercion has no influence on average contributions beyond the pure coercion effect, i.e. non-coerced subjects do not contribute significantly different amounts than subjects in a control VCM. Cettolin and Riedl argue that the lack of a cooperative intention may prevent unforced conditional cooperators from increasing their contributions.

[96] For the control treatment VCM+, the condition $1 < \gamma < n$ ensures that we have a social dilemma. See footnote 91 for a brief discussion.

[97] Van Lange et al. (1997) provide a review on the use of the ring test in the psychological literature. Economic applications of this measure can, for example, be found in Offerman et al. (1996), Park (2000), Brosig (2002), van Dijk et al. (2002) or very recently in Sutter et al. (2010).

questions, namely to what extent the behavior of TAs is driven by an intrinsic motivation or by opportunistic maximization behavior.

In the ring test, individuals choose 24 times between two possible pairs of payoffs for themselves and another person (see Appendix 1A for details). The recipient remains the same throughout the entire test and answers herself the same set of tasks (thereby, vice versa, influencing the first person's payoff). The test is fully incentivized since all 24 selected pairs are payoff relevant. However, the profit from the ring test is not revealed before the end of the second part (the VCM+ or the TAG, respectively) in order to avoid any income spill-over effects within the ring test or from the ring test to the VCM+/TAG.[98]

By calculating the sum of all 24 selected pairs, one can determine the overall amount of money allocated to the person herself (X) and the other person (Y). The ratio Y/X determines then a vector $\vec{A}$ and thus a certain angle in an X- Y-coordinate system. Dependent on this angle, subjects can be sorted into eight behavioral types (individualism, cooperation, altruism, martyrdom, masochism, sadomasochism, aggression and competition; see Figure 1A.1 in Appendix 1A) which reflect their social orientation. With the 24 choices one can also measure a participant's consistency in her payoff choices. When using the data from the ring test in our analysis, we focus only on TAs with a consistency measure of at least 2/3.[99] Moreover, we concentrate on two behavioral types, *individualistic* and *cooperative* types, as there is no single TA that is classified differently by the mechanism. This is not unusual because behavioral types that consistently follow other motivations are very rare.

Before the start of the first part, our subjects received written instructions only for the ring test, but they knew that there would be a second part in the experiment and that this part would be unrelated to the first part. Upon completion of the ring test, subjects received instructions for the second part: either the TAG or the VCM+. Instructions of both parts were read aloud to ensure common knowledge of the rules, and subjects were given plenty of time to ask questions in private before the start of each part.

At the end of the experiment, before private cash payment, subjects finally answered a couple of questions about their decisions in the experiment and a post-experimental questionnaire, including questions regarding socio-economic variables such as gender, age and major. The computer-based sessions were conducted at the experimental laboratory

[98] A subject's recipient in the ring test could by chance be also a member of the same team in the second part of the experiment, as we used an unrestricted random draw mechanism. However, this does not create any problems, since no feedback was provided before completion of the second part.

[99] Note that there exists no standard consistency threshold in the literature. While Park (2000) classifies only subjects with a consistency measure of 75% or more, Brosig (2002) uses a remarkably low threshold of 25%. We decided to implement a relatively high threshold. However, shifting this value downwards or even including all TAs does not yield different results.

MELESSA of the University of Munich between July 2010 and September 2010 using the experimental software z-Tree (Fischbacher, 2007) and the organizational software Orsee (Greiner, 2004). A total of 144 undergraduate students from all disciplines participated in six sessions with 24 participants each. Three sessions implemented treatment TAG, three sessions treatment VCM+. The six sessions provide us with 18 statistically independent observations for each of the two treatments. The sessions lasted up to 90 minutes including everything from the instructions to final payments, and the average earnings were 19.08 EUR, including a show-up payment of 4.00 EUR. No participant was allowed to take part in more than one session, and the assignment of subjects into treatments was random. Decisions were taken anonymously in cubicles, and communication among participants was prohibited.

3.3 Theoretical predictions

In the following, we formulate theoretical predictions for our two treatments. We start with straightforward hypotheses based on the assumptions of purely selfish and rational decision makers ("standard preferences"). In a next step, we then move to hypotheses based on two prominent models taking other-regarding preferences into account. Finally, we take care of the repeated interaction in our experiment by focusing on reputation formation.

3.3.1 Predictions based on standard preferences (homo oeconomicus model)

Our two treatments are finitely repeated games of perfect information. Assuming common knowledge of rationality and selfishness and using backward induction it is clear that in the TAG the TA will not return any positive amount to the OTMs in the second stage of period 10. Therefore, OTMs will not contribute anything to the public account in period 10 because any contribution would be "lost". The same rationale holds for all prior periods. Consequently, contributing nothing to the public account is a dominant strategy for all OTMs, i.e. $c_{i,t} = 0 \; \forall i \neq 1$ and t. The TA herself is forced to contribute $c_{1,t} = E = 20$ in all periods, but she would have an incentive to do so anyway because $\gamma > 1$. Regarding the distribution of the public account we have $d_{i,t} = 0 \; \forall i \neq 1$ and t whereas the TA receives in each period $d_{1,t} = \gamma C_t = 32$.

For treatment VCM+ the standard logic of the public goods game applies. Since $1 < \gamma < n$, the marginal per capita return from investing into the public account is smaller than one. Hence, it is a dominant strategy for OTMs to contribute nothing to the public account, i.e.

$c_{i,t} = 0\ \forall i \neq 1$ and t. The forced contributor, on the other hand, has to contribute $c_{1,t} = E = 20$ in each period, and the automatic equal distribution of the public account yields $d_{i,t} = \gamma C_t/n = 8\ \forall i, t$.

Proposition 3.1. *Under standard preferences, OTMs contribute zero in each period, irrespective of treatment. The TA in the TAG always allocates the entire public account to herself.*

3.3.2 Predictions based on other-regarding preferences

We focus on two prominent models that both belong to the class of outcome-based social preference models (at least in the way we model them): the inequity aversion model by Fehr and Schmidt (Fehr and Schmidt, 1999) and the welfare-oriented model by Charness and Rabin (Charness and Rabin, 2002).

3.3.2.1 Fehr and Schmidt (1999) preferences

The model by Fehr and Schmidt (1999) assumes that subjects suffer from inequity within their reference group. More precisely, a subject i benefits from her own payoff π_i but compares it with the payoff of the $n-1$ other members in her reference group. The corresponding utility function is the following:

$$U_i(\pi) = \pi_i - \alpha_i \frac{1}{n-1}\sum_{j \neq i} \max\{\pi_j - \pi_i, 0\} - \beta_i \frac{1}{n-1}\sum_{j \neq i} \max\{\pi_i - \pi_j, 0\} \qquad (3.3)$$

The vector $\pi = (\pi_1, \dots, \pi_n)$ denotes the monetary payoffs and α_i and β_i represent subject i's individual attitude towards inequity. The two weights are restricted to $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$. They control for the impact of utility losses from disadvantageous inequity (α_i) and advantageous inequity (β_i), respectively.[100]

If we assume that the TA in the TAG is inequity-averse and the team is the relevant reference group, then a TA might be willing to reduce payoff differences within the team by returning positive amounts to the OTMs. Note that the weight α_i does not play any role here, because the TA will never reduce the amount allotted to herself below the level of full payoff equalization as this reduces her own payoff *and* increases inequity. Thus, only the weight β_i matters for TA's decisions. If the TA distributes one point from the public account to an OTM

[100] Note that for $\alpha_i = \beta_i = 0$ the model collapses into the case of standard preferences.

instead of putting it into her own pocket, she will reduce her own payoff by 1 and decrease inequity, on average, by 4/3 (regarding the receiving OTM by two points and regarding both other OTMs by one point). Thus, returning positive amounts is optimal if $-1 + \beta_1 \cdot 4/3 \geq 0$ or $\beta_1 \geq 0.75$.

This yields the following equilibria in the one-shot game: If $\beta_1 < 0.75$, the TA takes the entire public account for herself, which implies zero contributions of OTMs irrespective of whether they are selfish or whether they are other-regarding, i.e. $c_i = d_i = 0 \; \forall i \neq 1$ and $d_1 = \gamma C = 32$. If $\beta_1 > 0.75$, the TA returns positive amounts to fully equalize period payoffs. All OTMs, therefore, have an incentive to contribute their full endowment, even when they are completely selfish and rational, and of course, the more so if they are other-regarding. Hence, we have $c_i = E = 20 \; \forall i \neq 1$, and $d_i = 32 \; \forall i$ as the only subgame-perfect equilibrium. If $\beta_1 = 0.75$, the TA is indifferent in the way she allocates the public account (as long as she is not worse off than one of the other team members). Hence, multiple equilibria exist in this case and cooperation between some or all team members may occur. Thus, TAs that are sufficiently averse to advantageous inequity ($\beta_1 \geq 0.75$) can generate full cooperation and payoff equalization in the one-shot version of the TAG. Regarding potential repeated game effects and reputation building, Section 3.3.2.3 will discuss details.

It is noteworthy that Fehr and Schmidt (1999) preferences can predict full cooperation in our VCM+ treatment. Using Proposition 4 of Fehr and Schmidt (1999, p. 839) it is, however, obvious that for our parameter values cooperation can only be achieved if all OTMs are sufficiently averse to advantageous inequity, i.e. $\gamma/n + \beta_i \geq 1$ or $\beta_i \geq 0.6 \; \forall i \neq 1$. Asymmetric equilibria in the one-shot game do not exist for our setup. According to the parameter distribution given in Fehr and Schmidt (1999, p. 844), the probability of having three OTMs with $\beta_i \geq 0.6$ in one team is $0.4^3 = 6.4\%$. As Fehr and Schmidt (1999) do not provide data for a threshold of 0.75, we cannot infer the probability of meeting a TA with $\beta_i \geq 0.75$ from their paper. From all calibration results that are available, it is clear that the probability of meeting a TA with sufficiently high β_i in order to induce full cooperation is higher than the 6.4%. Hence, full cooperation in the one-shot TAG treatment is expected to be more prevalent than in the VCM+ treatment. Again, the discussion of repeated game effects is relegated to Section 3.3.2.3.

Proposition 3.2. *With Fehr and Schmidt (1999) preferences, the TA in the TAG is willing to distribute positive amounts to OTMs if $\beta_1 \geq 0.75$, i.e. if she is sufficiently averse to advantageous inequity. Full cooperation and full payoff equalization within the team is an*

equilibrium in this case. If $\beta_1 < 0.75$, in the one-shot game the TA will take the entire benefit from the public account for herself, and no OTM has an incentive to contribute. Full cooperation can also be an equilibrium in the VCM+ treatment; however, it requires $\beta_i \geq 0.6$ for all OTMs.

3.3.2.2 Charness and Rabin (2002) preferences

Charness and Rabin (2002) assume that subjects care about social welfare. Their model includes a subject's own payoff and, additionally, two components of social welfare: the minimum payoff in a group (the "Rawlsian" motive) and the sum of all group members' payoffs (the efficiency concern). More precisely, the utility function in their general model (see their Appendix 1) with only outcome-based components looks as follows:[101]

$$U_i(\pi) = (1-\lambda_i)\pi_i + \lambda_i[\delta_i \min(\pi_1, \dots, \pi_n) + (1-\delta_i)(\pi_1 + \pi_2 + \cdots + \pi_n)] \qquad (3.4)$$

The vector $\pi = (\pi_1, \dots, \pi_n)$ denotes the monetary payoffs within the group of n subjects and λ_i and δ_i are individual weights (i.e. $\lambda_i, \delta_i \in [0, 1]$). The first weight, λ_i, captures how much an individual cares for social welfare relative to her own payoff.[102] The second weight, δ_i, controls for the influence of the "maximin"-aspect relative to the general efficiency concern.

Again, we first look at the one-shot game and relegate any discussion regarding repeated interaction to Section 3.3.2.3. As a TA's choice in the TAG is purely distributional, i.e. the sum of team members' payoffs is not affected by her decision, only the "Rawlsian" motive of social welfare matters for a TA's decision. TAs compare the utility loss from a reduction in own payoff, $1-\lambda_1$, with the utility gain from increasing the minimum payoff in the team ($\lambda_1\delta_1$). This implies that TAs never return amounts to OTMs beyond the level of full payoff equalization. Note further that the number of subjects s that lie at the minimum payoff matters, because it determines by how much the minimum can be raised with one point. If there is more than one individual at the minimum, each point has to be split equally among all affected subjects to obtain the maximum increase in the minimum payoff. Thus, returning positive amounts to OTMs is optimal for a TA as long as $1-\lambda_1 \leq \lambda_1\delta_1 \cdot 1/s$ or $\delta_1 \geq s \cdot (1-\lambda_1)/\lambda_1$.

As s cannot be smaller than 1, $0.5 \leq \lambda_1 \leq 1$ is a necessary condition to ensure reasonable values of δ_1. However, $\lambda_1 \geq 0.5$ (and δ_1 sufficiently large) makes positive returned amounts to OTMs optimal, only as long as there is a single OTM with minimum earnings. Once the

[101] Note that we consider here only the outcome-based version of the model and neglect the role of intentions as the more complex model with intentions does not seem suitable for deriving specific predictions in our setup.
[102] For $\lambda_i = 0$, the Charness and Rabin (2002) model nests standard preferences.

minimum is raised to the level of the second-lowest payoff or once there are two subjects with the same minimum earnings, the condition tightens to $\lambda_1 \geq 2/3$ (and δ_1 appropriately). Thus, in contrast to Fehr and Schmidt (1999), Charness and Rabin (2002) preferences can lead to a partial equalization of profits. Full payoff equalization in equilibrium will only obtain if λ_1 is large enough to make redistribution profitable in the case the points have to be split among all three OTMs, i.e. $\lambda_1 \geq 0.75$ (and δ_1 appropriately).

This implies the following: If $\lambda_1 \geq 0.75$ (and δ_1 appropriately), there is an equilibrium in which all OTMs contribute their full endowment even if they are completely selfish and rational and the more so if they are other-regarding, i.e. $c_i = E = 20 \; \forall i \neq 1$, and $d_i = 32 \; \forall i$.[103] If $\lambda_1 < 0.5$, selfish OTMs choose $c_i = 0$, while $E = 20$ is contributed by OTMs who care sufficiently about efficiency (requiring $\lambda_i \geq 0.625$ and δ_i sufficiently low[104]). If $0.5 \leq \lambda_1 < 0.75$, full cooperation will not be obtained with selfish and rational OTMs. However, partial cooperation with one or two OTMs contributing positive amounts is possible (the latter only for $\lambda_1 \geq 2/3$). Again, if all OTMs care sufficiently about efficiency, full cooperation will arise.

In the VCM+ treatment, OTMs have to care sufficiently for social welfare to have an incentive to contribute to the public account. Note that an increase in the contribution level decreases an OTM's own payoff by $1 - \gamma/n$, increases the minimum payoff in the team by γ/n and increases the sum of all team members' payoffs by $\gamma - 1$. Hence, contributing positive amounts is optimal if $(1 - \lambda_i)(1 - \gamma/n) \leq \lambda_i \delta_i \cdot \gamma/n + \lambda_i(1 - \delta_i)(\gamma - 1)$ or $\delta_i \leq 6 - 3/\lambda_i$. For δ_i to be non-negative, this requires $\lambda_i \geq 0.5$. Full cooperation by all group members will therefore only arise if all OTMs fulfill $\lambda_i \geq 0.5$ (and δ_i appropriately).

Proposition 3.3. *With Charness and Rabin (2002) preferences, the TA in the TAG might be willing to return positive amounts to OTMs if* $\lambda_1 \geq 0.5$ *(and* δ_1 *sufficiently high), i.e. if she is sufficiently "maximin"-oriented. However, full payoff equalization can only be achieved if* $\lambda_1 \geq 0.75$ *(and* δ_1 *appropriately). Full cooperation is also possible if all OTMs care sufficiently about efficiency. In the VCM+ treatment, full cooperation will only arise if all OTMs fulfill* $\lambda_i \geq 0.5$ (and δ_i *appropriately).*

[103] There is, of course, indifference of the TA between distributions in case of $\lambda_1 = 0.75$. This leads to multiple equilibria sustaining also contribution levels below 20.

[104] To see this, note that if a single OTM contributes one point to the public account, both the OTM's payoff and the minimum payoff is reduced by 1, whereas the sum of payoffs increases by $\gamma - 1$. Thus, contributing is advantageous if $(1 - \lambda_i) + \lambda_i \delta_i \leq \lambda_i(1 - \delta_i)(\gamma - 1)$ or $\delta_i \leq 1 - 1/(1.6\lambda_i)$. This implies $\lambda_i \geq 0.625$ (and δ_i appropriately). Note that the restriction on δ_i becomes weaker for further OTMs contributing one point (without changing the requirement on λ_i) as their contributions do not decrease the minimum anymore.

To sum up, in contrast to the case of standard preferences both models of other-regarding preferences predict (for appropriate parameter values) that TAs in the TAG return positive amounts to OTMs. Moreover, such behavior can induce full cooperation and payoff equalization within the team. Both models can also explain full cooperation in the VCM+ treatment. However, an equilibrium with full cooperation in the VCM+ requires that *all* OTMs have sufficiently strong other-regarding preferences, whereas in the TAG it is sufficient that the TA has strong enough other-regarding preferences. One noteworthy difference between the two discussed models is that in the Fehr and Schmidt (1999) model according to our parameterization, there are no asymmetric equilibria, whereas there are such equilibria for the Charness and Rabin (2002) model in both treatments.

3.3.2.3 Heterogeneous social preferences and repeated interaction (reputation model)

In a *repeated* game with heterogeneous social preferences, the argument that TAs return positive amounts to OTMs holds a fortiori. With repeated interaction, additionally, selfish TAs have an incentive to act as if they were other-regarding, because the future stream of income created by mimicking is larger than the costs of acting non-selfishly in a specific period.[105] This is true until the ultimate or until the pen-ultimate period, in which the opportunistic TAs that mimic other-regarding TAs start appropriating the benefits from the public account. By returning positive amounts to OTMs until the last or the second-to-last period, TAs induce higher contributions by the OTMs in future periods that the TA can subsequently pocket for herself. The argument holds only for the TAG treatment and not for the VCM+ treatment, but depending on the model there might also be additional contribution incentives in the latter treatment. We refrain from characterizing all equilibria in the repeated game because the argument has been used and formalized straightforwardly in connection with trust contracts (see, e.g., Fehr et al., 2007).

Note finally that both the Fehr and Schmidt (1999) and the Charness and Rabin (2002) model, taken literally, would yield a very high number of either zero or full contributions and no intermediate contribution amounts.

[105] We implicitly assume in this argument that players are not entirely sure about their opponent's type, i.e. we relax the common knowledge assumption.

Proposition 3.4. *With heterogeneous social preferences and repeated interaction, the TA in the TAG might change behavior across periods to profit from reputation effects. There is an incentive for completely selfish TAs to mimic the behavior of other-regarding TAs until the ultimate or the pen-ultimate period of the game.*

3.4 Experimental results

We start with a comparison of contributions and profits in the TAG and the VCM+ treatment in Section 3.4.1. Section 3.4.2 decomposes contribution behavior in the TAG further, and Section 3.4.3 analyzes the details of the TA's behavior as well as OTMs' optimal replies.

3.4.1 Contributions and profits in TAG and VCM+

Table 3.1 provides a first upshot of our main results regarding average contribution levels and average profits. It is immediately apparent that the TAG elicits higher contribution levels and, hence, leads to higher profits than the control treatment VCM+. Since there is always one team member that is forced to contribute the entire endowment, the difference between the treatments is solely driven by the contributions of the OTMs. The first row of Table 3.1 shows that mean contribution levels of OTMs differ by five points or 25% of the endowment. This difference is clearly significant ($p < 0.05$, Mann-Whitney U-test, $N = 36$).[106]

Table 3.1: Mean contributions and profits (in points) by treatment

		VCM+	TAG
Mean contribution	OTMs	9.88**	14.95**
	Forced contributors/TAs	20	20
	All members	*12.41***	*16.21***
Mean profit	OTMs	29.98***	26.54***
	Forced contributors/TAs	19.86***	39.30***
	All members	*27.45***	*29.73***

Note: Difference between VCM+ and TAG significant at: *** 1% level; ** 5% level; * 10% level.

Significance can also be shown by an OLS regression on contributions (cf. model 1 in Table 3.2) where only the treatment dummy *TAG* is included ($p < 0.01$, standard errors

[106] All non-parametric tests that we use in this chapter are two-sided tests.

clustered on the team level).[107] The significant difference in OTMs' contributions, by the nature of the game, translates into a significant difference in overall profits. In contrast to the prediction based on standard preferences, the TAG is thus more efficient than the VCM+.

Table 3.2: Contributions of OTMs (OLS regressions)

	Dependent variable: Contributions of OTMs	
	Model 1	Model 2
TAG dummy	5.069***	-0.520
	(1.834)	(1.727)
Period	-	0.404
		(0.529)
Period²	-	-0.107**
		(0.044)
Period * TAG	-	2.096***
		(0.757)
Period² * TAG	-	-0.154**
		(0.064)
Constant	9.881***	11.786***
	(1.501)	(1.284)
# Observations	1080	1080
R^2	0.088	0.158

Notes: *** Significant at 1% level; ** significant at 5% level; * significant at 10% level. Robust standard errors in parentheses (clustered on team level).

By looking at profits in more detail (see Table 3.1, rows 4-5), distributional issues become apparent. TAs in the TAG earn about 1/3 more than OTMs. However, OTM's average profit of 26.54 lies clearly above the endowment level, indicating that profits are more balanced than one might have expected according to standard theory. On the contrary, in the VCM+ treatment OTMs earn about 1/3 more than the forced contributors who only roughly earn their endowment level. A comparison between treatments, not surprisingly, shows that TAs in the TAG earn significantly more than forced contributors in the VCM+ (indeed, twice as much) ($p < 0.01$, Mann-Whitney U-test, $N = 36$). In contrast to this, OTMs in the TAG earn significantly less than OTMs in the VCM+ ($p < 0.01$, Mann-Whitney U-test, $N = 36$). However, this does not tell the entire story. Without any voluntary cooperation, OTMs would earn 20 in the TAG and 28 in the VCM+. We therefore can compare the profits to this baseline and get the actual gains by cooperation in the different treatments. The OTMs'

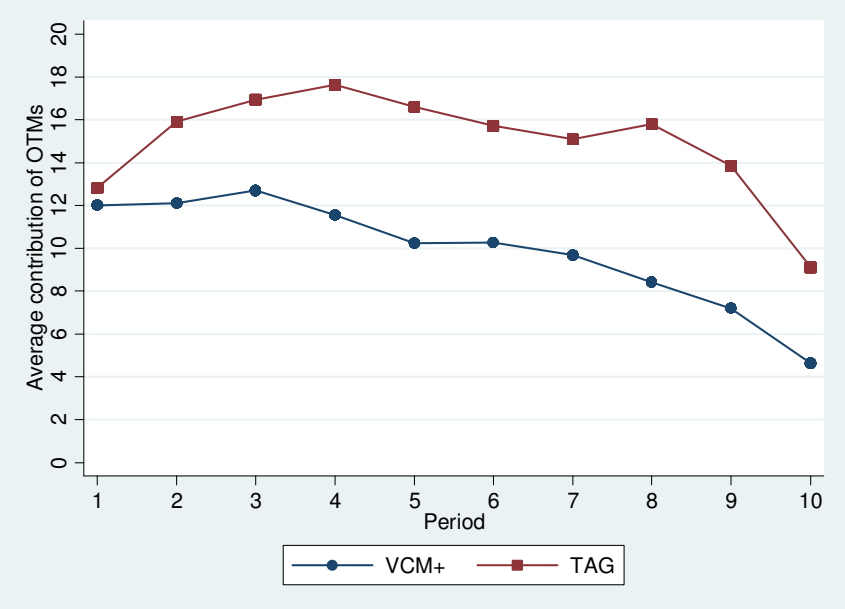
[107] Note that we present OLS rather than tobit regressions for contributions because they are more straightforward to interpret and they avoid the difficulties associated with interpreting interaction effects in nonlinear models (see Ai and Norton, 2003). Qualitatively, tobit regressions yield the same results.

gains of cooperation are, on average, 6.54 in the TAG and 1.98 in the VCM+, the difference is significant ($p < 0.01$, Mann-Whitney U-test, $N = 36$).

Result 3.1. *Mean contributions and profits are significantly higher in treatment TAG than in VCM+. In the TAG, TAs earn about 1/3 more than OTMs. OTMs' earnings lie clearly above the endowment level and the gains of cooperation are significantly higher than in VCM+.*

Figure 3.1 delineates average contributions of OTMs over time for both treatments. While average contributions are around 12 points and roughly the same in period 1 in the two treatments, from period 2 on, the difference between the two treatments becomes apparent. In the TAG, average contributions increase until they reach a level of almost 18 in period 4. Afterwards, average contributions decline slowly and remain still at a level of 14 in period 9, before they finally drop to 9 in period 10 due to a strong endgame effect. On the contrary, in the VCM+ average contributions decline almost linearly over time and end at a level of 4.5 in period 10.[108] This pattern suggests that TAs win OTMs' trust quickly by implementing appropriate first period decisions. They seem to be able to stabilize contributions on a high level in contrast to the VCM+, in which cooperation decays over time just as it is usually observed in standard VCMs.

Figure 3.1: Evolution of OTMs' average contributions across treatments



[108] A Mann-Whitney U-test shows that contribution differences in period 1 are indeed not significant ($p = 0.54$, $N = 108$). Significant differences, however, exist for periods 3-9 ($p < 0.05$, Mann-Whitney U-tests, $N = 36$).

In the regression of model 2 in Table 3.2 we not only include the variable *TAG* but add four variables capturing time trend differences between the two treatments. In addition to *Period* and *Period²* we use the two interaction terms *Period* * *TAG* and *Period²* * *TAG*. Results show that the treatment dummy becomes completely insignificant once we control for the time trend. For VCM+, we observe *Period* to be insignificant while *Period²* has a significantly negative sign. Hence, there is no evidence for a quadratic time trend in the VCM+ treatment. However, if we run a regression for VCM+ using *Period* only, this variable is highly significant ($p < 0.01$), indicating a clear decrease in cooperation over time. For the TAG, we have to consider the joint effect with the interaction terms. Note that both of them are significant. By adding up *Period* and *Period* * *TAG* as well as *Period²* and *Period²* * *TAG* we get the expected signs of a quadratic specification. Moreover, if we test for the combined effects to be zero, both null hypotheses can clearly be rejected (Wald tests, $p < 0.01$ in both cases). This confirms that treatment differences rely on differences in the time trend of contributions.

Result 3.2. *There is no significant difference in contributions between the two treatments in period 1. But from period 2 on, contributions in the two treatments develop differently. While contributions decline almost linearly in the VCM+ treatment, there is a quadratic time trend in the TAG that leads to higher contributions in all subsequent periods.*

Figure 3.2: Evolution of the number of teams with full cooperation across treatments

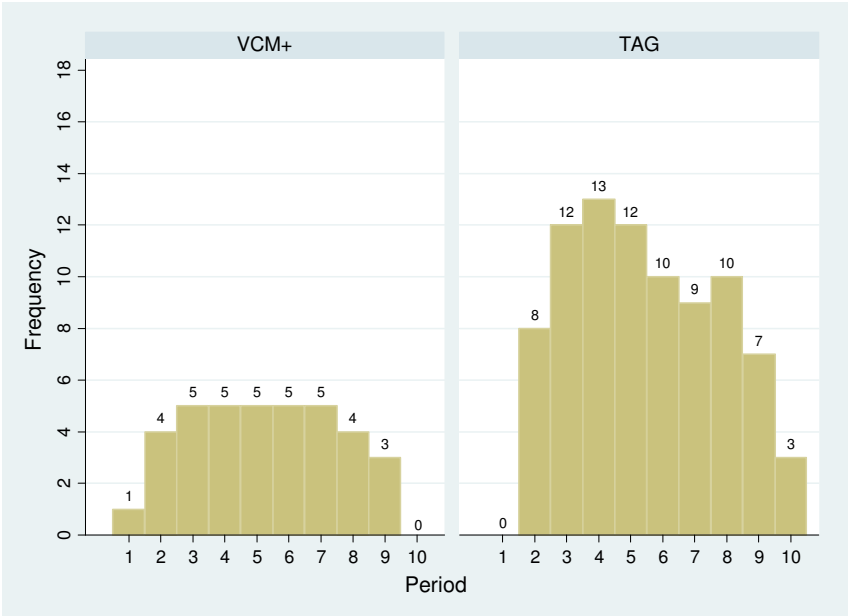


Figure 3.2 compares the number of teams with full cooperation in both treatments over time. It is clearly visible from period 2 on that the number of fully cooperating teams is roughly twice as high in the TAG than in the VCM+. This confirms the prediction from the Fehr and Schmidt (1999) model that full cooperation is easier to achieve in the TAG treatment. In fact, up to 2/3 of all teams manage to cooperate completely in intermediate periods of the TAG.

3.4.2 Explaining contribution behavior in the TAG

The OLS regressions in Table 3.3 concentrate on the TAG treatment and provide deeper insights into the dynamics that drive cooperative behavior of OTMs. All models contain only the periods 2-10, because first period contributions are not influenced by TA's decisions.[109]

In model 1 we include *Period* and *Period²* and add both individual i's contribution in the previous period $c_{i,t-1}$ (*Contribution (t-1)*) and the amount returned to i by the respective TA in the previous period $d_{i,t-1}$ (*Returned amount (t-1)*). Both lagged variables are highly significant and show a positive effect on next period's contributions. Especially, higher returned amounts, holding contributions constant, yield higher contributions in the subsequent period. This means that a more generous distribution decision by TAs increases future cooperation of OTMs. Interestingly, the quadratic time trend observed in Table 3.2 is not significant in this model and its coefficients are close to zero. We can show that this is not caused by the exclusion of period 1 as both time variables are large and highly significant ($p < 0.05$ each) in the absence of further covariates. Hence, we add model 2 in which we replace the quadratic trend by a linear one. Results reveal that contributions significantly decrease by about 0.5 points per period. Controlling for the lagged variables, therefore, turns the quadratic time trend into a linear one. This suggests that previous decisions by the TA cause the initial increase in contributions in the TAG.

Model 3 takes care of the fact that OTMs are not only informed about their own returned amount but also about the contributions and the returned amount of their team members. We control for social information by adding the lagged average contribution of the two other OTMs within the team (*Avg. contribution other OTMs (t-1)*) and the lagged average returned amount to these OTMs (*Avg. returned amount other OTMs (t-1)*). Results show that the latter variable has a significantly positive effect indicating that OTMs do not only take into account their own returned amount but also what TAs return to the other two team members.

[109] Tobit and random effects specifications yield very similar results.

From the specifications of models 1-3 we can conclude that contribution changes in the TAG depend on two main aspects: previous period's return behavior of the TA and the respective period. While the former carries information about the TA's type the latter seems to reflect a general belief about a slightly decreasing trustworthiness of TAs towards the end of the interaction, which could be explained by an anticipation of mimicking behavior of opportunistic TAs.[110]

Result 3.3. *Contributions of OTMs in the TAG depend negatively on period and positively on TA's previous period's returning behavior.*

Table 3.3: Contributions of OTMs in TAG (OLS regressions)

	Dependent variable: Contribution of OTMs in TAG, periods 2-10		
	Model 1	Model 2	Model 3
Period	0.035	-0.568***	-0.534***
	(0.663)	(0.100)	(0.088)
Period²	-0.050	-	-
	(0.057)		
Contribution (t-1)	0.205**	0.208***	0.271***
	(0.073)	(0.071)	(0.061)
Returned amount (t-1)	0.325***	0.329***	0.186***
	(0.047)	(0.046)	(0.025)
Avg. contribution other OTMs (t-1)	-	-	-0.075
			(0.109)
Avg. returned amount other OTMs (t-1)	-	-	0.239***
			(0.043)
Constant	6.531***	7.898***	5.685***
	(2.166)	(1.703)	(1.366)
# Observations	486	486	486
R^2	0.557	0.556	0.611

Notes: *** Significant at 1% level; ** significant at 5% level; * significant at 10% level. Robust standard errors in parentheses (clustered on team level).

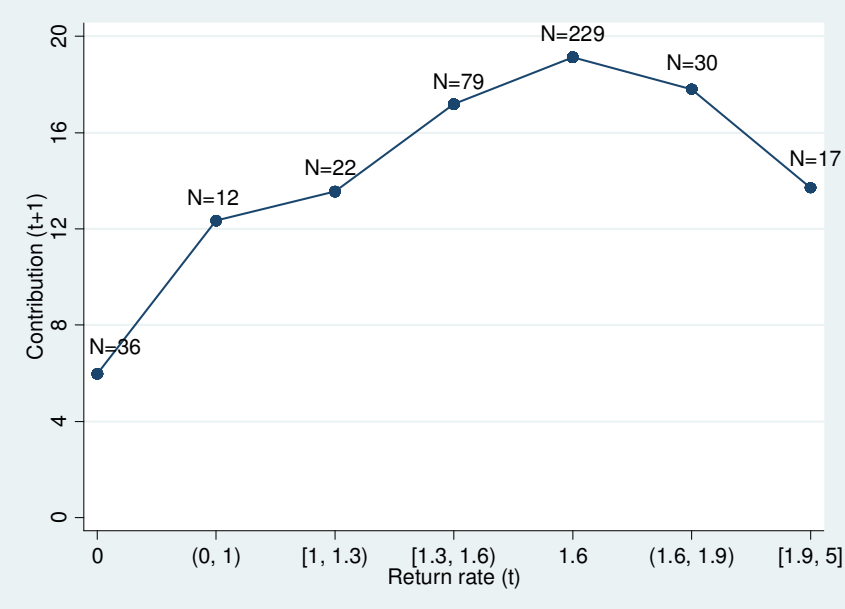
Finally, we present an alternative approach by only focusing on subjects that contribute positive amounts to the public account to be able to apply a relative measure of the benefits from an investment. Figure 3.3 shows mean contribution levels in period $t+1$ for different categories of the individual return rate by the TA in period t. Note that the individual return

[110] Note that both *Period* and the TA's previous period's return behavior stay highly significant once we include higher lags into the models. Significant effects can also be obtained by a fixed effects estimation. Finally, note that we also tried a linear, dynamic panel-data estimation method (Arellano and Bond, 1991). This method has the drawback that we cannot cluster on the team level and that we lose one further period. Nevertheless, we also find a significant influence of the two main aspects mentioned above.

rate r is defined as $r_{i,t} = d_{i,t}/c_{i,t}$. Not surprisingly, we find that OTMs contribute little in period $t+1$ if they do not get any return in the preceding period t. Increasing the returned amount to a rate of 1.6 clearly raises subsequent average contributions of OTMs and a return rate of 1.6 is nearly always followed by an OTM contributing the entire endowment. Interestingly, we find a negative effect for return rates larger than 1.6. Extraordinarily high *relative* returns, i.e. returns that exceed the amount generated by the respective investment, tend to decrease contribution levels in the subsequent period.[111] Such "over-generous" behavior seems to raise OTM's suspicion regarding the TA, but the number of observations is comparatively low.

Result 3.4. *For maximizing contributions in the subsequent period, it is the best strategy to return exactly 1.6 times the contributed amount. Higher individual return rates tend to decline contributions in the next period.*

Figure 3.3: Contributions in the next period for different categories of the individual return rate



3.4.3 Behavior of the TA and consequences for OTMs

Figure 3.4 gives a descriptive overview of the dynamics within each team in the TAG. It displays the average contribution levels of OTMs alongside with TAs' average returns to the

[111] A significant negative effect can also be shown by inserting a squared expression for the lagged returned amount in model 2 of Table 3.3. However, due to the small number of observations the decreasing effect is less robust when we introduce such a variable in estimation approaches such as fixed effects or the Arellano-Bond estimator.

team members.[112] The abbreviation on top of each panel indicates session (1-3), team (a-f) and provides information on the behavioral type of the TA according to the ring test in parentheses. The letter "I" indicates an individualistic TA, "C" represents a cooperative TA, and an "X" is displayed whenever the TA cannot be categorized by the ring test because of failing to meet the consistency standard. We have twelve individualistic, four cooperative and two non-classifiable TAs.[113]

As it can be discerned easily from Figure 3.4, no team starts with full cooperation in the TAG. OTMs seem to contribute cautiously in period 1, testing the reaction of their respective TA. However, first period behavior of TAs already leads to full cooperation in period 2 in eight out of 18 teams (44.44%). This number increases even a bit further and stays around 50-60% until period 8. While seven teams still cooperate fully in period 9, three teams even manage to cooperate fully until period 10 (see also Figure 3.2). Remarkably, only three teams (1e, 2b, 2c) never reach full cooperation in any of the ten periods.

Figure 3.4: Average contributions of and returns to OTMs in TAG over time by team

[112] Appendix 3B provides a similar graph showing average contribution levels of OTMs in the VCM+ treatment.
[113] The ring test provides evidence that there is no anti-social TA in our experiment. Hence, forcing the TA to contribute her full endowment is innocuous, in line with the discussion in Section 3.2.2 (see footnote 94).

We classify the teams in *high contribution* teams (1a, 1b, 2a, 2c, 2d, 2e, 2f, 3a, 3b, 3d, 3e), *low contribution* teams (1c, 1e, 3c, 3f), and *mixed contribution* teams (1d, 1f, 2b). *High contribution* teams are teams that show either average contribution levels of OTMs above ten (50% of the endowment) in each intermediate period 2-9 or that have a significantly increasing contribution pattern ending above 50% in period 9 (spearman rank correlation coefficient, $p < 0.05$). *Low contribution* teams are obtained by reversing the classification, while *mixed contribution* is the remaining category.[114]

Do TAs in *high contribution* teams behave differently than TAs in the other categories? If we look at average returns of TAs (consider the squared lines in Figure 3.4), it is obvious that the returned amounts in *high contribution* teams are indeed very high and a closer look at the data reveals that in almost all cases full cooperation goes along with equal profits for all team members. Six out of these eleven TAs (1a, 2a, 2d, 2e, 3b, 3e) return in each period, on average, more than the invested amount to their OTMs. Moreover, there is one TA (in team 2f) who does the same in all periods 1-9 but faces a zero average contribution of OTMs in period 10.[115] The other four TAs (1b, 2c, 3a, 3d) also return large amounts but appropriate the entire public account in the last or a late period.

On the contrary, TAs in teams with *low* or *mixed contribution* levels return either relatively low amounts in general (1c, 1e, 2b) or they return large amounts in the beginning, until full cooperation is achieved, but then take a large share of the public account for themselves already around period 5, thereby destroying cooperation in the subsequent period (1d, 1f, 3c, 3f). Interestingly, three of the latter TAs (1d, 3c, 3f) return large amounts in the following periods, presumably in order to re-increase OTMs' contributions. All three of them, finally, take the chance to appropriate a large share of the cooperation benefits a second time. Hence, TAs in *high contribution* teams allocate indeed differently than TAs in the other teams.

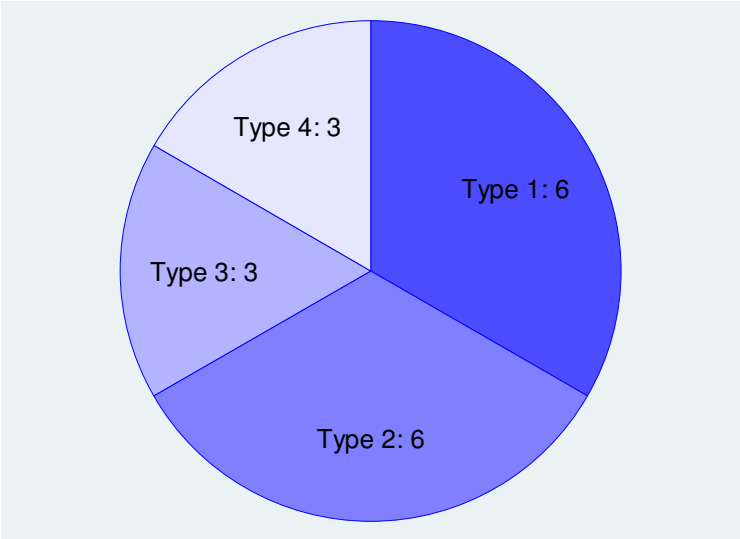
To sum up, four types of TAs appear in our data: TAs returning large amounts in each period (type 1), TAs returning large amounts except for a late period (or periods) where they take the entire public account for themselves (type 2), TAs taking a large share of the public account twice, precisely in a period around period 5 and a late period (type 3) and TAs returning small amounts in general (type 4). While the first two types generate high and stable

[114] Note that we observe only few teams in the *low contribution* category. This is different from what we usually observe in standard public goods games and also different from VCM+. Appendix 3B shows that the frequency of categories is significantly different between TAG and VCM+.

[115] Hence, we do not know whether this TA would also have returned more than the invested amount in the last period. We therefore carefully sort her into the category of type 2 in Figure 3.5 (see below), although she could also be of type 1.

levels of cooperation, the third type creates ups and downs in contribution levels and the fourth type generates a decrease in contributions, just as we usually observe in standard public goods games. Note that both type 2 and type 3 behavior is in line with mimicking strategies of selfish TAs (see our theoretic arguments in Section 3.3.2.3), while type 4 is selfish without mimicking cooperativeness. Figure 3.5 shows the distribution of types that we observe in our experiment.[116] Six of our 18 TAs belong to type 1.

Figure 3.5: Types of TAs in TAG



Looking at mean profits of TAs, the TA in team 1d performs best (an average of 50.6 points per period), followed by the TAs in teams 3c (47.5), 3d (46.8), 1f (45.6), and 1e (45.1). It is apparent that both TAs in teams 1d and 3c are of the third type. Thus, the strategy of appropriating the entire public account twice and returning large amounts in the other periods seems to be the most successful one in terms of maximizing the TA's profit. The TAs in teams 3d and 1f take the public account only once. Interestingly, the TA in team 1e, who returns almost nothing over all ten periods, earns the fifth largest amount, but nevertheless five points less, on average, than the best performing TA. However, if we compare overall team profits, team 1e clearly performs poorest with an average profit of only 24.5 points per team member. On the contrary, in the best performing teams 3b, 3d and 3e, team members earn, on average, 31.7 points per period. The latter teams have in common that their TAs are either of type 1 or 2.

[116] Note that the TA in team 1f is of type 2, although she creates only a mixed contribution pattern, according to the contribution classification of teams.

Result 3.5. *Heterogeneity between teams in terms of OTMs' contribution levels is caused by TAs' returning behavior. Most of the teams show high levels of cooperation because of the large fraction of TAs being either of type 1 or 2, i.e. returning large amounts until (almost) the end of the interaction. This is remarkable since it is not a TA's profit maximizing strategy from an ex post perspective.*

Can we explain the heterogeneity between TAs by social orientation? If we consider our ring test classification, it becomes obvious that only one of our four cooperative TAs (3b) resists the temptation of taking the entire public account until the end of the interaction. Thus, we cannot claim that being *cooperative* as a TA is a good indicator for non-exploitation of team members' trust. In addition, four of the six TAs returning more than invested in each period are classified as *individualistic*. Hence, there are individualistic TAs that are not just mimicking cooperative TAs but that become in fact trustworthy when put in the role of the TA. In line with this, the aggregated return rate for OTMs, defined as $\bar{r}_{OTM,t} = \frac{1}{3}\sum_{i=2}^{4} d_{i,t} / \frac{1}{3}\sum_{i=2}^{4} c_{i,t}$[117], is, on average, even slightly higher if the TA is individualistic than if she is cooperative (1.42, N = 112 vs. 1.35, N = 38) and this holds also for the last period. Hence, we do not find any descriptive evidence for a more trustworthy behavior of cooperative TAs. This result is confirmed by a cluster-robust OLS regression explaining a TA's relative appropriation of the public account by her social motivation as it yields insignificant results for the ring test dummy, both over all periods and for the last period.[118] This confirms that social orientation, surprisingly, does not matter for TA's decisions.

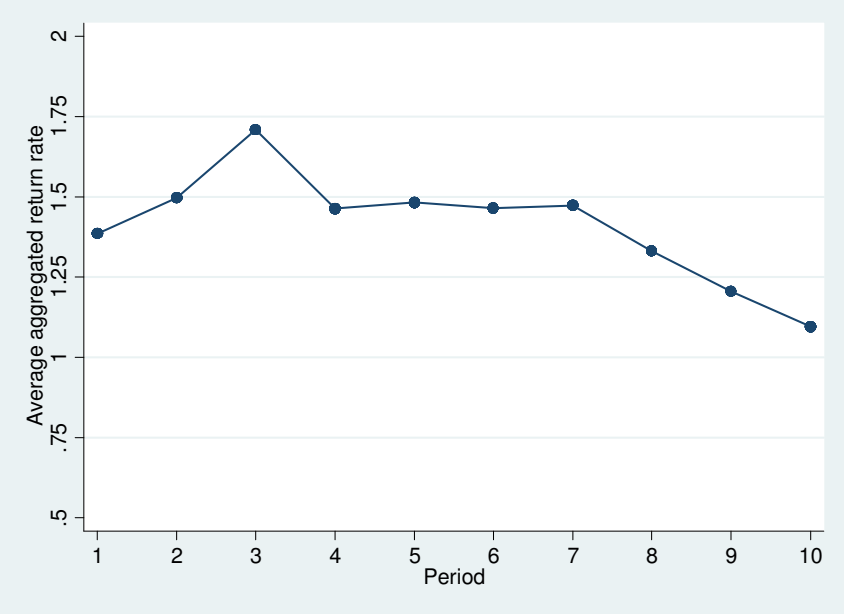
Result 3.6. *Surprisingly, heterogeneity in TA's distribution behavior cannot be explained by social orientation. Many individualistic TAs behave cooperatively even in the last period once they are responsible for the distribution of the public account.*

Overall, the average aggregated return rate is 1.42. This value is astonishingly high compared to predictions based on standard theory. Figure 3.6 shows that the mean of the aggregated return rate varies over time and, especially, decreases in the last three periods. However, even in period 10, it is slightly above one, indicating that TAs return more, on

[117] Note that the aggregated return rate is a weighted average of OTMs' individual return rates with the weight being a single OTM's relative contribution, i.e. $c_i / \sum_{i=2}^{4} c_i$.
[118] This holds irrespective of whether we control for the size of the public account or not. Using the exact angles of the vectors as obtained out of the ring test does not change the result either.

average, than OTMs contribute. Combining Figure 3.6 with the regression results in Tables 3.2 and 3.3, we can finally explain why we find a significant quadratic time trend in contributions in the TAG: While the increase in cooperation in the first half of the ten periods is caused by TAs' high and even increasing return rates, the decrease is due to a decrease in aggregated return rates from period 8 on (see Figure 3.6) *and* due to OTMs' beliefs about a decreasing trustworthiness of TAs (see the significant influence of *Period* in Table 3.3).

Figure 3.6: Evolution of the average aggregated return rate



If we look at the distribution of aggregated return rates, see row 1 of Table 3.4[119], it becomes obvious that TAs return, on average, exactly 1.6 times the contributed amount in more than 50% of the cases. Moreover, in about 30% of the cases, TAs implicitly *reward* OTMs for contributing but do not return the full benefit generated by OTMs' contributions. Interestingly, it is not predominantly the case within this category that TAs return only the investment plus an increment to barely motivate contributions in the subsequent period. Indeed, aggregated return rates between 1 and 1.3 are rare and appear only eight times. More frequently, in 13 cases, we observe aggregated return rates that are even larger than 1.6, implying that TAs sacrifice parts of their own share. Overall, positive reciprocity is obtained in 151 out of the 170 cases (88.8%) where OTMs, on average, contribute positive amounts to the public account. Furthermore, the level of implicit reward is quite substantial. Implicit *punishment*, on the contrary, is seen rarely. Partial punishment is almost not existent and there

[119] Row 1 excludes ten cases in which no OTM in the team contributes positive amounts because the aggregated return rate is not defined. Not surprisingly, TAs do usually not return positive amounts to OTMs in such a case.

are only 17 cases in which TAs take the entire share of the public account if mean contributions of OTMs are positive (plus nine cases in which the mean contribution is zero).

Concerning the distribution of profits within a team, it is noticeable that in 62 out of the 170 cases (36.47%) full payoff equalization across all team members (including the TA) is achieved. All of these cases exhibit full cooperation and a return rate of 1.6 to everybody. Remember that full payoff equalization was the central prediction for certain parameter values of a TA's utility function ($\beta_1 \geq 0.75$, $\lambda_1 \geq 0.75$) in both the Fehr and Schmidt (1999) and the Charness and Rabin (2002) model. However, if contributions between team members differ, full payoff equalization is not observed.[120] TAs overwhelmingly ensure by their returning behavior that low contributors earn less than high contributors. This happens either by returning the same individual return rate to each OTM or by raising the return rate with higher OTMs' contributions. Thus, TAs seem to follow a norm of effort-based inequity or maximin orientation once contributions differ, as the two theories predict implicitly.

Result 3.7. *The average aggregated return rate is relatively high but decreases from period 8 onwards. Thus, there are two reasons for the decrease in cooperation in the second half of the experiment: Subjects' beliefs about a reduced trustworthiness of TAs and an actual decrease in the aggregated return rate.*

Table 3.4: Frequency of return rates to OTMs on aggregated and individual level

	rate = 0	0 < rate < 1	1 ≤ rate < 1.6	rate = 1.6	rate > 1.6	*Sum*
	Full Punishment	*Partial Punishment*	*Partial Reward*	*Full Reward*	*Excessive Reward*	
aggregated return rate	17	2	48	90	13	*170*
individual return rate	45	13	103	244	49	*454*

Notes: The aggregated return rate is defined as $\bar{r}_{OTM,t} = \frac{1}{3}\sum_{i=2}^{4} d_{i,t} / \frac{1}{3}\sum_{i=2}^{4} c_{i,t}$, the individual return rate as $r_{i,t} = d_{i,t}/c_{i,t}$.

Row 2 of Table 3.4 presents the individual return rates to OTMs in the TAG (defined as $r_{i,t} = d_{i,t}/c_{i,t}$). Again, we focus here on observations in which OTMs contribute positive amounts. In addition, there are 86 cases in which the contribution of an OTM is zero.

[120] Note that in case of unequal contributions full payoff equalization would sometimes require the usage of two decimal places for the returned amount. However, there is no single observation in our data in which payoffs are equalized except for a remainder which cannot be split equally (having size 0.2). Moreover, this design issue should not matter in later periods as an almost equalization of profits already generates strong contribution incentives and, hence, full cooperation should appear in subsequent periods.

However, in almost all of these cases (actually, 78), a zero contribution results in a zero return by the TA.

In contrast, OTMs contributing positive amounts are predominantly faced by TAs rewarding their contribution behavior. In 293 cases (64.65%) the respective OTM receives the whole benefit generated by her investment or even more than that. Furthermore, if one adds the 103 observations which are below 1.6 but above or equal to 1, it turns out that the investment is profitable in about 87% of the cases. Hence, OTMs manage to benefit from their investments in almost every case. If we additionally account for the cases in which OTMs contribute zero and assume that this mistrust is justified then we find OTMs to benefit from an investment into the public account in 73% of cases. Thus, in the large majority of cases, it pays off for OTMs to contribute to the public account.

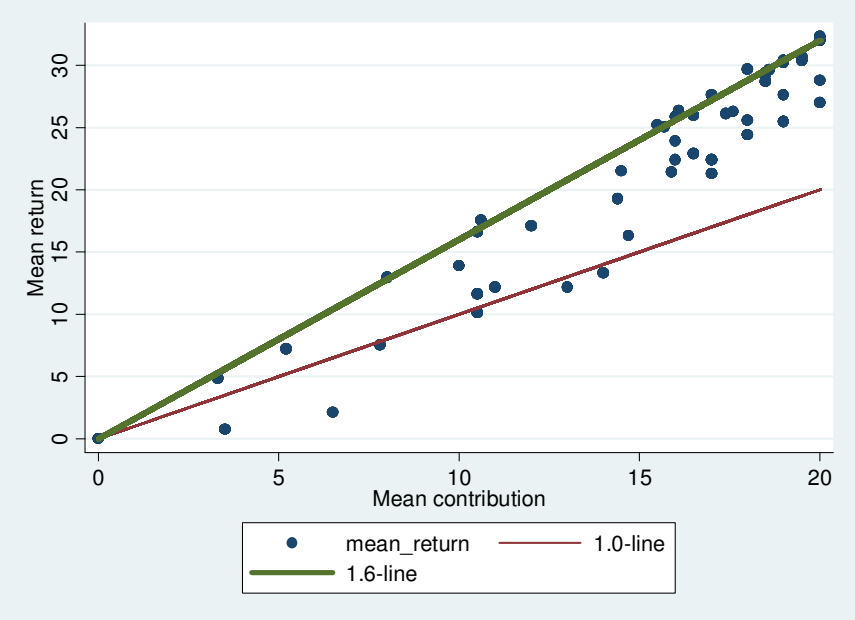
If we concentrate only on teams in which *at the same time* some OTMs contribute zero and others positive amounts we find the same picture again. TAs exclude free-riders from the benefits from cooperation since in 51 out of 56 cases they get a zero return. Thus, OTMs contributing zero cannot free-ride on their team members' contributions. At the same time, TAs implicitly reward the contributing OTMs by allocating to them a share larger than their investment (true in 82.09% of cases, N = 67). Hence, the *same* TAs use implicit punishment and implicit reward simultaneously when faced with different contributions.

Result 3.8. *In most of the cases, trust pays off for OTMs in the TAG. While non-contributors are excluded from the benefits from the public account, contributing positive amounts is profitable.*

Finally, Figure 3.7 shows a scatter plot comparing mean returns and mean contributions for each OTM in the TAG over all ten periods. Two reference lines are added. The *1.0-line* captures points where the mean return equals mean contribution. Subjects on this line therefore receive on average exactly the amount they invested into the public account. In contrast, the *1.6-line* consists of all points where subjects, on average, receive 1.6 times their invested amount. In particular, if a TA returns to an OTM the complete amount generated by the respective contribution in all periods, the resulting dot will lie on the *1.6-line*. The figure shows that only two observations lie clearly below the *1.0-line*. For these OTMs it did not pay to trust the TA. Moreover, there a few points on or near to the *1.0-line*. Those subjects do not profit from their investments into the public account but they do not lose substantial amounts either. Interestingly, there is only a single observation in the origin of the graph, i.e. complete

free riding is a very rare event in the TAG. In general, most of the points lie in the upper right corner near, on, or even slightly above the *1.6-line*. This shows again that over the whole course of the experiment almost all OTMs manage to gain large benefits from their investments into the public account and hence contribute large amounts.

Figure 3.7: Mean return and mean contribution for each OTM in the TAG



Result 3.9. *Over the course of the experiment, nearly all OTMs manage to profit from investing into the public account. As a large fraction of TAs returns large amounts, many OTMs contribute, on average, almost their full endowment level.*

3.5 Discussion and conclusion

We have analyzed a modified public goods game and implemented it experimentally in the laboratory. In our *team allocator game*, each team member – regardless of whether the member is an ordinary team member or the team allocator – can contribute to a public account. The sum of contributions is multiplied by an efficiency factor larger than one, but – in contrast to the standard public goods game – the public account is not distributed equally among all team members. Rather, the team allocator receives the entire amount and has full discretionary power over the allocation of the revenues from the account within the team.

More precisely, she can implement any distribution of the benefits from the public account over the ordinary team members and herself.

We provide three main empirical results: First, in contrast to theoretical predictions from standard preferences, we find that the level of contributions in the team allocator game is significantly higher than in an appropriate control treatment in which there is no team allocator, but one team member is forced to contribute her entire endowment. Second, we find that it is the team allocator's distribution behavior that influences together with the time horizon of the team interaction the development of contributions. Contributions increase in the returned amount, i.e. the reward channel is most effective in sustaining high levels of cooperation. Third, although there is some heterogeneity among the team allocators, on average, team allocators return remarkably high amounts to ordinary team members that invest into the public account. Non-contributors, however, are excluded from the benefits from cooperation. Hence, team allocators generate strong contribution incentives. Our results clearly refute predictions based on standard preferences. They are, however, largely in line with models of heterogeneous preferences and repeated interactions such as (effort-based) inequity aversion (Fehr and Schmidt, 1999) or a maximin-preference (Charness and Rabin, 2002).

The general implication of our results is that teams with a straightforward hierarchy can have an advantage over teams with equal members. They are more likely to overcome the social dilemma inherent to public good provision, i.e. team effort provision. This is the more remarkable since the described mechanism is costless: Implicit reward (and, to a much lesser degree, implicit punishment) works through the allocation process and does not bear any monetary costs such as formal or informal sanctions that have been studied widely. Allocation power in teams can, thus, be considered as a potential alternative of a sanctioning regime, because the latter is often much more efficiency-damaging.

However, it is difficult to predict how easily such a mechanism can really be implemented in a social dilemma environment. Thus, a natural extension of our setup is to implement an endogenous treatment in which subjects can vote on whether they want to have a team allocator or not. Another obvious extension would be to let subjects elect their team allocator. Recent literature on the impact of elected vs. randomly chosen leaders (e.g. Baldassarri and Grossman, 2011; Levy et al., 2011) suggests that legitimate authorities enhance team cooperation. As many real-life situations involve voting decisions on group leaders, our experiment most probably underestimates the true gain of endogenously formed hierarchies.

Appendix

3A Experimental instructions (originally in German)[121]

A warm welcome to an experiment on decision making!
Thank you for participating!

During the experiment you and all other participants will be asked to make decisions. Your decisions as well as the decisions of the participants you are matched with determine your earnings from the experiment according to the following rules.

Please stop talking to other participants from now on. If you have any questions after going through the instructions or while the experiment is taking place, please raise your hand, and one of the experimenters will come to you and answer your questions privately. In case the question is relevant for all participants, its answer is repeated aloud.

The whole experiment is computerized and will last approximately **90 minutes**. All your decisions and answers remain anonymous. You will not find out with whom you are matched in each of the experiment's parts and how much each of the other participants earns. We evaluate data from the experiment on aggregate level only and never link names to data from the experiment. At the end of the experiment, you will be asked to sign a receipt for your earnings. This has accounting purposes only.

The experiment consists of **two** parts. At the beginning of each part, you will receive the corresponding instructions for this part. The instructions will be read out loud and you will get time to ask questions. Please, do not hesitate to ask if anything is unclear to you. Your decisions in Part I of the experiment **do not** have any effects on Part II. In the interest of clarity, we will only use male terms in the instructions. They should be interpreted as being gender-neutral. For means of help, you will find a pen on your table.

While taking your decisions at the PC, there will be a clock counting down in the right upper corner of the screen. The clock serves as a guide for how much time you should need. You may exceed the time. The input screens will **not** be turned off when time has run out. However, the information screens on which no decision is required to be taken will be turned off when time has run out. Once you have taken a decision or have read through a screen, please confirm by clicking on the "OK" button.

Your earnings in the experiment will be calculated in **"points"**. At the end of the experiment, the "points" get converted into euro at the exchange rate announced in the respective part. In addition, you receive 4 euro for your arrival on time. Your total earnings from the experiment will be paid out to you privately and in cash at the end of the experiment.

[121] Baseline instructions describe treatment TAG. Differences in VCM+ are indicated by [**VCM+**].

Part I

In Part I of the experiment all participants are randomly assigned into groups of two. Nobody will find out with whom he forms a group – not during the experiment and not after the experiment either.
You have to take 24 decisions in this part of the experiment. In each decision you can choose between 2 options, A and B. Each option allocates a positive or negative payoff (earning) in points to you and to the other person in your group. The other person answers exactly the same questions. Your total payoff from Part I depends on your decisions *and* on the decisions taken by the other person in your group.

A decision example:

	Option A	Option B
Your payoff	10.00	7.00
Other's payoff	-5.00	4.00

- If you choose Option A you receive 10 points, and the other person loses 5 points. If the other person also chooses Option A, he, too, receives 10 points and you lose 5 points. In total, you therefore earn 5 points (10 points from your choice minus 5 points from the other person's choice). The other person earns 5 points (10 points – 5 points), too.
- In case you choose Option B and the other person chooses Option A, you earn 2 points (7 points from your choice minus 5 points from the other person's decision). The other person earns 14 points (10 points + 4 points).
- The remaining combinations (you choose A and the other person chooses B, or both persons choose B) are analogous to these two examples.

Overall you take 24 decisions like the one described above. Your total payoff is computed as follows: The 24 values for "your payoff" are summed up over your decisions. The 24 values for "other's payoff" are summed up over the other person's decisions. The sum of these two sums determines your total payoff from this part and is converted into euro at the end of the experiment as follows: **25 points = 3 euro** (1 point = 12 cent). This exchange rate is valid only for Part I of the experiment.

Note that you are not receiving information on each single decision taken by the other person in your group. Rather, you will find out only the sum of your decisions for "your payoff", the sum of the other person's decisions for "other's payoff" and your total payoff from Part I at the very end of the experiment. Note that you do not get any feedback immediately after Part I.

If there are any questions, please raise your hand now. We will come to you and answer your questions privately.

Part II

The points earned in Part II are converted into euro at the exchange rate of **25 points = 1 euro** (1 point = 4 cent) at the end of the experiment.

At the beginning of Part II, all participants are randomly assigned into groups of four. Nobody will find out with whom he forms a group – not during the experiment and not after the experiment either. Part II consists of **10 identical periods** and you remain matched with the **same persons throughout the entire Part II**.

Each participant is randomly given an **individual name** which, too, remains the same across all 10 periods, and which allows you to keep track of the behavior of your group members throughout the periods. The names are: Person 1, Person 2, Person 3 and Person 4.
Furthermore, a **member type** is assigned to each group member (A or B). Within each group there is one group member of type A and three group members of type B. The group members of type A and B differ in their decision possibilities. The type of each group member is publicly announced within the group and remains the same throughout the 10 periods.

The group member of type A is **randomly determined**. The probability of being of type A is 25 % for each group member. The remaining three group members are of type B.

Endowment and alternatives in each period
Each period consists of two stages, a **contribution stage** and a **distribution stage**

Contribution stage:

Each participant receives an initial endowment of **20 points** at the beginning of the contribution stage in each period. The 20 points are allocated to two alternatives, a group account and a private account, depending on the participant's type:

The group member of type A **is obliged** to put all of the 20 points into the group account. Thus, the group member of type A takes no decision during the contribution stage.
Group members of type B can **freely** choose how many points to contribute to the group account and how many points to contribute to the private account.

The group account:
Contributions to the group account from all group members are summed up. The sum is multiplied with 1.6 and distributed among the group members during the distribution stage (s.b.). For example, if the sum of all contributed points to the group account is 60, there are 60*1.6=96 points from the group account to be distributed to the group members in the distribution stage. If the sum of contributed points to the group account is 20, there are 20*1.6=32 points from the group account to be distributed in the distribution stage.

The private account:

The contribution of a group member to the private account turns solely and one-to-one into direct earning of the respective individual. For example, if a group member puts 6 points into the private account, he receives exactly 6 points from the private account to his earnings. If the contribution to the private account is 17, the group member earns exactly 17 points from the private account. The other group members do not receive anything in each case.

Distribution stage:

During the distribution stage, the group account gets divided among the four group members.

The group member of **type A** is in charge of the division**.** He distributes the group account among himself and the other group members. Group members of type B do not have any influence. Values with **at maximum one decimal place** are allowed for the distribution (please use a dot to separate digits).

[**VCM+**: The distribution is done **automatically**. Each group member receives 25% of the group account.]

The following table is exemplary and shows several distributions for the case that there are 60 points to be distributed. The first three distribution settings are possible. The fourth one is not possible as there are too few points (29) that are distributed. The fifth setting is not possible either as there are too many points (120) that are distributed.

	Distribution 1	Distribution 2	Distribution 3	Distribution 4	Distribution 5
Person 1	12.6	0	15	5	45
Person 2	10	0	15	8	15
Person 3	21	60	15	2	15
Person 4	16.4	0	15	14	45
	Possible	Possible	Possible	Too few points	Too many points

Naturally, the actual distribution chosen by the group member of type A can look completely different to the exemplary distributions 1–3. Any combination of numbers that adds up to the sum to be distributed is possible.

[**VCM+**: The following table is exemplary and shows the distribution for the case that there are 60 points to be distributed.

	Distribution
Person 1	15
Person 2	15
Person 3	15
Person 4	15

]

Earnings in one period:

Your earnings per period are the sum of the amount of your private account and the amount allocated to you from the group account.

Procedure:

On the first screen you get told about your individual name (Person 1, Person 2, Person 3 or Person 4) and which Person is of type A. The other group members are automatically of type B. Afterwards, all group members of type B get asked about how much of the 20 points they would like to contribute to the group account. The remainder is automatically allocated to the private account. Saving points for later periods is thus not possible. Only integer numbers between 0 and 20 (whereby 0 and 20 are possible choices, too) can be entered. The group member of type A is obliged to contribute 20 points to the group account and, consequently, does not get an input screen.
Afterwards, all group members get informed about contributions to the group account of all group members and the resulting sum to be distributed.
The group member of type A is then asked how he wants to divide the group account among the group members. The Windows Calculator can be used to help with calculations. It can be found by clicking on the calculator symbol on the screen.
[**VCM+**: Thereafter, the group account is divided among the group members.]
At the end of the period, all group members are informed about the contributions to the group account, the allocation from the group account, the contributions to the private account as well as the earnings of all group members in this period. Subsequently, the next period starts.

This part of the experiment is finished after 10 periods. The results from all periods are summed up and converted into euro.

Afterwards we will ask you to fill in a short questionnaire on the PC. The questions on individual persons relate to the names of Part II. There are reply options given for most of the questions. Free text entry is required by some questions. For free text entry questions, please write your answers in the corresponding blue text box on the PC screen, and confirm your entry by clicking the enter button. Your text will then appear above the blue text box.

You get told your feedback from Part I after you have filled in the questionnaire. After that, payment of your total earnings in the experiment takes place.

If there are any questions, please raise your hand now. We will come to you and answer your questions privately.

3B Further results

Figure 3B.1: Average contributions of OTMs in VCM+ over time by team

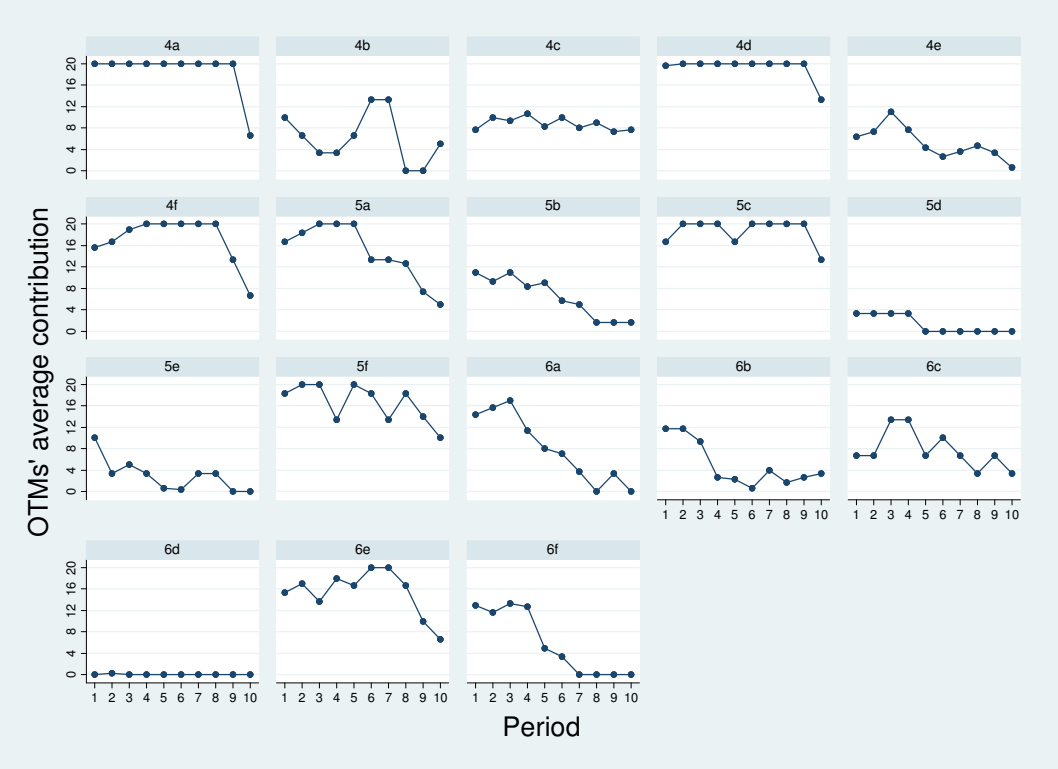


Figure 3B.1 shows the analogous to Figure 3.4 for the VCM+ treatment (sessions 4-6). Team 4a characterizes team a in session 4, etc. In this treatment, only five teams (4a, 4d, 4f, 5c, 5f) can be classified as *high contribution* teams. As usual in standard public goods games, the *low contribution* teams dominate. 9 out of the 18 teams (4e, 5a, 5b, 5d, 5e, 6a, 6b, 6d, 6f) fall into this category. The four remaining teams (4b, 4c, 6c, 6e) form the *mixed contribution* category. Table 3B.1 shows frequency of categories for the VCM+ and the TAG treatment (numbers for TAG as described in Section 3.4.3). Frequencies in the first two columns are significantly different using a χ^2 test ($p < 0.05$).[122]

Table 3B.1: Frequency of teams by category and treatment

	High contribution	Low contribution	Mixed contribution
TAG	11	4	3
VCM+	5	9	4
H0: No difference between *high contribution* and *low contribution* (χ^2 test (p-value))	< 0.05		

[122] A Fisher's exact test yields p = 0.06.

Chapter 4

The Role of Beliefs, Trust, and Risk in Contributions to a Public Good[123]

4.1 Introduction

Many situations in our daily lives possess properties of a public good (or a social dilemma). Examples range from teamwork over paying taxes or voting to the use of common goods. Being non-rival and non-excludable, public goods are plagued by free-riding problems in theory. However, a majority of involved agents do not free-ride in the provision of a public good, even when it is a dominant strategy for a rational and selfish individual to do so. Both in the field and in the experimental laboratory, we observe considerable heterogeneity with respect to cooperative behavior across individuals. At present, little is known about the determinants that shape cooperative behavior in social dilemmas and especially about the connection of cooperative behavior with obviously related behavioral tendencies such as trust or risk attitudes. In this chapter, we investigate the driving forces behind cooperation by using a standard public goods game in the experimental laboratory. More specifically, as far as we are aware, we are the first to provide a complete and entirely incentivized anatomy of the association between cooperative behavior, beliefs about others' cooperative behavior, trusting behavior, and decision-making under risk; in our experiment all decisions are incentivized.

Each of these links has been studied before in the experimental literature separately and subsets of them in combination. Recently, Thöni et al. (2009) investigated both self-reported trust and beliefs about others' contributions in an experiment among the Danish population and found that self-reported trust explains incentivized cooperative behavior to a significant extent. There is further evidence by Leonard et al. (2010), based solely on self-reported contribution behavior, that shows an association between trust and cooperation. Even earlier, Gächter et al. (2004) found a positive and significant effect of self-reported trust questions related to beliefs about people's fairness, helpfulness and trust in strangers on contributions in

[123] This chapter is joint work with Martin Kocher, Peter Martinsson and Conny Wollbrant.

a public goods experiment. However, they found no significant effect of the stated trust question on the actual trusting behavior.[124] On the other hand, Anderson et al. (2004) provided mixed evidence regarding the correlation between cooperation and self-reported trust in different domains of trust.[125]

For an individual with non-selfish preferences, contributing to a public good without knowing how much other group members are going to contribute can be viewed as a risky decision. Therefore, risk preferences might influence contributions to a public good. More risk-averse individuals with non-selfish preferences might choose to contribute less to the public good to compensate for the risk of others not contributing. However, this type of risk is a *social risk* (the risk that originates from the decisions of other human beings) rather than a *natural risk* (a random event, independent of human decisions). There is some evidence that humans perceive those risks differently but that attitudes towards social risk and natural risk are correlated (e.g., Bohnet et al., 2008) and we discuss this in greater detail in the conclusion of the chapter. Most existing studies relating risk and contributions to public goods use a measure of natural risk. In line with the notion that risk affects contributions to a public good, Charness and Villeval (2009), for instance, found that subjects who invested more in a risky asset contributed more to a public good. A similar result has been reported by Sabater-Grande and Georgantzis (2002), based on a multi-period prisoner's dilemma game.

Risk may also indirectly influence contributions as indicated by a few recent experiments that have focused on whether trust itself is determined by natural risk preferences. However, existing experimental results on the association between trust and natural risk are inconclusive. Whereas Schechter (2007) found a correlation of individual behavior in a trust game and a risk experiment in rural Paraguay, Bahry and Wilson (2004) and Eckel and Wilson (2004) did not find any relationship between elicited risk attitudes and the amount sent in a trust game.

Our experiment consists of three main parts that measure (i) cooperative behavior and beliefs about others' contributions to the public goods, (ii) (natural) risk preferences, and (iii) trusting behavior. All measures are observed on the individual level, and they are all

[124] It is a debated issue whether trust elicited in a trust experiment correlates with trust reported in surveys. Glaeser et al. (2000) compared results from trust experiments and stated trust, and found poor correlations between the amounts sent in the trust experiment and stated trust. They concluded "that most work using these survey questions needs to be somewhat reinterpreted" (p. 814). On the other hand, for example, Fehr et al. (2002), Bellemare and Kröger (2007) and Johansson-Stenman et al. (2011) found a significantly positive relationship.

[125] The relationship between trust and cooperation has been discussed in other fields for decades (e.g., Deutsch, 1958; Dawes, 1980). More recently and more generally, economists have established the importance of trust especially for economic growth (e.g., Knack and Keefer, 1997; Knack, 2002). Whereas the literature includes many definitions of social capital (see, e.g., the overview in Durlauf and Fafchamps, 2005), several emphasize trust as a key component (e.g., Bowles and Gintis, 2002; Putnam, 2000).

incentivized monetarily. Cooperative behavior is elicited by using a one-shot public goods experiment, which has the advantage of eliminating strategic motives. We apply the experimental design introduced by Fischbacher et al. (2001) based on the strategy method. Besides measuring unconditional cooperation, it elicits conditional contributions, i.e., how much a subject wants to contribute conditional on all possible average integer contributions by the other members in the subject's group, in an incentive-compatible way, and it allows us to classify cooperation types. The two predominant types are free-riders and conditional cooperators. We also included an incentivized question on beliefs about others' average contributions. Moreover, our subjects participate in a risk experiment using the same design as in Holt and Laury (2002) to elicit their attitudes toward natural risk and in a trust game similar to the design used by Berg et al. (1995) to elicit trusting behavior.

The results of our study indicate that beliefs about others' contributions and trust as elicited by the trust game are significantly associated with cooperative behavior, whereas (natural) risk preferences neither affect contributions nor trusting behavior significantly. Whereas the former association is not unexpected, given existing empirical results, the lack of association between natural risk and cooperation as well as the lack of association between natural risk and trust are noteworthy. Our measure of natural risk seems unable to explain any kind of social risk neither in the public goods game nor in the trust game. Furthermore, from the combination of individual contributions to the public good, individual beliefs on others' contributions and individual conditional contributions to the public good, we are able to infer that subjects do not perceive the contribution to a public good as risky.

The remainder of the chapter is organized as follows. Section 4.2 formalizes the way we think about the connection between cooperation, trust and risk attitudes. In Section 4.3 we describe the design of our experiment. The following section presents the results, and, finally, Section 4.5 concludes the chapter.

4.2 A model of cooperation and risk

While the main aim of this chapter is to establish empirical relationships, it is useful to formalize the connection between risk and cooperation.[126] Our simple model gives two immediate predictions that will be important for the empirical part. It predicts that pro-

[126] We do not model the relationship between risk preferences and trust separately because it seems intuitively obvious (see Ben-Ner and Putterman, 2001).

socially motivated agents will decrease their contributions as risk aversion increases, while risk aversion has no effect on contributions of agents who hold purely selfish motivations.

Let us assume that individual i faces the following utility maximization problem:

$$\max_{c_i} U = u(\pi_i) - \beta_i c(|c_i - \bar{c}|) \tag{4.1}$$

The first part of the utility function measures the individual's preferences over own material payoffs, π_i, with the common properties $u'(\pi_i) > 0$ and $u''(\pi_i) \leq 0$. That is to say, utility is concave in material payoffs so that the functional form can be interpreted in terms of the individual being either risk neutral or risk averse. The second part of the utility function captures non-selfish preferences of the individual. More specifically, with $\beta_i > 0$ she suffers disutility when her own contribution to the public good, c_i, deviates from the (expected) average contribution of others in her reference group, denoted $\bar{c}$. To model this, we assume a cost function that increases in the absolute difference between the agent's own contribution and the average contribution of others, $c = c(|c_i - \bar{c}|)$. We further assume that the function is convex in this argument, i.e., $c'(|c_i - \bar{c}|) > 0$ and $c''(|c_i - \bar{c}|) \geq 0$, with the additional property that $c(0) = 0$ to capture the fact that cost is zero whenever the agent's own contribution equals the (expected) average contribution of others ($c_i = \bar{c}$). The magnitude of the parameter $\beta_i \geq 0$ measures the relative importance of utility from monetary payoffs and disutility from deviating from the (expected) average contribution of others.[127] Maximizing the utility function with respect to the agent's contribution yields the following result (proof see Appendix 4A).

Proposition 4.1. *Given $\beta_i > 0$, (conditional) contributions decrease in higher levels of risk aversion.*

It is easy to see that modeling a link between cooperation and risk in a one-shot game only makes sense if one assumes some sort of other-regarding preferences. Selfish individuals have a dominant strategy to contribute zero; hence, they would never contribute for any level of risk aversion. Another point is noteworthy: Proposition 4.1 holds for conditional contributions and unconditional contributions (given expectations about others' contributions). It is, however, much more suitable in the context of unconditional

[127] Obviously, our model approach is related to models of inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000).

contributions. Conditional contributions – at least in the way we implement them as a one-shot public goods experiment based on the strategy method – are completely risk free. Still, the curvature of the utility function is sufficient to provide the result also for a decision environment not characterized by uncertainty.

Intuitively, when deciding on unconditional contributions, a subject does not precisely know how much other members of her group will contribute, and so her contribution decision is characterized by uncertainty. More risk-averse subjects will then require more in compensation for this uncertainty and will thus exhibit more selfish behavior resulting in lower contribution to the public good. In the absence of such uncertainty, as in the case of conditional contributions, the result arises because of a more risk-averse subject having a more concave utility function for utility from monetary payoffs. In striking the balance between social and purely monetary motivations, a more risk-averse subject will weight her monetary payoff more than a less risk averse subject. Thus, even if the conditional contribution decision is not subject to risk, risk preferences will still matter for contributions.[128]

4.3 Experimental design

Our experimental design consists of three different parts conducted in the following order: (i) a one-shot linear public goods experiment with the strategy vector method as well as an elicitation of beliefs on others' contributions, (ii) a risk attitude elicitation experiment, and (iii) a trust experiment. The decisions in all parts were monetarily rewarded, and it was clearly stated that the parts were independent of each other. Feedback was only given at the end of the whole experiment to avoid any cross-contamination of parts. The experimental instructions that we used can be found in Appendix 4B. All of the procedures described in the following were common knowledge among all participants.

[128] Although it might be more appropriate in other circumstances to consider concavity of the full utility function rather than concavity of the utility from monetary payoffs, our approach is primarily empirical as the risk measure we collect is derived based on individual gambles over monetary payoffs. One could easily incorporate concavity of U while retaining the result from Proposition 4.1.

4.3.1 One-shot public goods game

We used the one-shot public goods experiment based on the strategy method as developed by Fischbacher et al. (2001). In the experiment, the following linear payoff function for subject i was used:

$$\pi_i = 20 - c_i + 0.4 \sum_{j=1}^{4} c_j \tag{4.2}$$

Here, c_i denotes the contribution of subject i to the public good. Each group consists of four randomly matched subjects, and each subject receives an endowment of 20 experimental points. Each experimental point was exchanged for 0.33 euro. The marginal per capita return (MPCR) from investing in the public good is 0.4. Assuming that participants are rational and selfish, it is obvious that any $MPCR < 1$ yields a dominant strategy for every group member to free-ride, i.e., to contribute nothing to the public good. From a social perspective, it is optimal to contribute the whole endowment because $MPCR \cdot n > 1$, where n is the number of group members.

The details of the preference elicitation and the incentive mechanism in our experiment follow Fischbacher et al. (2001). Subjects are asked to make two decisions: first, an unconditional contribution to the public good, and thereafter a conditional contribution (a contribution schedule). The unconditional contribution is a single integer number that satisfies $0 \leq c_i \leq 20$. For the conditional contributions, subjects have to indicate how much they would contribute to the public good for any possible average contribution of the three other players within their group (rounded to integers). For each of the 21 possible averages ranging from 0 to 20 points, subjects must decide on a contribution between and including 0 and 20 (strategy method).

To ensure incentive compatibility, both the unconditional as well as the conditional contribution are potentially payoff relevant. For one group member in each group, who is randomly determined by the roll of a four-sided die,[129] the conditional contribution is relevant, whereas the unconditional contributions are relevant for the other three group members. More specifically, the three unconditional contributions within a group and the corresponding conditional contribution (for the specific average of the three unconditional contributions) determine the sum of money contributed to the public good. Individual earnings can then be calculated according to equation (4.2). Note that the strategy vector

[129] Each group member is assigned a number from one to four. The die is rolled by a randomly selected participant in the session and monitored by the experimenter.

protocol makes the standard simultaneous public goods game more akin to a sequential game and, hence, even more closely related to the trust game.

Furthermore, subjects were asked to guess the average unconditional contribution of the other three group members (rounded to integers). The guessing stage is implemented after the contribution stages and was not mentioned in the instructions. As in Gächter and Renner (2010), subjects were monetarily rewarded depending on the accuracy of their guesses. However, we use a slightly different incentive mechanism. If a subject's guess equals exactly the average unconditional contribution of the other three group members, the subject earns 9 experimental points from the guess; if there is a difference of 1 between the guess and the average, 6 points are earned; and a difference of 2 results in 3 points earned. Larger differences are neither rewarded nor punished.

4.3.2 Elicitation of natural risk attitudes

In the second part of the experiment, we used the design by Holt and Laury (2002) to measure individual risk attitudes. Each subject makes ten risky decisions without interacting with another player. In each decision they choose between Option X and Option Y, where both options include a lottery with the same probabilities but different payoffs. Option X is the relatively safer option because both possible lottery outcomes are between the outcomes of Option Y. Throughout the decisions, the payoffs are fixed, but the probability of receiving the higher payoff increases by 10 percentage points from 10% in decision 1 to 100% in decision 10 in both options. The exact amounts of money that were used can be found in Appendix 4C.

Depending on the subject's risk attitude, the subject should, moving down the decisions, switch at some point from Option X to Option Y (or in the unlikely case of extreme risk-loving always choose Option Y). Switching from Y to X or choosing always X is incompatible with consistent money maximizing behavior since switching back and forth violates monotonicity and sticking with Option X throughout gives a lower sure payoff for decision 10 than switching to Option Y (i.e., a violation of dominance). The point at which subjects switch from Option X (called the *safe* option, henceforth) to Option Y can then be used to calculate the degree of risk aversion. Higher values of the variable imply relatively higher levels of risk aversion. One of the ten lottery choices was randomly selected and played for real. Subjects could earn up to 3.85 euro in this part.

4.3.3 Trust game

The trust experiment followed the classical design by Berg et al. (1995), but each subject played both the role of sender and receiver (such as, for instance, in Burks et al., 2003). In the experiment, the sender is given an endowment of 20 experimental points, and he or she decides how much of the endowment (in integers) to send to the receiver. The amount sent by the sender is tripled before it reaches the receiver. The receiver finally decides on how much to return to the sender (the returned amount is not tripled). A rational and selfish individual, assuming common knowledge of rationality and selfishness, would send nothing to the receiver, as backward induction implies that a payoff-maximizing receiver has no incentive to send anything back. There is, however, a possibility for a Pareto improvement if the receiver returns at least one third of the tripled amount received.

The amount sent by the sender is typically seen as an indication of trust. Since we wanted to obtain trust measures for all subjects, all participants had to make decisions in both roles without knowing which role they would finally be playing. In the role of receiver, we used the strategy method like in the public goods part, i.e., subjects were asked to indicate how much they would send back for all the 21 possible amounts that they could receive.

For determining monetary payoffs, we randomly matched subjects into pairs with randomly assigned actual roles of senders and receivers. The monetary payoff was then determined by subjects' decisions, i.e., the amount sent by the sender and the amount indicated to send back by the receiver conditional on the amount sent. Each experimental point in the trust game was exchanged for 0.33 euro as in the public goods part.

4.3.4 Procedure

The computer-based experiments were conducted at the experimental laboratory MELESSA of the University of Munich in October 2009 and March 2010 using the experimental software z-Tree (Fischbacher, 2007) and the organizational software Orsee (Greiner, 2004). A total of 144 undergraduate students from all disciplines except economics participated in six sessions with 24 participants each. The sessions lasted up to 1½ hours, and the average payoff was 16.98 euro, including a show-up payment of 4.00 euro.

An experimental session started with instructions for the public goods game. At that time, subjects received instructions only for the public goods game, but they knew that there would be two more parts in the experiment and that these parts would be unrelated to the public goods game as well as to each other. Subjects received written instructions, which were read aloud, and they had the opportunity to ask questions in private. The public goods game only

began when all subjects correctly understood the procedures and after all subjects had passed through some computerized exercises, where they had to compute profits for different contribution levels in the game. At the end of the public goods part, beliefs about others' contributions were elicited. Upon completion, subjects received instructions for the second part, the risk attitude elicitation part, and finally, after the risk elicitation part, for the trust part. These instructions were also read aloud, and plenty of time was given to ask questions in private. We also took care that the matching of groups in the public goods game and the trust game was different, and this was clearly stated in the instructions. As already mentioned, the decisions and results of the different parts were only revealed at the end of the entire experiment to avoid any effects from earnings in one part on behavior in subsequent parts. Before payment, subjects answered a post-experimental questionnaire, among others including some questions related to socio-economic factors (and including a self-control measure used in Kocher et al., 2011). Finally, subjects were paid privately in cash and, then, were free to leave.

4.4 Results

We start with the presentation of descriptive results. Note that we have excluded fifteen subjects from our analysis that did not provide consistent answers in the risk experiment (i.e., that did switch back from Option Y to Option X (the safe option) or always chose the safe option, which is incompatible with consistent money maximizing behavior).[130] The average unconditional contribution to the public good is 6.67 points (33.35% of the endowment) and the corresponding guessed contribution by others is 7.28 points (36.40%). These levels correspond well to previous findings in German-speaking countries (e.g., Fischbacher and Gächter, 2010; Fischbacher et al., 2001; Kocher et al., 2008). In the trust game, 7.53 points are on average sent by the sender, and the corresponding level of 37.65% of the endowment as transfers to the responder also corresponds to what has been previously found (e.g., Camerer, 2003; Cárdenas and Carpenter, 2008). In the risk experiment, the safe option is chosen on average 6.14 times. A risk-neutral subject would choose the safe option four times, and thus our data indicate that subjects are on average risk averse. Our findings regarding risk are very similar to the results found by Holt and Laury (2002).

[130] We have also conducted all the analyses without excluding inconsistent subjects, where risk preferences were measured as the number of safe choices. Our results are unaffected both in magnitude and significance.

Using the design by Fischbacher et al. (2001), we categorize subjects into different types of contributors based on their submitted conditional contribution schedule. If a subject's own conditional contribution increases weakly monotonically with the average contribution of the other members, the subject is classified as a c*onditional cooperator*. Moreover, a subject is classified as a conditional cooperator if the relationship between own and others' average contributions is positive and significant at the 1% significance level based on the Spearman rank correlation coefficient (see the classifications used in, e.g., Fischbacher et al., 2001; Fischbacher and Gächter, 2010). *Hump-shaped contributors*, sometimes also called triangle contributors, are subjects who show weakly monotonically increasing (or increasing with a Spearman rank correlation coefficient at the 1% significance level) contributions up to a given level of others' contributions; above that level, their conditional contributions decrease based on a reversed classification as used to the inflection point. A *free-rider* is a subject who has a conditional contribution of zero for all levels of the other members' contributions. Finally, those who cannot be categorized into any of the above groups are referred to as *others*.

As shown in Table 4.1, we find that 20.16% of our subjects can be classified as free-riders, 58.13% as conditional cooperators, 11.63% as hump-shaped and 10.08% as others, which again is very similar to the proportions reported in, e.g., Fischbacher et al. (2001) and Kocher et al. (2008). In columns 3-6 of Table 4.1, we show descriptive statistics on the behavioral variables that we discussed above for the whole sample but now do separately for each type of contributor. As expected, the unconditional contribution differs significantly at the 1% level between the four types of contributors based on a Kruskal-Wallis test. Conditional cooperators on average contribute 8.11 points unconditionally, whereas free-riders only contribute 1.12 points. The average unconditional contributions for the hump-shaped and other types are 6.80 and 9.31 points, respectively.

In Table 4.1, we have a sub-section in which we only focus on conditional cooperators and free-riders for two reasons. First, they exhibit clear and consistent patterns of behavior, and, second, they comprise the majority (78.29%) of types in our sample. Not surprisingly, the unconditional contribution levels differ significantly between free-riders and conditional cooperators according to a Mann-Whitney test ($p < 0.01$).

We find similar differences – though smaller in magnitude – between the types when we investigate guessed contributions by others. The free-riders on average guessed that others would contribute 4.31 points compared to conditional contributors, who guessed 7.85 points. Therefore, free-riders also have a more pessimistic view of cooperativeness of others than conditional cooperators, or they fall prey to the false consensus effect despite the

incentivizing of the guess. In any case, we can reject the hypothesis of equality in guessed contributions both for all four types of contributors as well as for a pairwise comparison of free-riders and conditional contributors at the 1% significance level. Interestingly, in the trust game, the pattern of transfers is very similar to the contributions in the public goods game. Free-riders sent on average 2.58 points, compared to conditional cooperators, who sent 9.04 points. Again, statistical tests reject equality both for all four types of contributors and for a pairwise comparison of free-riders and conditional contributors at 1% significance level. However, when it comes to natural risk preferences, there are neither statistically significant differences between the four types of contributors ($p = 0.83$), nor for the pairwise comparison of free-riders and conditional cooperators ($p = 0.91$).[131]

Result 4.1. *The four types of contributors differ significantly in their unconditional contributions, their guessed contributions by others and their amounts sent in the trust game. However, there are no significant differences with regard to natural risk.*

Table 4.1: Descriptive statistics of the experiment and non-parametric tests (N = 129)

Type of subject	Proportion of subjects	Unconditional contribution	Guessed contribution by others	Amount sent (trust game)	Natural risk
Free-riders	20.16%	1.12	4.31	2.58	6.27
Conditional cooperators	58.13%	8.11	7.85	9.04	6.20
Hump-shaped contributors	11.63%	6.80	8.00	8.80	5.73
Others	10.08%	9.31	9.08	7.31	6.00
H0: No difference between types (Kruskal-Wallis test (p-value))		<0.01	<0.01	<0.01	0.83
H0: No difference between free-riders and conditional cooperators (Mann-Whitney test (p-value))		<0.01	<0.01	<0.01	0.91
All types	*100%*	*6.67*	*7.28*	*7.53*	*6.14*

Note: 15 subjects with inconsistent risk preferences are excluded.

In the following, we investigate the factors associated with being a specific contributor type using a multinomial logit model. In the analyses, we merge the *hump-shaped* and *others* types to one category *hump-shaped/others*. The three models in Table 4.2 assess the factors that influence the classification of being a free-rider, a conditional cooperator, and being of

[131] The raw correlations between unconditional contribution and the risk measure are -0.15 ($p = 0.20$) for conditional cooperators and 0.13 ($p = 0.51$) for free-riders.

the residual type. The reference group in the regressions consists of conditional cooperators, and thus the coefficients show how the different variables increase or decrease the likelihood of being classified as a free-rider or as hump-shaped/others compared to being classified as a conditional cooperator. We run three models, as we include belief and trust levels both separately as well as together.[132]

We also conducted a regression analyzing whether natural risk is associated with trust because there are significant and positive effects reported in previous research (e.g., Schechter, 2007). Such an analysis is important to decide whether we should allow for the effect in our econometric models. However, risk was insignificant at the 5% level (p = 0.25), and thus we only investigate the direct effect of risk on the likelihood of being of one of the types.

In all three models of Table 4.2, the coefficients of trust and beliefs are significantly negative for free-riders at the 5% significance level. In other words, both lower levels of trust as well as lower levels of beliefs in others' contributions are associated with being classified as a free-rider. In line with the descriptive results, natural risk does not significantly affect the likelihood of being classified as a certain type.

Table 4.2: Estimation results from multinomial logit model – contributor type

Dependent variable: Contributor type	Model 1		Model 2		Model 3	
	Free-riders	Hump-shaped/ Others	Free-riders	Hump-shaped/ Others	Free-riders	Hump-shaped/ Others
Belief about others' contribution	-0.28***	0.04	-	-	-0.27***	0.05
	(0.08)	(0.05)			(0.08)	(0.05)
Trust	-	-	-0.24***	-0.03	-0.20**	-0.04
			(0.09)	(0.03)	(0.09)	(0.03)
Natural risk	0.09	-0.16	-0.02	-0.17	0.07	-0.19
	(0.15)	(0.15)	(0.17)	(0.14)	(0.17)	(0.15)
Constant	0.033	-0.31	0.33	0.30	0.80	0.00
	(1.03)	(1.09)	(1.15)	(0.94)	(1.11)	(1.10)
Number of observations	129		129		129	

Notes: *** denotes significance at the 1% level, ** at the 5% level and * at the 10% level. Coefficients reported. Robust standard errors in parentheses. The reference group is conditional cooperators.

[132] As discussed in Thöni et al. (2009), there is an intuitively obvious correlation between trust and the stated belief regarding others' contribution: somebody who trusts others should have higher expectations in the cooperativeness of others. We follow the approach of Thöni et al. (2009) by estimating models that include only beliefs or only trust to avoid potential issues of multicollinearity and models that include both.

Result 4.2. *The likelihood of being classified as a free-rider rather than a conditional cooperator increases both in lower levels of trust and lower levels of beliefs in others' contributions. Natural risk does not matter significantly, both for the type classification and for the trust behavior.*

In Table 4.3, we show the results of how unconditional contributions in the public goods game are associated with belief, trust and natural risk preferences. Trust and beliefs are clearly associated with unconditional contribution behavior. In model 1 of Table 4.3, where we included the stated belief together with natural risk, the belief is significant ($p < 0.01$). In model 2, we included trust instead of the belief, and trust is significant in the regression ($p < 0.01$). In the third regression, where both belief and trust are included, we find that only the belief is significant ($p < 0.01$). However, trust and the stated belief in others' contributions are clearly associated. In contrast, when using the risk measure from the natural risk experiment task as an independent variable in a regression, the risk coefficient is insignificant in all models again. Therefore, natural risk preferences do not seem to influence behavior in the public goods game, and we have to refute the main hypothesis from our model. Two further observations are noteworthy. First, if only the risk measure is included in the regression, the coefficient does not become significant at all. Second, p-values of the coefficients for the risk measure decline a bit, when we only look at those subjects that were classified as conditional cooperators, but they are still far from being significant on conventional levels.

Table 4.3: Estimation results from OLS model – unconditional contributions

Dependent variable: Unconditional contribution	Model 1	Model 2	Model 3
Belief about others' contribution	1.09***	-	1.05***
	(0.07)		(0.09)
Trust	-	0.34***	0.08
		(0.09)	(0.07)
Natural risk	-0.11	-0.06	-0.08
	(0.21)	(0.34)	(0.21)
Constant	-0.59	4.53**	-1.04
	(1.26)	(2.22)	(1.26)
Number of observations	129	129	129

Notes: *** Denotes significance at the 1% level, ** at the 5% level and * at the 10% level. Robust standard errors in parentheses. The results are very similar for a tobit regression model.

Result 4.3. *Trust and beliefs are positively associated with unconditional contributions whereas there is no significant impact of natural risk.*

One way of testing our theoretical model more directly – at the cost of having to specify a couple of parameters – is to use the first-order condition for equation (4.1) and calculate β_i, which measures the relative importance of utility from monetary payoffs and disutility from deviating from the expected average deviation of others. One has to make some assumptions on the functional form of the utility function. We use a CRRA utility function, where the risk parameters are recovered from the Holt and Laury-task, and a quadratic cost-of-deviation function between one's own contribution and the average contribution of the other players. Because the latter is unknown for a subject when deciding, we plug in the individual expectation. The individual β_i should predict social inclination, and we find that a higher β_i is highly significantly associated with the probability of being a conditional cooperator.[133] However, the connection is closest, when we calculate β_i without taking risk preferences into account. This indicates that (i) the spirit of our model is capturing important aspects of cooperative behavior and that (ii) attitudes towards natural risk do not seem to help in understanding behavior in the social dilemma.

Yet another way of looking at the association between natural risk and cooperation is to compare implied unconditional contributions with actual unconditional contributions. Implied unconditional contributions are defined as the unconditional contribution one would expect from an individual taking his or her beliefs about others' average contributions and the according number in the contribution schedule. If somebody, for instance, expects an average contribution of the other three group members of 5 points, and if the same person indicates in the contribution schedule that he or she is going to contribute 3 points to the public good in case the average contribution of others is 5 points, then the implied unconditional contribution is 3 points. Because the contribution table is completely deterministic and thus risk-free – one can condition one's contribution on the contribution of others, and there is no uncertainty involved – the implied unconditional contribution includes the assumption that subjects think their belief is correct. If conditional cooperators are not entirely sure about their guesses and are on average risk averse, one could expect that the average implied unconditional contribution should be higher than the average actual unconditional contribution because the latter involves social risk. This is, however, not the case in our data. If anything, on average actual unconditional contributions for conditional cooperators are slightly higher than average implied unconditional contributions.[134] On the other hand, they should be the same for free-riders, and this is also what we find.

[133] We run probit regressions (not reproduced here) with a dummy variable (conditional cooperator = 1) as the dependent and β_i as well as controls as independent variables.
[134] Strategic reputational aspects should not play a role because we have a one-shot interaction in our experiment.

4.5 Conclusion

By using a laboratory experiment, we have investigated how beliefs about others' contributions, trust, and risk preferences together play a role in shaping contributions in a public goods experiment. According to Fischbacher et al. (2001), we classify subjects into different contribution types. Previous findings documenting that conditional cooperation is a widespread behavioral type are supported by our experimental results. We further find that beliefs about others' contributions and trust elicited by a trust game are significantly associated with public good contributions, whereas natural risk preferences neither affect contributions to the public good nor affect trust behavior in our experiment. Our findings regarding the correlation between trust and cooperation are similar to those in Thöni et al. (2009) despite the fact that we use an incentivized trust game.

The result that trust and cooperation are highly correlated is not surprising. It is intuitively clear that voluntary contribution to a public good involves a certain level of trust in the contribution of others. The association between trust and cooperation can be seen in both actual trusting behavior and in stated beliefs. Interestingly, free-riders not only contribute and trust less but also have less optimistic expectations about other' contributions, in line with the false consensus effect.

It is surprising that the attitude towards natural risk does not seem to play a role at all in shaping trust or in explaining cooperation in our experiments. It seems that social risk in a contribution decision is indeed something different to natural risk, as has already been indicated by Bohnet et al. (2008). However, in contrast to Bohnet et al., we do not find any association between natural risk and trust/cooperation. Of course, we cannot exclude the possibility that our risk measure does not measure actual attitudes towards natural risk correctly. However, we have been using a widely accepted and often used method for eliciting natural risk preferences that has been validated by others. A natural extension to our experiment would be to employ the method developed by Bohnet and co-authors, and to re-assess the relationship between cooperation, trust and risk preferences based on the explicit distinction of natural and social risk. However, the comparison of implied unconditional contributions with actual unconditional contributions from our data seems to indicate that our subjects do not perceive the decision on unconditional contributions in the social dilemma as socially risky at all.

The literature on trust and the literature on cooperation in economics, specifically in experimental economics, have been distinct to a certain extent. Our results provide another

piece of evidence suggesting that one should see them as strongly related concepts and that it would be helpful to further improve the economists' knowledge of the interactions between cooperation, beliefs in others' behavior and trust. For policy makers, our results highlight the importance of high levels of trust as an important ingredient for achieving high degrees of voluntary cooperation in a society. Therefore, this indicates that building trust is an important activity for policies aiming at increasing the contributions to public goods. Such a strategy especially appears to create a virtuous circle of cooperation among the often large number of conditional cooperators, who by their behavior will both contribute more to public goods as well as reduce the speed of decay of contributions to public goods over time (see Fischbacher and Gächter, 2010, on the dynamic effects of the interaction of free-riders and conditional cooperators).

Trust building is an important alternative to previously tested institutions in public goods games that have focused on increasing contributions. For instance, monetary punishment (e.g., Bochet et al., 2006; Fehr and Gächter, 2000; Ostrom et al., 1992) and exclusion by voting (e.g., Cinyabuguma et al., 2005; Maier-Rigaud et al., 2010) have the potential of substantially increasing contributions to a public good. In the case of monetary punishment, the overall effect on efficiency has shown to be negative in the short run, whereas in a long run public goods experiment with 50 periods Gächter et al. (2008) reports positive effect as the degree of punishment decreases over time. Of course, ultimately it depends on how subjects use the punishment option, which partly can be related to the number of periods they interact, but group culture effects have been shown to be important here as well (Hermann et al., 2008; Gächter et al., 2010). Trust building in reality is also a costly activity. However, the effect of trust is supposedly more long term compared to the sharp reduction of contributions to public goods when the monetary punishment possibility is revoked (e.g., Fehr and Gächter, 2000). Few of the contribution-enhancing mechanisms have been applied to trust games, and trust has not been considered as a mechanism that one can influence exogenously. However, building trust – even though economists do not yet understand well enough how it works (one study addressing this issue is Näf and Schunk, 2009) – seems to be an interesting alternative mechanism to decentralized sanctioning. Future research in economics could strengthen its focus on trust building and its institutional prerequisites.

Appendix

4A Proofs

In what follows we suppress subscript i on the payoff function π for clarity of exposition.

Proof of Proposition 4.1

Recall the agent's maximization problem from (4.1):

$$\max_{c_i} U = u(\pi) - \beta_i c(|c_i - \bar{c}|) \tag{4A.1}$$

Given a linear public goods technology, i.e., $\pi = \pi(c_i, \bar{c})$, $\pi_1'(c_i, \bar{c}) < 0$, $\pi_2'(c_i, \bar{c}) > 0$ and $\pi_{11}''(c_i, \bar{c}) = \pi_{22}''(c_i, \bar{c}) = 0$, the first order condition is $u'(\pi)\pi_1' - \beta_i c'(|c_i - \bar{c}|) = 0$, and so the implicit function becomes $f = u'(\pi)\pi_1' - \beta_i c'(|c_i - \bar{c}|)$.

The derivative of c_i^* with respect to $\bar{c}$ is then

$$\frac{\partial c_i^*}{\partial \bar{c}} = -\frac{\frac{\partial f}{\partial \bar{c}}}{\frac{\partial f}{\partial c_i}} = -\frac{u''\pi_2' + \pi_{12}''u' - \beta_i c''(|c_i - \bar{c}|)}{u'' + u'\pi_{11}'' - \beta_i c''(|c_i - \bar{c}|)}, \tag{4A.2}$$

which is greater than zero if $\beta_i c''(|c_i - \bar{c}|) > u''\pi_2' + \pi_{12}''u'$.

Using the definition of the Arrow-Pratt measure of risk aversion $\lambda = -u''/u'$ and the linearity of the public goods technology ($\pi_{12}'' = 0$) we can write $(\beta_i/\lambda) \cdot (c''(|c_i - \bar{c}|)/u') + \pi_2' > 0$ which is positive.

In addition, given $\beta_i > 0$, the left hand side of the condition increases in the fraction β_i/λ, i.e., the sensitivity to others' average contribution over risk aversion, which proves the proposition.

4B Experimental instructions (originally in German)

Welcome to the experiment and thank you for participating!

Please do not talk to other participants.

General

This is an experiment on economic decision making. You will earn "real" money that will be paid out to you in cash at the end of the experiment. During the experiment all participants will be asked to make decisions. Your decisions and the decisions of other participants determine your earnings from the experiment according to the following rules.

The experiment will last two hours. If you have any questions or if anything is unclear, please raise your hand, and one of the experimenters will come to you and answer your questions privately.

During the experiment a part of your earnings will be calculated in **points**. At the end of the experiment all points that you earn will be converted into euro at the exchange rate of

1 point = 0.33 euro (3 points = 1 euro).

In the interest of clarity, we will only use male terms in the instructions.

Anonymity

You will learn neither during nor after the experiment, with whom you interact(ed) in the experiment. The other participants will neither during nor after the experiment learn, how much you earn(ed). We never link names and data from experiments. At the end of the experiment you will be asked to sign a receipt regarding your earnings which serves only as a proof for our sponsor. The latter does not receive any other data from the experiment.

Means of help

You will find a pen at your table which you, please, leave behind on the table when the experiment is over. While you make your decisions, a clock will run down at the top of your computer screen. This clock will give you an orientation how long you should need to make your decisions. But you can nevertheless exceed this time. The input screens will not be dismissed once time is over. However, the pure output screens (here you do not have to make a decision) will be dismissed.

Experiment

The experiment consists of three parts. You will receive instructions for a part after the previous part has ended. The parts of the experiment are completely independent; decisions in one part have no consequences for your earnings in later parts. The sum of earnings from the different parts will constitute your total earnings from the experiment.

Part I

The decision situation

The basic decision situation will be explained to you in the following. Afterwards you will find control questions on the screen which should raise your familiarity with the decision situation.

You will be a member of a group consisting of **4 people**. Each group member has to decide on the allocation of 20 points. You can put these 20 points into your **private account** or you can put them **fully or partially** into a **group account**. Each point you do not put into the group account will automatically remain in your private account.

Your income from the private account:

You will earn one point for each point you put into your private account. For example, if you put 20 points into your private account (and therefore do not put anything into the group account) your income will amount to exactly 20 points out of your private account. If you put 6 points into your private account, your income from this account will be 6 points. No one except you earns something from your private account.

Your income from the group account:

Each group member will profit equally from the amount you put into the group account. On the other hand, you will also get a payoff from the other group members' in-payments into the group account. The income for each group member out of the group account will be determined as follows:

Income from group account =
Sum of all group members' contributions to the group account $\times 0.4$

If, for example, the sum of all group members' contributions to the group account is 60 points, then you and the other members of your group each earn 60 x 0.4 = 24 points out of the group account. If the four group members contribute a total of 10 points to the group account, you and the other members of your group each earn 10 x 0.4 = 4 points out of the group account.

Total income:

Your total income is the sum of your income from your private account and that from the group account:

Income from your private account (= 20 – contribution to group account)
+ Income from group account (= 0,4 $\times$ *sum of contributions to group account)*
= Total income

Before we proceed, please try to solve the control questions on your screen. If you want to compute something, you can use the Windows calculator by clicking on the respective symbol on your screen.

Procedure of Part I

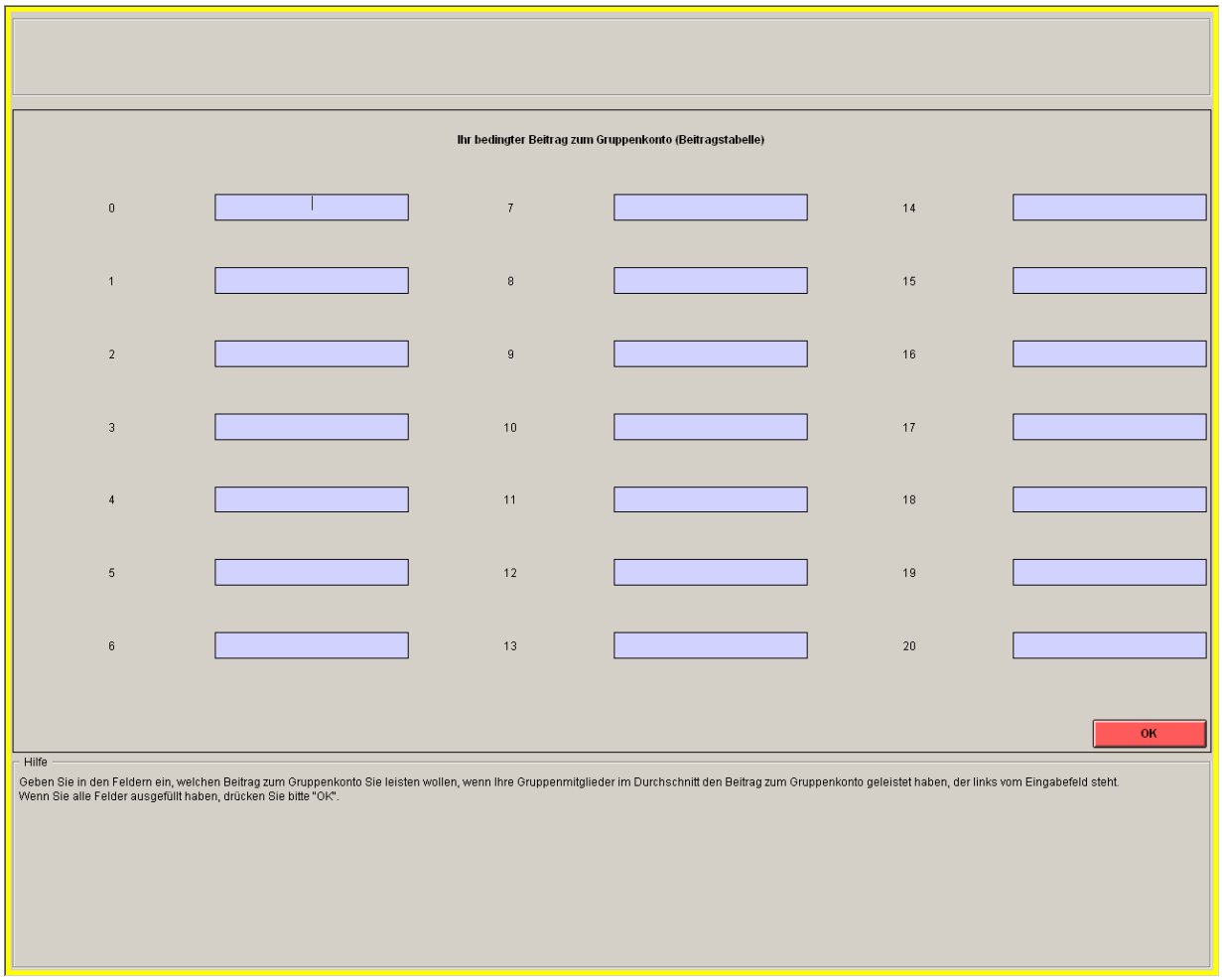
Part I includes the decision situation just described to you. The decisions in Part I will only be made **once**.

On the first screen you will be informed about your **group membership number**. This number will be of relevance later on. If you have taken note of the number, please click "next".

Then you have to make your decisions. As you know, you will have 20 points at your disposal. You can put them into your private account or you can put them into the group account. Each group member has to make **two types** of contribution decisions which we will refer to below as the **unconditional contribution** and the **contribution table**.

- In the **unconditional contribution** case you decide how many of the 20 points you want to put into the group account. Please insert your unconditional contribution in the respective box on your screen. You can insert integer numbers only. Your contribution to the private account is determined automatically by the difference between 20 and your contribution to the group account. After you have chosen your unconditional contribution, please click "next".

- On the next screen you are asked to fill in a **contribution table**. In the contribution table you indicate **how much <u>you</u> want to contribute to the group account for each possible average contribution of the <u>other</u> group members** (rounded to the next integer). Thus, you can condition your contribution on the other group members' average contribution. The contribution table looks as follows:



The numbers in each of the left columns are the possible (rounded) average contributions of the **other** group members to the group account. This means, they represent the amount each of the other group members' has put into the group account on average. You simply have to insert into the input boxes how many points you want to contribute to the group account – conditional on the indicated average contribution. **You have to make an entry into each input box**. For example, you will have to indicate how much you contribute to the group account if the others contribute 0 points to the group account on average, how much you contribute if the others contribute 1, 2, or 3 points on average, etc. You can insert any integer numbers from 0 to 20 in each input box. Once you have made an entry in each input box, please click "OK".

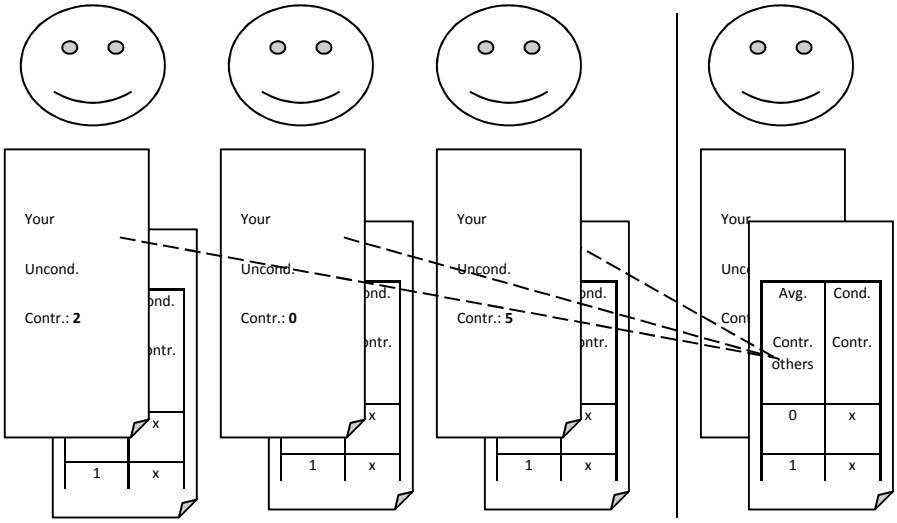
After all participants of the experiment have made an unconditional contribution and have filled in their contribution table, a random mechanism will select a group member from every group. Only **the contribution table** will be the payoff-relevant decision for the **randomly determined subject**. Only the **unconditional contribution** will be the payoff-relevant decision for the **other three group members** not selected by the random mechanism. You obviously do not know whether the random mechanism will select you when you make your unconditional contribution and when you fill in the contribution table. You will therefore have to think carefully about both types of decisions because both can become relevant for you. Two examples should make this clear.

Example 1: Assume that **the random mechanism selects you. This implies that your relevant decision will be your contribution table**. The unconditional contribution is the relevant decision for the other three group members. Assume they made unconditional contributions of 0, 2, and 5 points. The average rounded contribution of these three group members, therefore, is 2 points ((0+2+5)/3 = 2.33).
If you indicated in your contribution table that you will contribute 1 point to the group account if the others contribute 2 points on average, then the total contribution to the group account is given by 0+2+5+1=8 points. All group members, therefore, earn 0.4×8=3.2 points out of the group account plus their respective income from the private account.
If, instead, you indicated in your contribution table that you would contribute 19 points if the others contribute two points on average, then the total contribution of the group to the group account is given by 0+2+5+19=26. All group members therefore earn 0.4×26=10.4 points out of the group account plus their respective income from the private account.

Example 2: Assume that **the random mechanism did not select you,** implying that **the unconditional contribution is taken as the payoff-relevant decision** for you and two other group members. Assume your unconditional contribution to the group account is 16 points and those of the other two group members are 18 and 20 points. The average unconditional contribution of you and the other two group members, therefore, is 18 points (= (16+18+20)/3).
If the group member whom the random mechanism selected indicates in her contribution table that she will contribute 1 point to the group account if the other three group members contribute on average 18 points, then the total contribution to the group account is given by 16+18+20+1=55 points. All group members will therefore earn 0.4×55=22 points out of the group account plus their respective income from the private account.
If, instead, the randomly selected group member indicates in her contribution table that she contributes 19 points to the group account if the others contribute on average 18 points, then the total contribution to the group account is given by 16+18+20+19=73 points. All group members will therefore earn 0.4×73=29.2 points out of the group account plus their respective income from the private account.

The random selection of the participants will be implemented as follows. A randomly selected participant will throw a 4-sided dice **after** all participants have made their unconditional contribution and have filled in their contribution table. She enters the thrown number into the computer thereby being monitored by the experimenter who confirms the correctness of the entry by password. The thrown number will then be compared with the group membership number, which was shown to you on the first screen. If the thrown number equals your group membership number, then your contribution table is payoff-relevant for you and the unconditional contribution is payoff-relevant for the other three group members. Otherwise, your unconditional contribution is the relevant decision for you.

The following figure visualizes the situation in example 1. You are the person on the right side with group membership number 3. Number 3 was thrown and therefore your conditional contribution is payoff-relevant. For the other three group members the unconditional contribution is payoff-relevant.



You will make all your decisions only **once**. After the end of Part I you will get the instructions of Part II. How much you have earned in Part I will be revealed at the end of the experiment.

Part II

In Part II you will receive **10 decision problems**. You do not interact with another person in this part. In each of the problems you can choose between **two alternative lotteries**. Your decisions are only valid after you have made a decision for all problems and after you have clicked on the OK-button in the lower part of your screen. Take your time for your decisions because your choice determines your earnings from the second part according to the rules described below.

Here is an example for such a decision problem:

Lottery X	**Lottery Y**	**Your choice**
You receive 2 EUR with probability 8/10 or 1.60 EUR with probability 2/10	You receive 3.85 EUR with probability 8/10 or 0.10 EUR with probability 2/10	□ Lottery X □ Lottery Y

Your earnings will be determined in the following way: First, the computer chooses one of the 10 decision problems randomly and with equal probability. The lottery that you chose for this decision problem will then be simulated in the way that the computer draws a random number between 0 and 10.
For example: Assume that the computer randomly chooses the decision problem from the table above, and your choice was lottery X. Then, the computer simulates lottery X, and you receive either 2 EUR (with probability 8/10 = 80%) or 1.60 EUR (with probability 2/10 = 20%) as your earnings from Part II of the experiment. You will receive the high payoff if the randomly chosen number is smaller or equal to 8 (80% probability) and the low payoff if the random number is bigger than 8 (20% probability).
If, however, the computer chooses a decision problem with a 40% probability of receiving the high payoff, then each random number below or equal to 4 will result in the high payoff whereas all numbers bigger than 4 lead to the low payoff, etc.

Please note that we are talking about euro-amounts here and not about points! The euro-amount that you will earn in Part II will be added to the in euro converted points from the other parts.
You will make your decisions only **once**. After the end of Part II you will get the instructions of Part III. How much you have earned in Part II will be revealed at the end of the experiment.

Part III

The decision situation

At the beginning of Part III all participants will be randomly matched into **groups of two**. In each pair **both participants will slip into the roles A and B**. Afterwards, it will be determined randomly for whom role A and for whom role B is payoff-relevant. **Your interaction partner will be no one who was member of your group in Part I**! On the screen you first have to make decisions in the role of participant A and afterwards in the role of participant B.

Participant A has an endowment of 20 points. Participant B has no endowment. Participant A has to decide how many of the 20 points she wants to send to participant B. **She can send every integer number X between 0 and 20** (0 and 20 are also possible). Participant A will keep the residual (20-X), while the **amount X sent to participant B is tripled**. This means that for each point participant A sends to B, B will receive three points.

Participant B has to decide how many points of the tripled amount she wants to send back to A. **She can send back every integer number Y between 0 and 3X** (0 and 3X are also possible). Participant B will keep the residual (3X-Y). Note, that the **amount Y sent back to participant A is not tripled.**

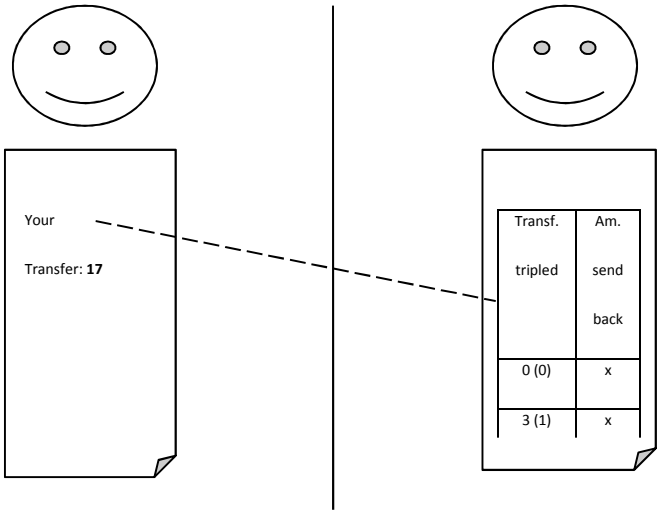
Procedure of Part III

In each pair both subjects will first slip into role A and afterwards into role B. In the role of participant A you will decide about the transfer to participant B. In the role of participant B you will decide how much you want to send back to A for each possible integer transfer between 0 and 20. The corresponding screen will look as follows:

On the left side of each input box you can see the amount that you have at your disposal. This amount is three times the transfer by participant A. The transfer of participant A itself is denoted in brackets. If you want to compute something, you can click on the calculator symbol.

After all participants of the experiment have chosen a transfer and have filled in the table, the computer will determine randomly which group member is assigned to role A and role B respectively. Payoff-relevant then are only the decisions made in the assigned roles. This means in particular that the value extracted from the contribution table is the value the person in role B has chosen for the actual transfer of the person in role A.
When you make your decisions you do not know whether role A or role B will be payoff-relevant for you. Therefore it is reasonable to think carefully about your decisions in both roles.
As a reminder, here are the payoffs for both participants: A will receive the residual of her endowment (20-X) plus the amount Y sent back by B (in sum: 20-X+Y). B will receive three times the amount sent by A (3X) minus the amount Y sent back to A (in sum: 3X-Y).

As an example, please consider the following figure. The figure shows the transfer of participant A (17). In this case participant B obtains 51 points (17x3). From these 51 points she sends 12 points back to participant A.



You will make your decisions only **once**. At the end of the experiment you will learn your payoff-relevant role and how much you have earned in Part III.

After Part III is finished we will ask you to fill in a short questionnaire on the screen. Afterwards you will learn for each part separately how much you have earned. Then the experiment ends. There are neither more parts nor any repetitions. Finally, you will be informed about your total earnings from the experiment and paid out.

4C Measuring individual risk attitudes with the Holt and Laury (2002) design

Table 4C.1: The ten paired lottery-choice decisions

Option X	Option Y	Expected payoff difference
1/10 of €2.00, 9/10 of €1.60	1/10 of €3.85, 9/10 of €0.10	€1.17
2/10 of €2.00, 8/10 of €1.60	2/10 of €3.85, 8/10 of €0.10	€0.83
3/10 of €2.00, 7/10 of €1.60	3/10 of €3.85, 7/10 of €0.10	€0.50
4/10 of €2.00, 6/10 of €1.60	4/10 of €3.85, 6/10 of €0.10	€0.16
5/10 of €2.00, 5/10 of €1.60	5/10 of €3.85, 5/10 of €0.10	-€0.18
6/10 of €2.00, 4/10 of €1.60	6/10 of €3.85, 4/10 of €0.10	-€0.51
7/10 of €2.00, 3/10 of €1.60	7/10 of €3.85, 3/10 of €0.10	-€0.85
8/10 of €2.00, 2/10 of €1.60	8/10 of €3.85, 2/10 of €0.10	-€1.18
9/10 of €2.00, 1/10 of €1.60	9/10 of €3.85, 1/10 of €0.10	-€1.52
10/10 of €2.00, 0/10 of €1.60	10/10 of €3.85, 0/10 of €0.10	-€1.85

Note that risk neutral people choose option X for the first four lotteries and switch to option Y afterwards. Risk averse people will switch to option Y later whereas risk-loving individuals switch to Y before the fourth lottery.

Chapter 5

Gender and Cooperative Preferences[135]

5.1 Introduction

The research on gender differences in cooperation is one of the areas where existing empirical results could not be more inconclusive. While some studies provide evidence for women to be more cooperative in social dilemmas, others find the opposite, and still others report no effects at all. In public goods games, lower voluntary contributions of women are reported by Sell and Wilson (1991) and Brown-Kruse and Hummels (1993). Replicating the latter study, however, Cadsby and Maynes (1998) do not find any gender differences. Sell et al. (1993), Solow and Kirkwood (2002), Chermak and Krause (2002) and Andreoni and Petrie (2008) also find no significant differences in the contributions of men and women. Contrary results are reported by Stockard et al. (1988), Nowell and Tinkler (1994) and Seguino et al. (1996), who show that women contribute more than men. Frank et al. (1993) observe women to cooperate significantly more than men in a prisoner's dilemma game. Ortmann and Tichy (1999) find the same result, but only for the first round, a time effect also documented by Mason et al. (1991).[136]

Whereas it is conceivable that some of the inconclusiveness might arise from different empirical protocols and different experimental designs in different studies, it is still concerning and difficult to explain that the evidence is so mixed. In this chapter we argue that inconclusive evidence so far can be explained when looking at three aspects together: contributions in a social dilemma, cooperative preferences and beliefs. We experimentally implement a linear public goods game in the laboratory in order to study gender differences. It is rational for selfish individuals to contribute nothing to the public good, whereas collective rationality implies full contribution. In contrast to most of the existing literature, we elicit not only unconditional contributions to the public good, but also cooperative preferences (conditional contributions) and beliefs in an incentivized way. Our design is similar to the one

[135] This chapter is joint work with Nadja Furtner, Martin Kocher, Peter Martinsson and Conny Wollbrant.
[136] For reviews of the literature on gender effects in social dilemmas see Eckel and Grossman (2008) and Croson and Gneezy (2009).

pioneered by Fischbacher et al. (2001) and later validated by Fischbacher and Gächter (2010). We are not aware of any other paper that looked at gender differences in cooperation in a similar way. The advantages of our approach will become apparent when we discuss our results, but some are obvious: disentangling differences in beliefs from differences in cooperative preferences is important for understanding existing results. In addition, eliciting a full contribution schedule allows for a much finer-grained picture of potential differences in cooperation between men and women than just unconditional contributions as it, for instance, eliminates strategic uncertainty.

In order to address the criticism of small sample size leading to non-robust results and to make sure that we are not picking up chance results, we take the results from the current study and look at a set of existing experimental data. These experiments are not completely identical to the main one presented here, but they are broadly comparable and thus allow for putting our results to another test based on a much larger sample.

It does not do justice to the literature outside economics to claim that there is no theoretical foundation for a potential gender difference in cooperation. Gilligan (1982), for instance, provided a theoretical basis for interpreting gender differences in behavior through variations in moral development. Her main argument is that men and women on average approach moral problems differently: while women are more likely to evaluate moral problems under the perspective of how certain actions influence interpersonal relationships and the well-being of others, men have a greater tendency to focus on justice and individual rights. This leads to a fundamental conflict in social dilemmas such as the voluntary provision of public goods. On the one hand, free-riding is the right of each individual and therefore justifiable. On the other hand, it harms others who do not free-ride. Following Gilligan's arguments, women are expected to be more cooperative and therefore to free-ride less frequently than men. The theory also suggests that women make a greater effort to adapt their behavior to the behavior of others. In public goods games this implies being conditionally cooperative, i.e. to cooperate when others cooperate as well.

The results from our experiment show that women contribute significantly larger amounts unconditionally. Moreover, they are not only more optimistic about the contribution behavior of their group members but also have a significantly different conditional contribution pattern as obtained from the contribution schedule. Precisely, we find that women are more frequently classified as conditionally cooperative, whereas men are more often free-riders. Together, belief and type distribution fully explain the observed gender gap in unconditional contributions. When checking these results for robustness by going back to a

set of studies that allow analyzing the same research questions, we confirm that the difference in cooperative preferences between men and women is very robust. However, the finding that women have more optimistic beliefs is less universal. Hence, variations in beliefs – probably induced by subtle cues or subtle differences in experimental designs – can explain why women in some studies contribute less than men unconditionally, although they have in general a stronger preference for cooperation.

The rest of the chapter is structured as follows: In Section 5.2 we describe the design of our experiment. In Section 5.3 we present and discuss our results. Section 5.4 includes the robustness check with data from other studies. Finally, our conclusion follows in Section 5.5.

5.2 Experimental design

5.2.1 One-shot public goods game

To investigate both cooperative behavior and underlying preferences for cooperation, we use the design of the one-shot public goods experiment developed by Fischbacher et al. (2001) and, additionally, ask for beliefs. Hence, the design consists of three stages: (i) an *unconditional contribution,* (ii) a *contribution schedule* and (iii) the elicitation of subject's *belief* about others' average unconditional contribution.[137]

All subjects are randomly matched into groups of four members. Regarding the first stage, subjects receive an initial endowment of 20 experimental points and have to decide simultaneously on how to spend it.[138] Subjects can either keep the points for themselves or invest them into a public good. The invested amount, an integer that satisfies $0 \leq c_i \leq 20$, is henceforth referred to as the *unconditional contribution.* The sum of all contributions to the public good is multiplied by 1.6 and divided equally among all group members. This leads to the following linear payoff function for subject i:

$$\pi_i = 20 - c_i + 0.4 \sum_{j=1}^{4} c_j \tag{5.1}$$

where c_i denotes the contribution of subject i and the sum of c_j denotes the contributions of all group members to the public good. The marginal per capita return (MPCR) from investing

[137] See also Fischbacher and Gächter (2010) or Gächter and Renner (2010) for the elicitation of beliefs in such a setup.
[138] Each experimental point earned in stage 1 or 2 was later exchanged for 0.33 euro.

in the public good is 0.4. Hence, from an individual perspective, free-riding (i.e. $c_i = 0$) is a dominant strategy for every subject. Since the sum of marginal returns is larger than 1, however, contributing the whole endowment is the optimal choice from a social perspective. To avoid any strategic effects, the decision is taken only once.

At the second stage, each subject has to fill in a contribution schedule. It includes all possible average contributions of the three other players within a group, ranging from 0 to 20 points. For any of these 21 possible averages, subjects are asked to indicate how much they would contribute to the public good (strategy method). These contributions are referred to as c*onditional contributions* as they state how much a subject is willing to contribute conditional on the average contribution of her three group members.

Both the unconditional and the conditional contributions are potentially payoff relevant. Incentive compatibility is ensured with the following mechanism: In each group one group member is randomly determined by the roll of a four-sided dice.[139] For this group member the conditional contribution is relevant, whereas for the other three group members the unconditional contributions are relevant. More specifically, the three unconditional contributions within a group and the corresponding conditional contribution (for the specific average of the three unconditional contributions) determine the sum of money contributed to the public good. Individual earnings can then be calculated according to equation (5.1).

In the third stage, we elicit the beliefs by asking subjects how much they guess their group members have contributed unconditionally, on average (rounded to integers). Note that the guessing stage was not described in the instructions to avoid any behavioral biases of the unconditional contribution. Like in Fischbacher and Gächter (2010) and Gächter and Renner (2010) we pay subjects for the accuracy of their guesses to create real monetary incentives. We implement the following payment schedule: In case the guess of a subject completely fits the average unconditional contribution of her group members, she is rewarded by 9 points. If it differs by one point (two points) the reward amounts to 6 (3) points. Any deviation larger than two points from the true average contribution level leads to zero earnings in this stage.

[139] At the beginning of the experiment, each group member is assigned a number from one to four. After all decisions were taken, one of the participants is randomly selected by the computer. The selected participant then rolls the dice monitored by the experimenter. This procedure ensures that the participants have no doubt in the random choice of the group member for whom the conditional contribution counts.

5.2.2 Procedure

We conducted the computer-based experiments in October 2009 and March 2010 at the experimental laboratory MELESSA of the University of Munich, using the experimental software z-Tree (Fischbacher, 2007) and the organizational software Orsee (Greiner, 2004). A total of 144 undergraduate students (89 women and 55 men) from all disciplines except economics participated in six sessions with 24 subjects each. Unlike most previous studies we did not impose a special gender composition to avoid that the results are influenced by psychological effects. There were two further parts in the experiment, but these parts took place after the public goods game and were both unrelated to it.[140] In order to avoid any effects from earnings on subsequent behavior, all decisions and results of the different parts were only revealed at the end of the entire experiment. The sessions lasted up to 1½ hours, and the average payoff was 16.98 euro, including a show-up payment of 4.00 euro. The average payoff from the public goods game was 8.66 euro.

At the beginning of each experimental session, subjects received written instructions (see Appendix 5B). The instructions were read aloud and the opportunity to ask questions in private was given. To ensure that all subjects correctly understood the procedure, they had to pass through some computerized exercises, where profits for different contribution levels in the game had to be computed. We started the public goods game only after the exercises were passed successfully. Directly before payment, subjects answered a post-experimental questionnaire including some questions related to socio-economic factors. Among those was the question of a subject's gender. Except for this question, gender is never mentioned or made salient in the experiment. In fact, our post-experimental questionnaires generally contain the question for gender. No subject could have guessed that the interest of our study is in gender differences in cooperation. Finally, subjects were paid in private.

5.3 Results

5.3.1 Gender and unconditional cooperation

Starting with the unconditional contribution, we find that our subjects contribute on average 6.75 points (33.8% of their endowment) to the public good. Likewise they expect their group members to contribute on average 7.24 points (36.2%). These levels correspond well with

[140] For details see Kocher et al. (2011) or Chapter 4. Note that Chapter 4 and Chapter 5 use data from the same experiment.

previous findings in German-speaking countries (e.g. Fischbacher et al., 2001; Fischbacher and Gächter, 2010).

As shown in Table 5.1, we find considerable differences between male and female participants. The unconditional contributions of women are significantly higher than those of men (Mann-Whitney U-test, $p < 0.02$, $N = 144$). On average, women contribute 7.60 points (38.0% of their endowment) to the public good, whereas men contribute only 5.38 points (26.9%). This corresponds to a difference of approximately 10% of the endowment. It has to be added however, that women on average also believe in a higher contribution by others than men. The corresponding guesses are 7.73 (38.7%) and 6.45 (32.3%) points, respectively. This difference is weakly significant according to a Mann-Whitney U-test ($p < 0.06$). Interestingly, on average only men contribute slightly less than they expect their group members to contribute.

Result 5.1. *Women's unconditional contributions are significantly higher than men's unconditional contributions. Women also hold a more optimistic belief about their group members' contribution levels than men.*

Table 5.1: Descriptive statistics of the experiment (N = 144)

	Unconditional contribution	Guessed contribution by others
All subjects	*6.75 (33.8%)*	*7.24 (36.2%)*
Men	5.38 (26.9%)	6.45 (32.3%)
Women	7.60 (38.0%)	7.73 (38.7%)
H0: No difference between men and women (Mann-Whitney U-test (p-value))	<0.02	<0.06

The slight self-serving bias, i.e. expecting others to contribute more than oneself on average, is consistent with the findings of Fischbacher and Gächter (2010) and is caused to a large degree by free-riders (more on this in Section 5.3.3). The underlying reasons for the observed gender difference in unconditional contributions are not immediately clear. It seems to be the case that the differing beliefs play a crucial role. However, it could also be the case that men and women just have different preferences for cooperation. Do men contribute as much as women if their beliefs are the same? Our first approach to analyze this question is the following tobit regression (see Table 5.2). In line with the results of previous studies we find the belief to have a positive and highly significant impact on the unconditional contribution level. The more an individual expects others to contribute, the more s(he) also contributes

herself/himself. However, even if we control for the belief, women contribute more than men (at least weakly significantly), suggesting that differences in the belief are not the only explanation for the observed gender gap. Finally, we also include an interaction term for gender and beliefs but find no evidence that an increase in the belief has different impacts on unconditional contributions across sexes.

Result 5.2. *Subject's unconditional contribution increases in the belief about others' contributions. Controlling for beliefs, women still contribute significantly more than men.*

Table 5.2: Explaining unconditional contributions

	Model 1		Model 2	
	Coef.	p-value	Coef.	p-value
Belief	1.41***	0.00	1.58***	0.00
Woman	1.57*	0.10	3.60*	0.08
Belief * Woman	-	-	-0.27	0.26
Constant	-5.46***	0.00	-6.67***	0.00
N	144	-	144	-

Notes: Censored tobit regressions. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level.

5.3.2 Gender and cooperative preferences

Following Fischbacher et al. (2001), we categorize subjects into four different types of contributors based on their submitted conditional contribution schedule. A subject is classified as *conditional cooperator* if either her conditional contribution increases weakly monotonically with the average contribution of the other group members or if the relationship between her own and the others' average contributions is positive and significant at the 1% significance level based on the Spearman rank correlation coefficient (see the classification used in Fischbacher et al., 2001 and Fischbacher and Gächter, 2010). *Hump-shaped contributors*, sometimes also called triangle contributors, are subjects who show weakly monotonically increasing contributions (or increasing with a Spearman rank correlation coefficient at the 1% significance level) up to a given level of others' contributions (the inflection point); above that level their conditional contributions decrease based on reversed classification. A *free-rider* is a subject who has a conditional contribution of zero for all levels of the other members' average contributions. Finally, those who cannot be categorized into any of the above groups are referred to as *others*.

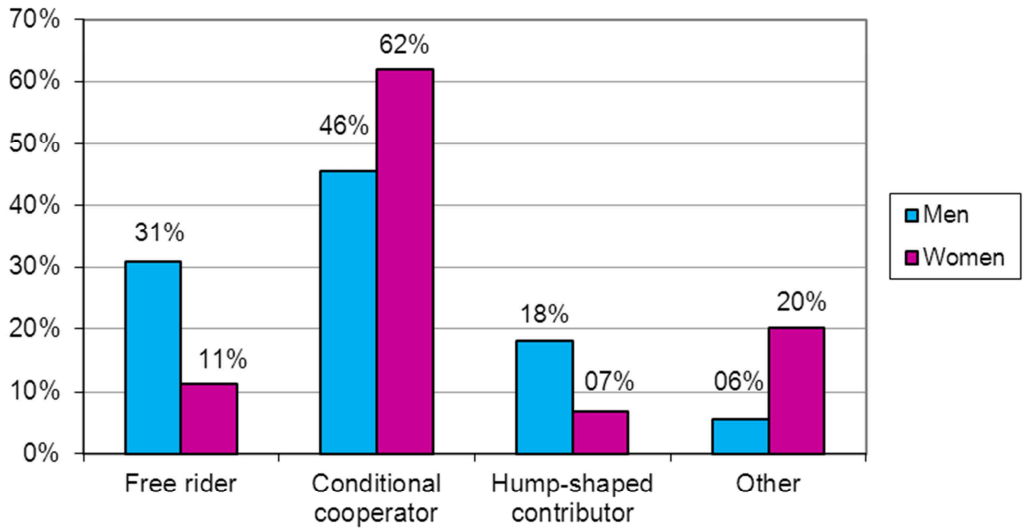
We find that overall 18.8% of our subjects can be classified as free-riders, 55.5% as conditional cooperators, 11.1% as hump-shaped and 14.6% as others. This distribution is very similar to the proportions reported in other studies about conditional cooperation, e.g. Fischbacher et al. (2001) and Kocher et al. (2008).

As it can be seen in Figure 5.1, we find again substantial differences between men and women. 61.8% of the women, but only 45.5% of the men can be classified as conditional cooperators. On the other hand, 30.9% of the men but only 11.2% of the women are free-riders. Men are also more likely to be classified as hump-shaped contributors (18.2% vs. 6.7%), whereas non-interpretable contribution patterns are shown more often by women (5.5% vs. 20.2%). The distribution of types differs significantly between sexes at the 1% level (Fisher's exact test).

Thus, we have convincing evidence that subject's basic inclination to cooperate differs across sexes. Women have a much more common preference for conditional cooperation than men. This results is consistent with the theory by Gilligan (1982) discussed in the introduction.

Result 5.3. *Women can much more frequently be classified as conditionally cooperative. On the contrary men are more often free-riders.*

Figure 5.1: Distribution of types by gender



5.3.3 The combined effect of beliefs and cooperative preferences

The previous results raise the question about connections between the gender differences found in the unconditional contribution stage and in the conditional contribution stage. For example, it could be the case that men are in general less cooperative with respect to unconditional contributions. Then we would expect the average unconditional contribution of male subjects to be lower than the one of female subjects for each classified type (with exception of the free-riders). This is, however, not the case as Table 5.3 illustrates. We find no significant differences (according to Mann-Whitney U-tests) between men and women *within* any particular category. The similarity in the unconditional contributions of men and women becomes particularly apparent in the group of the conditional cooperators. In this group, men contribute on average only 0.02 points less unconditionally than women (8.12 vs. 8.14 points). Regarding beliefs, the same observation is made, and there are also no significant differences.

However, using Kruskal-Wallis tests we find significant differences regarding the unconditional contribution as well as the guessed contribution of others *between* the types for both men and women. It is not surprising that free-riders exhibit the lowest unconditional contributions of all types. On average, male and female free-riders contribute 1.41 and 0.50 points, respectively. On the other hand, they are also most pessimistic about the cooperativeness of others, with average guessed contributions of only 3.88 and 4.80 points, respectively. With respect to the actual level of 6.75 points it can be inferred that free-riders clearly underestimate average unconditional contributions of others. In contrast, all other types tend to overestimate the cooperativeness of others. Conditional cooperators contribute on average 8.12 (men) or 8.14 points (women) and expect contributions of others of 8.00 and 7.82 points, respectively. Hence, on average, they invest (at least) as much into the public good as they expect the other group members to contribute. Unlike one might expect, they do not seem to demand a strategic uncertainty premium or exhibit a self-serving bias. The average unconditional contributions of male and female hump-shaped contributors are 4.90 and 8.83 points respectively with guessed contributions of others of 6.80 and 9.17 points. Finally, subjects classified as *others* contribute on average 6.67 and 9.44 points unconditionally and expect others to contribute 7.00 and 8.61 points, respectively.

Result 5.4. *There are no significant differences between men and women within each type category.*

Table 5.3: Unconditional contributions and beliefs by gender and type

Type of subject	Number of subjects		Unconditional contribution			Guessed contribution by others		
	Men	Women	Men	Women		Men	Women	
All types	*55*	*89*	*5.38*	*7.60*	**	*6.45*	*7.73*	*
Free-rider	17	10	1.41	0.50	°	3.88	4.80	°
Conditional cooperator	25	55	8.12	8.14	°	8.00	7.82	°
Hump-shaped	10	6	4.90	8.83	°	6.80	9.17	°
Other	3	18	6.67	9.44	°	7.00	8.61	°
H0: No difference between types (Kruskal-Wallis test (p-value))			<0.01	<0.01		<0.05	<0.05	

Note: *** Difference significant at 1% level, ** significant at 5% level, * significant at 10% level, ° not significant (based on Mann-Whitney U-tests).

Table 5.4 extends the regressions of Table 5.2 by subject's type. We take *free-rider* as the base category and find, not surprisingly, that all other types unconditionally contribute significantly larger amounts. Moreover, the coefficient for the belief is again highly significant. Most importantly, however, our gender dummy becomes far from being significant (p = 0.89 in model 1) suggesting that we have now fully explained the gender gap in the unconditional contribution. To be precise, the difference in the unconditional contribution between men and women is due to *both* differences in the belief *and* differences in the distribution of cooperation types across gender.

Result 5.5. *The gender gap in the unconditional contribution is caused both by differences in the belief and by differences in the underlying cooperative preference.*

Table 5.4: Explaining unconditional contributions (2)

	Model 1		Model 2	
	Coef.	p-value	Coef.	p-value
Belief	1.25***	0.00	1.33***	0.00
Woman	-1.34	0.89	-1.14	0.57
Belief * Woman	-	-	-0.13	0.57
Type:				
Conditional cooperator	7.65***	0.00	7.57***	0.00
Hump-shaped	5.31***	0.00	5.28***	0.00
Other	8.29***	0.00	8.25***	0.00
Constant	-9.56***	0.00	-9.13***	0.00
N	144	-	144	-

Notes: Censored tobit regressions. *** Significant at 1% level, ** significant at 5% level, * significant at 10% level. Type: Base category is free-rider.

5.4 Robustness check: Data from related studies

In the light of contradictory results in previous studies on gender and cooperation, it is natural to ask to what extent our experimental results are robust. We have a quite homogenous subject pool of 144, mostly German undergraduate students, and it could of course be that our results are subject-pool- or country-specific. Therefore, we use our empirical results from the previous section and see how far we can corroborate them when taking data sets of recent public good experiments that some of the authors of this chapter conducted for other reasons than to study gender effects. The main finding of this robustness exercise is that our central result – women are relatively more frequently conditional cooperators, whereas men are relatively more frequently free-riders – is overwhelmingly confirmed.

More precisely, in the following we present the results of three different studies that were conducted in five different countries, namely Austria, Colombia, Japan, the USA, and Vietnam. Kocher et al. (2008) investigated conditional cooperation in Austria, Japan and the USA. Martinsson et al. (2009) conducted the experiments in Colombia and finally Martinsson and Villegas-Palacio (2010) provided the data from Vietnam. It is important to note that all of these studies are not completely identical but broadly comparable with our main experiment from the previous section. In particular, we checked that the data and the results of these studies were not affected by potential order effects, in connection with treatment variation in some of the experiments, and that subject's anonymity was preserved throughout the whole experiment.

Table 5.5: Number of observations separated by gender and country[141]

Country	Men	Women	All
Austria	23	12	35
Colombia	68	27	95
Japan	30	6	36
USA	19	17	36
Vietnam	28	20	48
All	*168*	*82*	*250*

[141] We have no information about the gender of one participant in both Austria and Colombia. These two persons are excluded from the analysis. The study in Japan was conducted at the Technical University of Tokyo which explains the very low number of female participants.

Table 5.5 provides an overview of the number of observations these studies rely on, separated by gender and country. It shows that the number of experimental participants within each country ranges from 35 to 95, with a number of female participants between 6 and 27.

As the number of participants is too small to draw robust conclusions for each country separately and as it is not our aim to study culture- or country-specific effects, we do not focus on a given country in the following. In contrast, we rather pool the data to obtain a large data set that allows us to assess how universal our results are. Nevertheless, more detailed information on country-specific data is included in Appendix 5A.

Combining all data we find again considerable differences in the conditional contribution patterns of men and women. Table 5.6 illustrates that 65% of women can be classified as conditional cooperators, but only 54% of men. On the other hand, 20% of men behave as free-riders, but only 9% of women. Regarding the other categories note that in particular the category of hump-shaped contributors is small. Overall, the distribution of types differs significantly between men and women (Fisher's exact test, $p = 0.05$). Comparing these results with the data from our main experiment we find that gender differences in the type distribution are a very robust result.

Result 5.6. *Gender differences regarding the type distribution – men being relatively more frequently free-riders and women being relatively more frequently conditional cooperators - are a robust finding.*

Table 5.6: Fraction and absolute number of types in pooled data

Type of subject	Men	Women
Free-rider	0.20	0.09
	(33)	(7)
Conditional cooperator	0.54	0.65
	(90)	(53)
Hump-shaped	0.10	0.05
	(16)	(4)
Other	0.17	0.22
	(29)	(18)
All	*1.00*	*1.00*
	(168)	*(82)*

Finally, we cast a brief glance at unconditional contributions and beliefs in these studies. Note that the results are less comparable to our experiment. The two main reasons are that (i)

beliefs are not elicited in the study of Kocher et al. (2008), and (ii) the studies of Martinsson et al. (2009) and Martinsson and Villegas-Palacio (2010) use unconditional contributions and beliefs that range from 0 to 60 points.

If we compute average unconditional contributions over all studies (dividing contributions in the latter two studies by three) we find that women do not contribute significantly larger amounts than men. On the contrary, average contributions of women (6.59) are even below those of men (7.51).[142] This sounds surprising when compared to the fact that women have, on average, the more cooperative underlying preference. However, this puzzle can be explained by the belief. Indeed, if we look only at the data of the studies by Martinsson et al. (2009) and Martinsson and Villegas-Palacio (2010), for which we have elicited beliefs, i.e., if we focus on Colombia and Vietnam, we find evidence that the belief influences the differences between studies. For example, in Colombia men contribute slightly more unconditionally although they are more often classified as free-riders and less frequently as conditionally cooperative. However, they also hold, on average, more optimistic beliefs in contrast to our main experiment. In Vietnam, the gender difference in the beliefs is smaller and, as a consequence, the result that men contribute larger amounts vanishes. To sum up, beliefs vary over studies and seem to drive different observations with regard to the unconditional contribution although the distribution of conditional type patterns and the resulting gender differences in cooperative preferences are astonishingly robust.

5.5 Conclusion

In our experiment we try to contribute to an ongoing debate on gender differences in cooperativeness. Experimental evidence so far was mixed and inconclusive. Our experimental design provides an anatomy of cooperative behavior and thus allows explaining conflicting evidence in existing studies. More precisely, we use a one-shot public goods game based on the Fischbacher et al. (2001) procedure. In detail, we elicit (i) a subject's unconditional contribution, (ii) her conditional contribution for each potential average contribution level of the other group members and, additionally, (iii) her belief about the average unconditional contribution of her group members. With this setup we are able to disentangle two potential

[142] The difference is not significant. The same result can be obtained by focusing only on the data of Kocher et al. (2008). Here, average unconditional contributions are 6.71 (women) and 7.99 (men) which is not a significant difference either.

determinants of a gender gap: differences in the underlying cooperative preferences and differences in the belief about group members' behavior.

Our results indicate that both aspects matter. We find that, on average, women contribute unconditionally 10% more to the public good than men. However, they also hold a more optimistic belief about the contributions of their group members and show a more cooperative preference as they can be more often classified as conditionally cooperative and less often as free-riders than men. We find that controlling for both aspects fully explains the gender difference in the unconditional contribution.

To control for the robustness of our results we compare them with data from similar experimental studies conducted at different places in the world. The overview shows that differences in the type distribution with women being more often classified as conditionally cooperative and men more often as free-riders are a very robust finding. However, it is not always the case that there is also a difference in the unconditional contribution. This could be caused by differences in beliefs across different studies.

Having our central results in mind, at least three aspects deserve more attention in the future. First, determinants of beliefs are not well-understood, in general. However, differences in beliefs contribute to the gender gap in cooperative behavior, and thus more work is needed that analyzes the foundations of beliefs. Second, country- or culture-specific effects as well as the context of the social dilemma (framing effects, etc.) could have a strong influence on the gender gap through beliefs or through other sources. Third, our results have clear implications for repeated public goods games. Experiments in the future can test whether the patterns of behavior across gender in repeated public goods games are well-predicted based on the connection of beliefs and underlying cooperative preferences.

Appendix

5A Further results

Table 5A.1: Distribution of types, unconditional contributions and beliefs by gender and country

Type of subject	Austria		Colombia		Japan		USA		Vietnam	
	Men	Women	Men	Women	Men	Women	Men	Women	Men	Women
Free-rider	0.26 (6)	0.17 (2)	0.19 (13)	0.04 (1)	0.33 (10)	0.50 (3)	0.16 (3)	0.00 (0)	0.04 (1)	0.05 (1)
Conditional cooperator	0.52 (12)	0.50 (6)	0.56 (38)	0.67 (18)	0.43 (13)	0.33 (2)	0.74 (14)	0.94 (16)	0.46 (13)	0.55 (11)
Hump-shaped	0.13 (3)	0.08 (1)	0.09 (6)	0.07 (2)	0.13 (4)	0.00 (0)	0.00 (0)	0.00 (0)	0.11 (3)	0.05 (1)
Other	0.09 (2)	0.25 (3)	0.16 (11)	0.22 (6)	0.17 (3)	0.17 (1)	0.11 (2)	0.06 (1)	0.39 (11)	0.35 (7)
All	*1.00 (23)*	*1.00 (12)*	*1.00 (68)*	*1.00 (27)*	*1.00 (30)*	*1.00 (6)*	*1.00 (19)*	*1.00 (17)*	*1.00 (28)*	*1.00 (20)*
Mean unconditional contribution	7.87	5.50	8.00	7.42	7.93	5.50	8.21	8.00	5.07	5.30
Mean belief	-	-	6.17	4.91	-	-	-	-	6.42	6.08

Table 5A.1 displays country specific data on types, unconditional contributions and beliefs. In order to get comparable results, we classify some individuals from the study of Kocher et al. (2008) differently than in the original paper. In the original study, subjects were classified as conditional cooperators only if they submitted a contribution schedule that was monotonically increasing with the average contribution of the other group members. All other studies use a slightly different classification mechanism where subjects are also classified as conditional cooperators if they have a highly significant (at the 1% level) positive Spearman rank correlation between their own and others' contributions. This classification was also used by Fischbacher et al. (2001) and Fischbacher and Gächter (2010). Hence, we reclassify one (male) individual from the USA and three individuals (two men, one woman) from Austria as conditional cooperators instead of others. The fact that only one woman but three more men are reclassified as conditionally cooperative goes against our result, which would be even stronger otherwise. Furthermore, note that unconditional contributions and beliefs in Colombia and Vietnam are divided by three as the allowed contribution in these experiments ranged from 0 to 60 points.

5B Experimental instructions (originally in German)[143]

Welcome to the experiment and thank you for participating!

Please do not talk to other participants.

General

This is an experiment on economic decision making. You will earn "real" money that will be paid out to you in cash at the end of the experiment. During the experiment all participants will be asked to make decisions. Your decisions and the decisions of other participants determine your earnings from the experiment according to the following rules.

The experiment will last two hours. If you have any questions or if anything is unclear, please raise your hand, and one of the experimenters will come to you and answer your questions privately.

During the experiment a part of your earnings will be calculated in **points**. At the end of the experiment all points that you earn will be converted into euro at the exchange rate of

1 point = 0.33 euro (3 points = 1 euro).

In the interest of clarity, we will only use male terms in the instructions.

Anonymity

You will learn neither during nor after the experiment, with whom you interact(ed) in the experiment. The other participants will neither during nor after the experiment learn, how much you earn(ed). We never link names and data from experiments. At the end of the experiment you will be asked to sign a receipt regarding your earnings which serves only as a proof for our sponsor. The latter does not receive any other data from the experiment.

Means of help

You will find a pen at your table which you, please, leave behind on the table when the experiment is over. While you make your decisions, a clock will run down at the top of your computer screen. This clock will give you an orientation how long you should need to make your decisions. But you can nevertheless exceed this time. The input screens will not be dismissed once time is over. However, the pure output screens (here you do not have to make a decision) will be dismissed.

Experiment

The experiment consists of three parts. You will receive instructions for a part after the previous part has ended. The parts of the experiment are completely independent; decisions in one part have no consequences for your earnings in later parts. The sum of earnings from the different parts will constitute your total earnings from the experiment.

[143] The experimental instructions equal those of Chapter 4 as both studies emerged from the same experiment.

Part I

The decision situation

The basic decision situation will be explained to you in the following. Afterwards you will find control questions on the screen which should raise your familiarity with the decision situation.

You will be a member of a group consisting of **4 people**. Each group member has to decide on the allocation of 20 points. You can put these 20 points into your **private account** or you can put them **fully or partially** into a **group account**. Each point you do not put into the group account will automatically remain in your private account.

Your income from the private account:

You will earn one point for each point you put into your private account. For example, if you put 20 points into your private account (and therefore do not put anything into the group account) your income will amount to exactly 20 points out of your private account. If you put 6 points into your private account, your income from this account will be 6 points. No one except you earns something from your private account.

Your income from the group account:

Each group member will profit equally from the amount you put into the group account. On the other hand, you will also get a payoff from the other group members' in-payments into the group account. The income for each group member out of the group account will be determined as follows:

Income from group account =
Sum of all group members' contributions to the group account $\times 0.4$

If, for example, the sum of all group members' contributions to the group account is 60 points, then you and the other members of your group each earn 60 x 0.4 = 24 points out of the group account. If the four group members contribute a total of 10 points to the group account, you and the other members of your group each earn 10 x 0.4 = 4 points out of the group account.

Total income:

Your total income is the sum of your income from your private account and that from the group account:

Income from your private account (= 20 – contribution to group account)
+ Income from group account (= 0,4 $\times$ *sum of contributions to group account)*
= Total income

Before we proceed, please try to solve the control questions on your screen. If you want to compute something, you can use the Windows calculator by clicking on the respective symbol on your screen.

Procedure of Part I

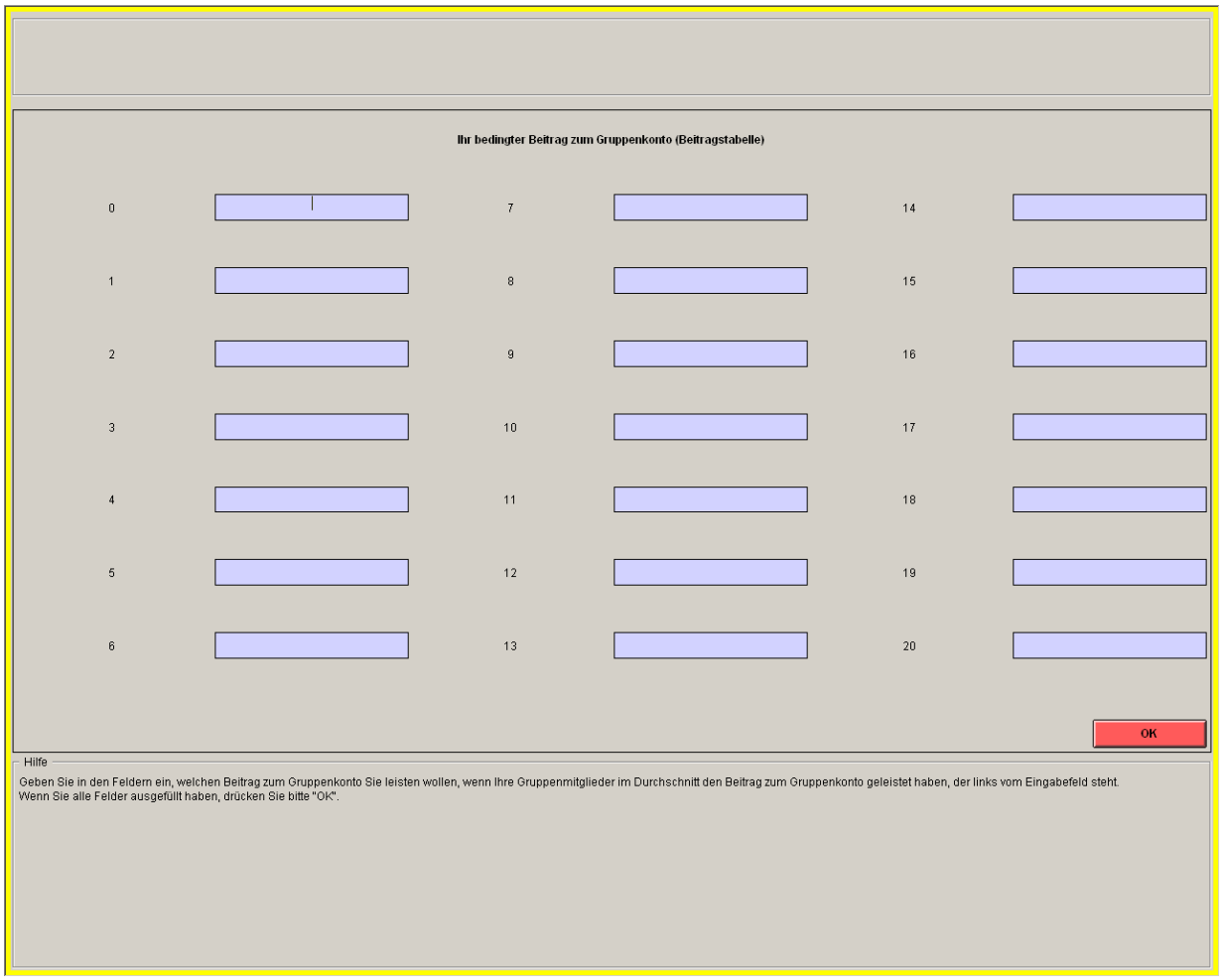
Part I includes the decision situation just described to you. The decisions in Part I will only be made **once**.

On the first screen you will be informed about your **group membership number**. This number will be of relevance later on. If you have taken note of the number, please click "next".

Then you have to make your decisions. As you know, you will have 20 points at your disposal. You can put them into your private account or you can put them into the group account. Each group member has to make **two types** of contribution decisions which we will refer to below as the **unconditional contribution** and the **contribution table**.

- In the **unconditional contribution** case you decide how many of the 20 points you want to put into the group account. Please insert your unconditional contribution in the respective box on your screen. You can insert integer numbers only. Your contribution to the private account is determined automatically by the difference between 20 and your contribution to the group account. After you have chosen your unconditional contribution, please click "next".

- On the next screen you are asked to fill in a **contribution table**. In the contribution table you indicate **how much <u>you</u> want to contribute to the group account for each possible average contribution of the <u>other</u> group members** (rounded to the next integer). Thus, you can condition your contribution on the other group members' average contribution. The contribution table looks as follows:



The numbers in each of the left columns are the possible (rounded) average contributions of the **other** group members to the group account. This means, they represent the amount each of the other group members' has put into the group account on average. You simply have to insert into the input boxes how many points you want to contribute to the group account – conditional on the indicated average contribution. **You have to make an entry into each input box**. For example, you will have to indicate how much you contribute to the group account if the others contribute 0 points to the group account on average, how much you contribute if the others contribute 1, 2, or 3 points on average, etc. You can insert any integer numbers from 0 to 20 in each input box. Once you have made an entry in each input box, please click "OK".

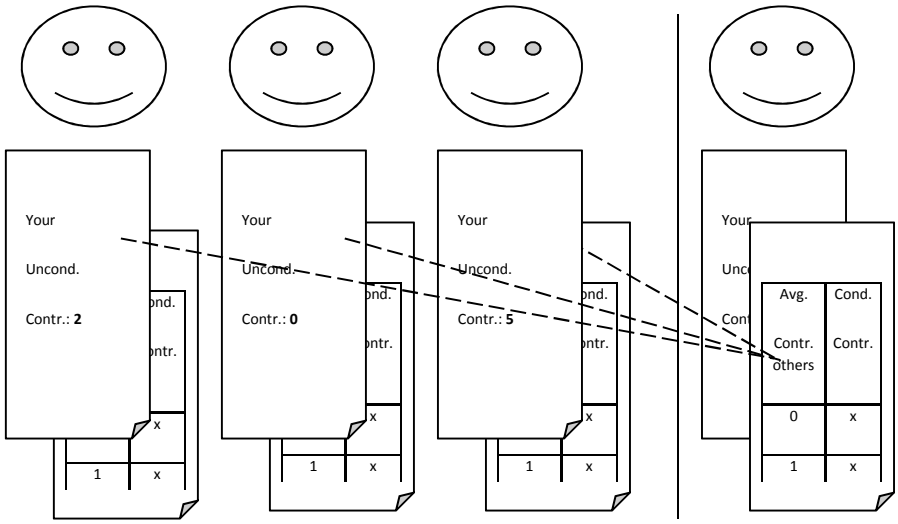
After all participants of the experiment have made an unconditional contribution and have filled in their contribution table, a random mechanism will select a group member from every group. Only **the contribution table** will be the payoff-relevant decision for the **randomly determined subject**. Only the **unconditional contribution** will be the payoff-relevant decision for the **other three group members** not selected by the random mechanism. You obviously do not know whether the random mechanism will select you when you make your unconditional contribution and when you fill in the contribution table. You will therefore have to think carefully about both types of decisions because both can become relevant for you. Two examples should make this clear.

Example 1: Assume that **the random mechanism selects you. This implies that your relevant decision will be your contribution table**. The unconditional contribution is the relevant decision for the other three group members. Assume they made unconditional contributions of 0, 2, and 5 points. The average rounded contribution of these three group members, therefore, is 2 points ((0+2+5)/3 = 2.33).
If you indicated in your contribution table that you will contribute 1 point to the group account if the others contribute 2 points on average, then the total contribution to the group account is given by 0+2+5+1=8 points. All group members, therefore, earn 0.4×8=3.2 points out of the group account plus their respective income from the private account.
If, instead, you indicated in your contribution table that you would contribute 19 points if the others contribute two points on average, then the total contribution of the group to the group account is given by 0+2+5+19=26. All group members therefore earn 0.4×26=10.4 points out of the group account plus their respective income from the private account.

Example 2: Assume that **the random mechanism did not select you,** implying that **the unconditional contribution is taken as the payoff-relevant decision** for you and two other group members. Assume your unconditional contribution to the group account is 16 points and those of the other two group members are 18 and 20 points. The average unconditional contribution of you and the other two group members, therefore, is 18 points (= (16+18+20)/3).
If the group member whom the random mechanism selected indicates in her contribution table that she will contribute 1 point to the group account if the other three group members contribute on average 18 points, then the total contribution to the group account is given by 16+18+20+1=55 points. All group members will therefore earn 0.4×55=22 points out of the group account plus their respective income from the private account.
If, instead, the randomly selected group member indicates in her contribution table that she contributes 19 points to the group account if the others contribute on average 18 points, then the total contribution to the group account is given by 16+18+20+19=73 points. All group members will therefore earn 0.4×73=29.2 points out of the group account plus their respective income from the private account.

The random selection of the participants will be implemented as follows. A randomly selected participant will throw a 4-sided dice **after** all participants have made their unconditional contribution and have filled in their contribution table. She enters the thrown number into the computer thereby being monitored by the experimenter who confirms the correctness of the entry by password. The thrown number will then be compared with the group membership number, which was shown to you on the first screen. If the thrown number equals your group membership number, then your contribution table is payoff-relevant for you and the unconditional contribution is payoff-relevant for the other three group members. Otherwise, your unconditional contribution is the relevant decision for you.

The following figure visualizes the situation in example 1. You are the person on the right side with group membership number 3. Number 3 was thrown and therefore your conditional contribution is payoff-relevant. For the other three group members the unconditional contribution is payoff-relevant.



You will make all your decisions only **once**. After the end of Part I you will get the instructions of Part II. How much you have earned in Part I will be revealed at the end of the experiment.

Part II

(omitted because of irrelevance)

Part III

(omitted because of irrelevance)

After Part III is finished we will ask you to fill in a short questionnaire on the screen. Afterwards you will learn for each part separately how much you have earned. Then the experiment ends. There are neither more parts nor any repetitions. Finally, you will be informed about your total earnings from the experiment and paid out.

Bibliography

Ai, C., Norton, E. (2003), Interaction terms in logit and probit models. *Economics Letters* 80: 123-129.

Alm, J., McClelland, G. H., Schulze, W. D. (1999), Changing the social norm of tax compliance by voting. *Kyklos* 52: 141-171.

Anderson, C. M., Putterman, L. (2006), Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior* 54: 1-24.

Anderson, L. R., Mellor, J. M., Milyo, J. (2004), Social capital and contributions in a public-goods experiment. *American Economic Review* 94: 373-376.

Andreoni, J. (1988), Why free ride? Strategies and learning in public goods experiments. *Journal of Public Economics* 37: 291-304.

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

Andreoni, J. (1995), Cooperation in public goods experiments: Kindness or confusion? *American Economic Review* 85: 891-904.

Andreoni, J., Croson, R. T. A. (2008), Partners versus strangers: Random rematching in public goods experiments. In: Plott, C. R., Smith, V. L. (Eds.). Handbook of Experimental Economics Results, Amsterdam, North-Holland.

Andreoni, J., Gee, L. (2011), Gun for hire: Does delegated enforcement crowd out peer punishment in giving to public goods? NBER Working Papers 17033, National Bureau of Economic Research, Inc.

Andreoni, J., Harbaugh, W., Vesterlund, L. (2003), The carrot or the stick: Rewards, punishments and cooperation. *American Economic Review* 93: 893-902.

Andreoni, J., Petrie, R. (2008), Beauty, gender and stereotypes: Evidence from laboratory experiments. *Journal of Economic Psychology* 29: 73-93.

Arellano, M., Bond, S. (1991), Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277-297.

Bahry, D., Wilson, R. K. (2004), Trust in transitional societies: Experimental results from Russia. Working Paper, Rice University.

Baldassarri, D., Grossman, G. (2011), Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences* 108: 11023-11027.

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

Ben-Ner, A., Putterman, L. (2001), Trusting and trustworthiness. *Boston University Law Review* 81: 523-551.

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

Bochet, O., Page, T., Putterman, L. (2006), Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior and Organization* 60: 11-26.

Bochet, O., Putterman, L. (2009), Not just babble: Opening the black box of communication in a voluntary contribution experiment. *European Economic Review* 53: 309-326.

Bohnet, I., Greig, F., Herrmann, B., Zeckhauser, R. (2008), Betrayal aversion. *American Economic Review* 98: 294-310.

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

Botelho, A., Harrison, G. W., Pinto, L., Rutström, E. E. (2005), Social norms and social choice. Working Papers 30, Núcleo de Investigação em Microeconomia Aplicada (NIMA), Universidade do Minho.

Botelho, A., Harrison, G. W., Costa Pinto, L. M., Rutström, E. E. (2009), Testing static game theory with dynamic experiments: A case study of public goods. *Games and Economic Behavior* 67: 253-265.

Bowles, S., Gintis, H. (2002), Social capital and community governance. *Economic Journal* 112: 419-436.

Brandts, J., Schram, A. (2001), Cooperation and noise in public goods experiments: Applying the contribution function approach. *Journal of Public Economics* 79: 399-427.

Brosig, J. (2002), Identifying cooperative behavior: some experimental results in a prisoner's dilemma game. *Journal of Economic Behavior and Organization* 47: 275-290.

Brosig, J., Weimann, J., Ockenfels, A. (2003), The effect of communication media on cooperation. *German Economic Review* 4: 217-241.

Brown-Kruse, J., Hummels, D. (1993), Gender effects in laboratory public goods contribution: Do individuals put their money where their mouth is? *Journal of Economic Behavior and Organization* 22: 255-267.

Burks, S. V., Carpenter, J. P., Verhoogen, E. (2003), Playing both roles in the trust game. *Journal of Economic Behavior and Organization* 51: 195-216.

Burlando, R., Hey, J. D. (1997), Do Anglo-Saxons free-ride more? *Journal of Public Economics* 64: 41-60.

Cadsby, C., Maynes, E. (1998), Gender and free riding in a threshold public goods game: Experimental evidence. *Journal of Economic Behavior and Organization* 34: 603-620.

Cárdenas, J. C., Carpenter, J. (2008), Behavioral development economics: Lessons from field labs in the developing world. *Journal of Development Studies* 44: 337-364.

Cárdenas, J. C., Rodríguez, L. A., Johnson, N. (2009), Collective action for watershed management: Field experiments in Colombia and Kenya. Documentos CEDE 006649, Universidad de Los Andes-CEDE.

Carpenter, J. (2007a), The demand for punishment. *Journal of Economic Behavior and Organization* 62: 522-542.

Carpenter, J. (2007b), Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior* 60: 31-51.

Carpenter, J., Kariv, S., Schotter, A. (2010), Network architecture and mutual monitoring in public goods experiments. IZA Discussions Papers 5307, Institute for the Study of Labor.

Casari, M. (2005), On the design of peer punishment experiments. *Experimental Economics* 8: 107-115.

Casari, M., Luini, L. (2009), Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior and Organization* 71: 273-282.

Cason, T. N., Khan, F. U. (1999), A laboratory study of voluntary public goods provision with imperfect monitoring and communication. *Journal of Development Economics* 58: 533-552.

Cassar, A., Rigdon, M. L. (2011), Trust and trustworthiness in networked exchange. *Games and Economic Behavior* 71: 282-303.

Cettolin, E., Riedl, A. (2011), Partial coercion, conditional cooperation, and self-commitment in voluntary contributions to public goods. CESifo Working Paper 3556.

Charness, G., Rabin, M. (2002), Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117: 817-869.

Charness, G., Villeval, M.-C. (2009), Cooperation and competition in intergenerational experiments in the field and the laboratory. *American Economic Review* 99: 956-978.

Charness, G., Yang, C.-L. (2008), Endogenous group formation and public goods provision: Exclusion, exit, mergers, and redemption. Working Paper, University of California Santa Barbara.

Chaudhuri, A. (2011), Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics* 14: 47-83.

Chermak, J. M., Krause, K. (2002), Individual response, information, and intergenerational common pool problems. *Journal of Environmental Economics and Management* 43: 47-70.

Cinyabuguma, M., Page, T., Putterman, L. (2005), Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics* 89: 1421-1435.

Croson, R., Gneezy, U. (2009), Gender differences in preferences. *Journal of Economic Literature* 47: 448-474.

Dal Bó, P., Foster, A., Putterman, L. (2010), Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review* 100: 2205-2229.

Dawes, R. (1980), Social dilemmas. *Annual Review of Psychology* 31: 169-193.

Decker, T., Stiehler, A., Strobel, M. (2003), A comparison of punishment rules in repeated public good games. *Journal of Conflict Resolution* 47: 751-772.

Denant-Boemont, L., Masclet, D., Noussair, C. (2007), Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 133: 145-167.

Deutsch, M. (1958), Trust and suspicion. *Journal of Conflict Resolution* 2: 265-279.

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

Durlauf, S. N., Fafchamps, M. (2005), Social capital. In: Aghion, P., Durlauf, S. N. (Eds.). Handbook of Economic Growth Vol. I., Amsterdam, North-Holland: 1639-1699.

Eckel, C. C., Grossman, P. J. (2008), Differences in the economic decisions of men and women: Experimental evidence. In: Plott, C., Smith, V. (Eds.). Handbook of Experimental Economics Results Vol. 1, New York: 509-519.

Eckel, C. C., Wilson, R. K. (2004), Is trust a risky decision? *Journal of Economic Behavior and Organization* 55: 447-465.

Egas, M., Riedl, A. (2008), The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B – Biological Sciences* 275: 871-878.

Ertan, A., Page, T., Putterman, L. (2009), Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review* 53: 495-511.

Falk, A., Fehr, E., Fischbacher, U. (2005), Driving forces behind informal sanctions. *Econometrica* 73: 2017-2030.

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

Fehr, E., Fischbacher, U., Rosenbladt, B., Schupp, J., Wagner, G. G. (2002), A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative surveys. *Schmollers Jahrbuch* 122: 519-542.

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

Fehr, E., Gächter, S. (2002), Altruistic punishment in humans. *Nature* 415: 137-140.

Fehr, E., Klein, A., Schmidt, K. (2007), Fairness and Contract Design. *Econometrica* 75: 121-154.

Fehr, E., Schmidt, K. (1999), A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114: 817-868.

Feld, L. P., Tyran, J.-R. (2002), Tax evasion and voting: An experimental analysis. *Kyklos* 55: 197-221.

Fischbacher, U. (2007), z-Tree - Zurich toolbox for readymade economic experiments – experimenter's manual. *Experimental Economics* 10: 171-178.

Fischbacher, U., Gächter, S. (2010), Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review* 100: 541-556.

Fischbacher, U., Gächter, S., Fehr, E. (2001), Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters* 71: 397-404.

Frank, R. H., Gilovich, T., Regan, D. T. (1993), Does studying economics inhibit cooperation? *Journal of Economic Perspectives* 7: 159-171.

Gächter, S., Herrmann, B. (2011), The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review* 55: 193-210.

Gächter, S., Herrmann, B., Thöni, C. (2004), Trust, voluntary cooperation, and socio-economic background: Survey and experimental evidence. *Journal of Economic Behavior and Organization* 55: 505-531.

Gächter, S., Herrmann, B., Thöni, C. (2010), Culture and cooperation. *Philosophical Transactions of the Royal Society B* 365: 2651-2661.

Gächter, S., Nosenzo, D., Renner, E., Sefton, M. (2010), Who makes a good leader? Cooperativeness, optimism and leading-by-example. *Economic Inquiry*: forthcoming.

Gächter, S., Renner, E. (2010), The effects of (incentivized) belief elicitation in public good experiments. *Experimental Economics* 13: 364-377.

Gächter, S., Renner, E., Sefton, M. (2008), The long-run benefits of punishment. *Science* 322: 1510.

Gilligan, C. (1982), In a different voice. Harvard University Press, Cambridge, MA.

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

Greiner, B. (2004), An online recruitment system for economic experiments. In: Kremer, K., Macho, V. (Eds.). Forschung und wissenschaftliches Rechnen 2003, GWDG Bericht 63, Göttingen: 79-93.

Griesinger, D. W., Livingston, J. W. (1973), Toward a model of interpersonal motivation in experimental games. *Behavioral Science* 18: 173-188.

Guillen, P., Schwieren, C., Staffiero, G. (2006), Why feed the Leviathan? *Public Choice* 130: 115-128.

Gürerk, Ö., Irlenbusch, B., Rockenbach, B. (2006), The competitive advantage of sanctioning institutions. *Science* 312: 108-111.

Güth, W., Levati, M. V., Sutter, M., Van der Heijden, E. (2007), Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics* 91: 1023-1042.

Herrmann, B., Thöni, C., Gächter, S. (2008), Antisocial punishment across societies. *Science* 319: 1362-1366.

Hobbes, T. (2008 [1651]), Leviathan. Gaskin, J. C. A. (Ed.), Oxford University Press, Oxford.

Holt, C. A., Laury, S. K. (2002), Risk aversion and incentive effects. *American Economic Review* 92: 1644-1655.

Isaac, R. M., McCue, K., Plott, C. (1985), Public goods provision in an experimental environment. *Journal of Public Economics* 26: 51-74.

Isaac, R. M., Walker, J. (1988), Group size effects in public goods provision: The voluntary contributions mechanism. *Quarterly Journal of Economics* 103: 179-199.

Janssen, M., Holahan, R., Lee, A., Ostrom, E. (2010), Lab experiments for the study of social-ecological systems. *Science* 328: 613-617.

Joffily, M., Masclet, D., Noussair, C., Villeval, M.-C. (2011), Emotions, sanctions and cooperation. IZA Discussion Papers 5592, Institute for the Study of Labor.

Johansson-Stenman, O. Mahmud, M., Martinsson, P. (2011), Trust, trust games and stated trust: Evidence from rural Bangladesh. *Journal of Economic Behavior and Organization* forthcoming.

Johnson, N. D., Mislin, A. A. (2011), Trust Games: A Meta-Analysis. *Journal of Economic Psychology* 32: 865-889.

Kamei, K., Putterman, L., Tyran, J.-R. (2011), State or nature? Formal vs. informal sanctioning in the voluntary provision of public goods. Working paper 2011-3, Brown University.

Keser, C., van Winden, F. (2000), Conditional cooperators and voluntary contributions to public goods. *Scandinavian Journal of Economics* 102: 23-39.

Knack, S. (2002), Social capital and the quality of government: Evidence from the United States. *American Journal of Political Science* 46: 772-785.

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

Kocher, M. G., Cherry, T., Kroll, S., Netzer, R., Sutter, M. (2008), Conditional cooperation on three continents. *Economics Letters* 101: 175-178.

Kocher, M. G., Martinsson, P., Myrseth, K., Wollbrant, C. (2011), Strong, bold, and kind: Self-control and cooperation in social dilemmas. Working Papers in Economics 523, University of Gothenburg.

Kosfeld, M., Okada, A., Riedl, A. (2009), Institution formation in public goods games. *American Economic Review* 99: 1335-1355.

Kroll, S., Cherry, T. L., Shogren, J. F. (2007), Voting, punishment, and public goods. *Economic Inquiry* 45: 557-570.

Kube, S., Traxler, C. (2011), The interaction of legal and social norm enforcement. *Journal of Public Economic Theory* 13: 639-660.

Ledyard, J. (1995), Public goods: A survey of experimental research. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics, Princeton University Press, Princeton.

Leonard, T., Croson, R. T. A., de Oliveria, A. C. M. (2010), Social capital and public goods. *Journal of Socio-Economics* 39: 474-481.

Levati, M. V., Sutter, M., van der Heijden, E. (2007), Leading by example in a public goods experiment with heterogeneity and incomplete information. *Journal of Conflict Resolution* 51: 793-818.

Levy, D. M., Padgitt, K., Peart, S. J, Houser, D., Xiao, E. (2011), Leadership, cheap talk and really cheap talk. *Journal of Economic Behavior and Organization* 77: 40-52.

Liebrand, W. B. G. (1984), The effect of social motives, communication and group size on behaviour in an n-person multi-stage mixed-motive game. *European Journal of Social Psychology* 14: 239-264.

Maier-Rigaud, F. P., Martinsson, P. and Staffiero, G. (2010), Ostracism and the provision of a public good: Experimental Evidence. *Journal of Economic Behavior and Organization* 73: 387-395.

Markussen, T., Putterman, L., Tyran, J.-R. (2011), Self-organization for collective action: An experimental study of voting on formal, informal, and no sanction regimes. Working Paper 2011-4, Brown University.

Martinsson, P., Villegas-Palacio, C. (2010), Does disclosure crowd out cooperation? Working Papers in Economics 446, University of Gothenburg.

Martinsson, P., Villegas-Palacio, C., Wollbrant, C. (2009), Conditional cooperation and social group: Experimental results from Colombia. Working Papers in Economics 372, University of Gothenburg.

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

Masclet, D., Noussair, C., Villeval, M.-C. (2011), Threat and punishment in public good experiments. CIRANO Working Papers 2011s-08, CIRANO.

Masclet, D., Villeval, M.-C. (2008), Punishment, inequality, and welfare: A public good experiment. *Social Choice and Welfare* 31: 475-502.

Mason, C. F., Phillips, O. R., Redington, D. B. (1991), The role of gender in a non-cooperative game. *Journal of Economic Behavior and Organization* 15: 215-235.

Näf, M., Schunk, D. (2009), Once bitten, twice shy: On the causal effect of prior experiences on trusting behaviour. Working Paper, University of London, Royal Holloway.

Nikiforakis, N. (2008), Punishment and counter-punishment in public-good games: Can we really govern ourselves? *Journal of Public Economics* 92: 91-112.

Nikiforakis, N. (2010), Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior* 68: 689-702.

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

Noussair, C., Tucker, S. (2005), Combining monetary and social sanctions to promote cooperation. *Economic Inquiry* 43: 649-660.

Nowell, C., Tinkler, S. (1994), The influence of gender on the provision of a public good. *Journal of Economic Behavior and Organization* 25: 25-36.

Offerman, T., Sonnemans, J., Schram, A. (1996), Value orientations, expectations and voluntary contributions in public goods. *Economic Journal* 106: 817-845.

Ortmann, A., Tichy, L. K. (1999), Gender differences in the laboratory: Evidence from prisoner's dilemma games. *Journal of Economic Behavior and Organization* 39: 327-339.

Ostrom, E., Walker, J., Gardner, R. (1992), Covenants with and without a sword: Self-governance is possible. *American Political Science Journal* 86: 404-417.

Ostrom, E., Walker, J., Gardner, R. (1994), Rules, games, and common-pool resources. University of Michigan Press, Ann Arbor.

Page, T., Putterman, L., Unel, B. (2005), Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency. *Economic Journal* 115: 1032-1053.

Palfrey, T., Prisbrey, J. (1997), Anomalous behavior in public goods experiments: How much and why? *American Economic Review* 87: 829-846.

Park, E.-U. (2000), Warm-glow versus cold-prickle: A further experimental study of framing effects on free-riding. *Journal of Economic Behavior and Organization* 43: 405-421.

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

Putterman, L., Tyran, J.-R., Kamei, K. (2011), Public goods and voting on formal sanction schemes. *Journal of Public Economics* 95: 1213-1222.

Rabin, M. (1993), Incorporating fairness into game theory and economics. *American Economic Review* 80: 1281-1302.

Rege, M., Telle, K. (2004), The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics* 88: 1625-1644.

Reuben, E., Riedl, A. (2009), Public goods provision and sanctioning in privileged groups. *Journal of Conflict Resolution* 53: 72-93.

Rivas, M. F., Sutter, M. (2011), The benefits of voluntary leadership in experimental public goods games. *Economics Letters* 112: 176-178.

Rockenbach, B., Milinski, M. (2006), The efficient interaction of indirect reciprocity and costly punishment. *Nature* 444: 718-723.

Sabater-Grande, G., Georgantzis, N. (2002), Accounting for risk aversion in repeated prisoners' dilemma games: An experimental test. *Journal of Economic Behavior and Organization* 48: 37-50.

Schechter, L. (2007), Traditional trust measurement and the risk of confound: An experiment in rural Paraguay. *Journal of Economic Behavior and Organization* 62: 72-92.

Sefton, M., Shupp, R., Walker, J. (2007), The effect of rewards and sanctions in the provision of public goods. *Economic Inquiry* 45: 671-690.

Seguino, S., Stevens, T., Lutz, M. (1996), Gender and cooperative behavior: Economic man rides alone. *Feminist Economics* 2: 1-21.

Sell, J., Wilson, R. K. (1991), Levels of information and contributions to public goods. *Social Forces* 70: 107-124.

Sell, J., Griffith, W., Wilson, R. K. (1993), Are women more cooperative than men in social dilemmas? *Social Psychology Quarterly* 56: 211-222.

Selten, R. (1975), Re-examination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4: 25-55.

Solow, J. L., Kirkwood, N. (2002), Group identity and gender in public goods experiments. *Journal of Economic Behavior and Organization* 48: 403-412.

Stockard, J., van de Kragt, A. J. C., Dodge, P. J. (1988), Gender roles and behavior in social dilemmas: Are there sex differences in cooperation and in its justification? *Social Psychology Quarterly* 51: 154.

Sutter, M., Haigner, S., Kocher, M. (2010), Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies* 77: 1540-1566.

Thöni, C., Tyran, J-R., Wengström, E. (2009), Microfoundations of social capital. Discussion Paper 09-24, Department of Economics, University of Copenhagen.

Tyran, J.-R., Feld, L. P. (2006), Achieving compliance when legal sanctions are non-deterrent. *Scandinavian Journal of Economics* 108: 135-156.

Ule, A., Schram, A., Riedl, A., Cason, T. N. (2009), Indirect punishment and generosity toward strangers. *Science* 326: 1701-1704.

van Dijk, F., Sonnemans, J., van Winden, F. (2002), Social ties in a public good experiment. *Journal of Public Economics* 85: 275-299.

van Lange, P. A. M., de Bruin, E. M. N., Otten, W., Joireman, J. A. (1997), Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology* 4: 733-746.

Yamagishi, T. (1986), The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51: 110-116.

Yamagishi, T. (1988), Seriousness of social dilemmas and the provision of a sanctioning system. *Social Psychology Quarterly* 51: 32-42.

Zelmer, J. (2003), Linear public goods games: A meta-analysis. *Experimental Economics* 6: 299-310.