DISCUSSION PAPER SERIES

IZA DP No. 8533

Institution Formation and Cooperation with Heterogeneous Agents

Sebastian Kube
Sebastian Schaube
Hannah Schildberg-Hörisch
Elina Khachatryan

October 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Institution Formation and Cooperation with Heterogeneous Agents

Sebastian Kube
University of Bonn and IZA

Sebastian Schaube
University of Bonn

Hannah Schildberg-Hörisch
University of Bonn and IZA

Elina Khachatryan
University of Kassel

Discussion Paper No. 8533
October 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8533
October 2014

ABSTRACT

Institution Formation and Cooperation with Heterogeneous Agents

Driven by an ever-growing number of studies that explore the effectiveness of institutional mechanisms meant to mitigate cooperation problems, recent years have seen an increasing interest in the endogenous implementation of these institutions. In this paper, we test within a unified framework how the process of institution formation is affected by three key aspects of natural environments: i) heterogeneity among players in the benefits of cooperation, ii) (a)symmetry in players' institutional obligations, and iii) potential trade-offs between efficiency and equality in payoff allocations. We observe social preferences to be limiting the scope for institution formation. Inequality-averse players frequently object to institutions that fail to address differences in players' benefits from cooperation – even if rejecting the institution causes monetary losses to all players. Relating our findings to previous studies on institution formation, we discuss potential advantages and drawbacks of stipulating unanimous support for implementing institutions that foster cooperation.

JEL Classification: C90, D02, D62, D63, H41

Keywords: institution formation, unanimity voting, cooperation problems, heterogeneous agents, inequality aversion

Corresponding author:

Hannah Schildberg-Hörisch
Institute of Applied Microeconomics
University of Bonn
Adenauerallee 24-42
53113 Bonn
Germany
E-mail: schildberg-hoerisch@uni-bonn.de

1 Introduction

"[...] a set of rules used in one physical environment may have vastly different consequences if used in a different physical environment."

(Ostrom, 1990, p.22)[1]

Cooperation problems are ubiquitous in many areas in economics, ranging from teamwork or hold-up problems in managerial economics, over community governance or property rights security in development economics, natural resource management or climate protection in environmental economics, trade obstacles or treaty formation in international economics, to tax compliance and the provision of public goods in public economics. Each example certainly has its own distinctive issues, but when it comes to mitigating the underlying cooperation problems, there is usually a common approach: the modification of individuals' incentive-compatibility constraints, such that "free-riding" is no longer the dominant strategy (e.g., Shavell and Polinsky (2000)). These modifications (implicitly or explicitly) impose restrictions on individuals' choice sets, which raises the question whether they will be implemented in the first place (e.g., Gürerk et al. (2006), Tyran and Feld (2006), Kosfeld et al. (2009), Bierbrauer and Hellwig (2011), Markussen et al. (2014)). In the present paper, we will shed light on this central question — asking in particular to which extent i) the heterogeneity of the involved players and ii) the (a)symmetry of the restrictions affects their implementation.

Consider the following example that we use throughout the paper, namely the provision of a public good. If members of a society are perfectly identical and all benefit equally from overcoming this social dilemma, one might expect them to mutually agree on establishing an institution that eliminates the social dilemma.[2] However, controversies might arise when members are heterogeneous and have different stakes in overcoming the social dilemma. In particular when equality considerations are taken into account, the exact content of the institution is key to successful implementation. Symmetric institutions, in which all members have the same obligations, might be rejected in favor of asymmetric institutions with member-specific obligations — even if this implies monetary losses for all members.

[1] Reported, inter alia, in Decker et al. (2003).

[2] Of course, expected benefits must exceed the costs of implementing the institution. Throughout the paper, we take this for granted by assuming the institution to be costless — notwithstanding that the case of positive costs would be interesting to study (e.g., Kamei et al. (2014)).

To causally identify how institution formation is affected by selected aspects of natural environments, we conducted a series of laboratory experiments. The basic underlying game, a public-good game, is a prominent workhorse for studying cooperation problems. Each player receives an endowment and has to decide on its allocation between private consumption and contributions to a public good. Provision of the public good creates benefits for all group members and is socially efficient in terms of the sum of monetary payoffs.[3] However, the individual marginal return from the public good is below the marginal return from private consumption, such that free-riding incentives exist which jeopardize public good provision. To offer players the opportunity to endogenously mitigate the cooperation problem, we add an additional stage that is played prior to the public good game. At this first stage, players decide on implementing an institution using unanimity voting. If all players in the group vote in favor of the institution, they are committed to certain efficiency-enhancing contribution levels in the subsequent public good game.[4] If at least one player votes against the implementation of the institution, the regular public good game is played and each player can freely decide how much to contribute in the second stage.

Players in our setup thus start in the absence of institutions and subsequently decide on the implementation of a joint institution to foster cooperation. In such an initial, lawless state of nature that is characterized by sovereign players facing a social dilemma, it seems natural to use unanimity voting for deciding on the implementation of institutions.[5] In fact, unanimous decision-making is the easiest possible, if not the only, voting procedure that players do not have to explicitly agree upon prior to voting. It does not require players to give up sovereignty, since each player can veto any decision. This is different for non-unanimous voting rules, such as majority voting, where players need to forfeit part of their sovereignty and which therefore typically only emerges after a joint history of cooperation.[6]

[3]Throughout the paper, efficiency refers to monetary payoffs.

[4]Essentially, the institution consists of two elements: i) It states a certain obligation for each player, i.e., the exact amount that he is required to contribute in the second stage, and ii) it installs a deterrent sanctioning technology, i.e., players' contributions are monitored and a player receives harsh punishment when deviating from the required contribution. For reasons of simplicity, the second component is not an explicit part of the experiment. Instead, it is implicitly modeled by restricting a player's choice set in the second stage to the required contribution (see Kosfeld et al. (2009) or Gerber et al. (2013) for similar approaches).

[5]The idea of an initial state of nature that is characterized by sovereign agents in a lawless environment goes back to Rousseau (1762) and Hobbes (1651).

[6]Cooperation in past periods may foster trust and reciprocal behavior among players, which may make them

Since our focus is on how institution formation is affected i) by heterogeneity in players' benefits from cooperation, and ii) by the (a)symmetry of obligations, we vary these factors in a controlled manner while fixing the decision rule to unanimity voting in all treatments. First, in some treatment conditions (*Homogeneous types*), all players are of the same type and, thus, receive the same benefits from the public good, while in other conditions (*Heterogeneous*), there are two types that differ in their marginal benefits. Second, we vary the content of the institution. All players are either obliged to contribute their entire endowment to the public good (*Symmetric Institution*), or obligations differ between the two player types (*Asymmetric*). While the symmetric institution implies efficient public good provision, but inequality in payoffs for heterogeneous players, obligations in the asymmetric case are chosen such that final payoffs are equalized. This setup allows us to clearly identify the roles of inequality aversion and efficiency concerns in the process of institution formation.

We find that inequality considerations can hamper the formation of efficient institutions meant to foster cooperation. With heterogeneous player types, those with low marginal benefits frequently object to the symmetric institution (about 40% reject it). The same is observed for homogeneous player types with asymmetric institutions (about 45% reject it). On the other hand, support is high when the institution implements equal payoff allocations: the asymmetric institution seems perfectly acceptable for heterogeneous player types, as does the symmetric institution for homogeneous types. In both cases, more than 90% of all votes are in favor of the implementation.

With respect to the sum of monetary payoffs, we observe that efficiency is always lower when institution formation failed than when the institution was implemented. The symmetric institution for homogeneous player types performs best (average efficiency is above 90% of the maximally obtainable sum of payoffs). Compared to this, under heterogeneity both the symmetric and the asymmetric institution lead to lower rates of efficiency, albeit for different reasons. In the former case, average efficiency is lower because the symmetric institution is frequently rejected. In the latter case, heterogeneous player types frequently implement the asymmetric institution, but average efficiency is lower since total obligations and the level of public good provided are

willing to forfeit part of their sovereignty. To give just one example, international organizations, most notably the League of Nations as the precursor of what is now the United Nations, used to apply the unanimity voting rule for voting on matters of substance before World War II. It was only during the post-war growth in international coordination through permanent organizations that non-unanimous voting rules were increasingly applied.

lower. The asymmetric institution for homogeneous player performs worst.

The striking differences in average efficiency and implementation rates between treatments underline at least three important issues. First, our results stress that inequality aversion can have a strong impact on the process of institution formation. In most of the existing studies on institution formation, introducing social preferences to the theoretical models usually leads to stronger support for the institution; be it because more players want to be part of a coalition than is predicted under standard preferences (e.g. Kosfeld et al. (2009), McEvoy et al. (2011)), or because the institution to be implemented allows them to reduce free-riders' payoffs (e.g., Markussen et al. (2014)). By contrast, in those cases where inequality-aversion makes a difference in our setup, inequality-averse players are predicted to be less inclined to support the formation of the institution — a phenomenon that has not been discussed so far in the corresponding literature. As can be seen in our data, this easily leads to situations where players forego monetary payoffs by objecting to efficient institutions; in particular given the requirement of unanimous decisions.

However, and this is the second point we would like to stress, the use of unanimity voting for implementing institutions must not always be detrimental to efficiency. On the contrary, it can even help to foster cooperation.[7] Already Wicksell (1964) discusses that institutions based on unanimity or consensus voting can be ideally suited to overcome the canonical problem of free-riding. Unanimity makes individual activism implicitly conditional on the activism of all other parties involved. This mitigates the dilemma of institution formation: those who agree on implementing an institution do not face the subsequent risk of free-riding by non-supporting, and thus non-participating, players (see also Maggi and Morelli (2006)). Consequently, there is no drawback in supporting institutions that are based on unanimous decisions; either all players participate and the institution is formed, or the institution is not created at all. This can be clearly seen when comparing our data to related studies that implicitly allow players to "opt out" of institutions (Kosfeld et al. (2009), Gerber et al. (2013)). While efficient and equitable institutions are frequently not implemented in those other studies, we observe that such institutions receive strong support and are implemented most of the times when unanimity is required.

[7] Apart from this, there is also another desirable feature of unanimity. It is easy to agree on a principle of unanimity, since every party has veto power and freedom of choice is thus granted (at least ex ante, before an institution is implemented). Moreover, recent evidence implicitly suggests that many people value unanimous decisions, and that they have a strong preference for involving all players in the decision-making process (see Decker et al. (2003), Sutter et al. (2010), or Linardi and McConnell (2011), and the references therein).

Of course, this is not to say that unanimity will lead to stronger cooperation all the time. The unanimity voting rule grants de facto veto rights to every party involved. Therefore, it is crucial that the institution to be voted on addresses idiosyncratic interests amongst the involved parties. We see this in our study, since homogeneous players frequently reject asymmetric institutions, and heterogeneous players regularly reject symmetric institutions. Support for the latter is also found in lab experiments by Banks et al. (1988), and Kesternich et al. (2014), as well as in the survey evidence reported in Reuben and Riedl (2013). The importance of fixing appropriate institutional obligations beforehand is also reflected in the literature that studies homogeneous players' acceptance thresholds on minimum contribution requirements in public good games (Birnberg et al. (1970), Dannenberg et al. (2010), Rauchdobler et al. (2010)). Taken together, the evidence strongly suggests that prior to the ultimate voting about the implementation of an institution, great care has to be taken ex ante in designing the institution.

Finally, the institution at hand is build around a centralized authority with a deterrent sanctioning technology, but also other institutional mechanisms could be implemented to foster cooperation (e.g. Falkinger and Fehr (2000), Andreoni and Gee (2012)). One could even think about implementing decentralized sanctioning regimes. Of course, the seminal papers by Ostrom et al. (1992), Fehr and Gächter (2000), and Fehr and Gächter (2002) started with the basic idea of mutual monitoring and punishment among the members of a group; focusing in particular on the question whether certain behavioral norms can emerge, even in the absence of formal institutions with a centralized structure. Still, there are some studies where players do vote over the implementation of decentralized sanction regimes (Putterman et al. (2011), Markussen et al. (2014), Kamei et al. (2014)). Those studies exclusively focus on majority voting and homogeneous agents. It might be interesting to reconsider their results in our setup with heterogeneous players, or to see how behavior would change when using unanimity voting procedures.

The outline of the paper is as follows. Section 2 describes the experiment design. In Section 3, theoretical predictions for subjects' behavior will be derived, using both standard and social preferences (inequality aversion). Section 4 presents and discusses the empirical results. Section 5 concludes.

2 Experiment

In natural environments, the complexity of the process of institution formation makes it particularly difficult to draw causal conclusions about the conditions under which institutions come into being. As a starting point, we therefore use the controlled environment of laboratory experiments to study central aspects of the endogenous formation of institutions. In this section, we present the design of our experiment and describe the implemented procedures.

Design

Our design builds on a standard public goods game (VCM game), a frequently used workhorse to study elements of social dilemmas in the lab (e.g., Isaac and Walker (1988)). Each player has a private endowment E. Players simultaneously decide on the amount c_i that they contribute to a public good, with $0 \leq c_i \leq E$, $i = 1, ..., n$. The benefits from the public good are enjoyed by all players, independent of their individual contribution c_i. In some treatments, players are heterogeneous, i.e., not all players benefit from the public good to the same extent. To model heterogeneity, we allow the marginal per capita return (MPCR) γ_i from the public good to vary across players.[8] Given the contributions of all players $(c_1, ..., c_n)$, player i's material payoff π_i is thus given by

$$\pi_i = E - c_i + \gamma_i \textstyle\sum_{i=1}^{n} c_i.$$

In all treatments, parameters for γ_i are chosen such that players face a social dilemma. Efficiency, defined as the sum of payoffs of all players, is maximized if all players contribute their entire endowment. Yet, from an individual perspective, each player's material payoff is maximized by not contributing to the public good, regardless of the other players' contributions. Formally, this implies $\sum_{i=1}^{n} \gamma_i > 1$ and $\gamma_i < 1 \; \forall i$.

We form groups of three players ($n = 3$). Between treatments, we vary two components. First, we vary the composition of players' types γ_i. In some treatments (HOM), players are homogeneous, i.e., all players are of the same type and thus receive the same benefits from the public good ($\gamma_i = 2/3$). In other treatments (HET), players are heterogeneous: two players have a high return from the public good ($\gamma_i = 3/4$) and one player has a low return ($\gamma_i = 1/2$).[9] The

[8] To give just two among many possible examples, nation states differ in their benefit from climate protection or researchers at different stages of their career benefit from joint publications to a different extent.

[9] We choose a single player with a lower return because this setup is sufficient to illustrate the potential weakness of unanimity voting, i.e., already a single player can prevent successful institution formation by vetoing.

different marginal per capita returns are chosen as to keep total efficiency gains constant between treatments ($2 \cdot 3/4 + 1 \cdot 1/2 = 3 \cdot 2/3$).

Second, we vary availability and content of the institution. In the benchmark treatments (VCM), there is no institution formation stage and players play a regular public goods game. In the main treatments, there is an institution formation stage first, followed by a contribution stage. In the institution formation stage, a single institution is available and can be implemented via unanimity voting, i.e., the institution is implemented if and only if all players vote in favor of adopting the institution. If the institution is rejected, the regular public goods game without any restrictions on contributions is played. The institution consists of two elements. First, it states each player's obligation $\bar{c}_i$, the amount that each player has to contribute to the public good in the second stage. Second, it installs a deterrent sanctioning technology, i.e., a player receives a sufficiently harsh punishment when deviating from his obligation to make it payoff maximizing to contribute according to the individual obligation. For reasons of simplicity, the sanctioning scheme is not explicitly modeled in the experiment. Instead, if the institution has been implemented in the first stage, effective sanctioning is implicitly modeled by restricting players' choice set in the second stage to the level of the individual obligation (see also Kosfeld et al. (2009) for a similar approach). Voting and the implementation of the institution are costless.[10]

The main treatments vary in the type of institution that is available. In general, treatments are designed to reflect a tradeoff between efficiency and equality of payoffs. In treatments with the symmetric institution (SYM), all players are obliged to contribute their entire endowment to the public good if the institution has been implemented. The symmetric institution maximizes the sum of payoffs of all players and, thus, induces the efficient outcome. In treatments with the asymmetric institution (ASYM), one player is required to contribute 8 units, while the two others are obliged to contribute all 20 units to the public good. In treatments with heterogeneous players, the obligation is 20 for the high types, and 8 for the low types. Obligations are chosen such that the asymmetric institution implies equal payoffs for both types of players (36 each), which comes at an efficiency cost. In contrast, with heterogeneous players, the symmetric institution implies inequality in final payoffs (45 for the high types and 30 for the low type). If the asymmetric institution is combined with homogeneous players, one randomly chosen player has to contribute 8 units, while the other two players are obliged to contribute 20 units. The design results in the

[10]These are simplifying assumptions. Qualitatively, the theoretical predictions do not change as long as the gains in individual material payoffs due to implementing the institution outweigh the individual implementation costs.

Table 1: Treatments

Institution Player types	VCM	SYM	ASYM
HOM	$\gamma = 2/3$ no obligations	$\gamma = 2/3$ $\bar{c} = 20$ $\Pi = 40$	$\gamma = 2/3$ $c(\bar{c} = 20) = 20$, $c(\bar{c} = 8) = 8$ $\Pi(\bar{c} = 20) = 32$, $\Pi(\bar{c} = 8) = 44$
HET	$\gamma_h = 3/4$, $\gamma_l = 1/2$ no obligations	$\gamma_h = 3/4$, $\gamma_l = 1/2$ $\bar{c} = 20$ $\Pi_h = 45$, $\Pi_l = 30$	$\gamma_h = 3/4$, $\gamma_l = 1/2$ $\bar{c_h} = 20$, $\bar{c_l} = 8$ $\Pi_h = \Pi_l = 36$

2×3-treatment matrix shown in Table 1.

Procedures

The computerized experiments (using z-Tree; Fischbacher (2007)) were run at the BonnEcon-Lab of the University of Bonn, Germany in 2012. Student subjects were recruited randomly from all majors (using Orsee; Greiner (2004)) and were randomly assigned to one of the six treatments (between-subject design). For each treatment, we ran two sessions with 24 subjects each. In each session, subjects first received written instructions (see Appendix B). To create common knowledge, instructions were read out aloud to the subjects. Afterwards, subjects answered a set of control questions and could pose clarifying questions to ensure understanding of the game's structure and payoffs. Subjects then played the game repeatedly for 20 periods. Interaction took place within the same group of three subjects (partner matching protocol), it was anonymous and decisions were taken in private at the computer. After each voting stage, subjects received feedback on the voting result and the voting behavior of the other two subjects in their matching group. After each contribution stage, subjects were informed about their own payoff and the payoffs and contributions of the other two subjects in their group. After all 20 periods, subjects answered a questionnaire covering socio-demographic characteristics and social preferences. In particular, we used the strategy method to elicit responders' minimal acceptable offer in a non-incentivized, 10 Euro Ultimatum Game and dictator behavior in a non-incentivized, 10 Euro Dictator Game. Each session lasted about 80 minutes. Accumulated earnings were converted at a rate of 40 tokens = 1 Euro. Total earnings per subject ranged between 10 Euro and 22.5 Euro, with an average of about 16.4 Euro.

Altogether, we had 282 subjects, and observations on 5640 individual decisions. Given the allocation of subjects to the six treatments, repeated interaction in 20 periods and matching groups of 3, we have 16 independent observations per treatment.[11] 39% of our subjects are male, their age ranges from 16 to 42, with an average age of 22 years.

3 Behavioral Predictions

For each treatment, we characterize players' equilibrium behavior under two alternative assumptions concerning the shape of the utility function. First, we assume that each player's utility function coincides with the monetary payoff of the game, π_i, i.e., that players have standard preferences. Second, we assume that at least (some) players have social preferences as defined in Fehr and Schmidt (1999): in addition to valuing own monetary payoff, a player suffers from inequality in monetary payoffs, i.e., from others being worse or better off than himself. In our treatment with heterogeneous benefits from the public good, players might vote against implementing an institution that obliges all players to contribute equally to the public good in order to avoid inequality in payoffs. Hence, we consider the model of Fehr and Schmidt (1999) as a natural choice to derive predictions for our setup. In the remainder of this section, we will provide an intuition for the behavioral predictions for each treatment under the two alternative assumptions on the shape of players' utility functions using the parameters of our design. More general proofs are provided in Appendix A.

Table 2 summarizes the behavioral predictions for players with standard preferences. In basic VCM games, they are predicted not to contribute to the public good at all. Whenever $\gamma_i < 1$, contributing does not pay off from an individual perspective. Condition $\gamma_i < 1$ is met for all players in treatments HOM-VCM ($\gamma = 2/3$) and HET-VCM ($\gamma_l = 1/2$ and $\gamma_h = 3/4$).

In all two-stage treatments, predictions are derived using backward induction. Let U^{INST} denote utility when the institution has been implemented, with INST=SYM for the symmetric and INST=ASYM for the asymmetric institution. In the contribution stage, players will compare the utility they receive with the respective institution being in place, U^{INST}, to the utility of the VCM game that is played if the institution has not received unanimous support in the voting stage, U^{VCM}. Unanimity voting ensures that, whenever $U^{INST} \geq U^{VCM}$, it is a best response

[11] Exceptions are treatments HET-VCM and HOM-ASYM, for which we have 15 independent observations since some subjects did not show up.

Table 2: Behavioral Predictions Based on Standard Preferences

Institution Player type		VCM	SYM	ASYM
HOM	voting	-	implement institution	implement institution
	contribution	$c = 0$	$c = 20$	$c(\bar{c} = 20) = 20$, $c(\bar{c} = 8) = 8$
HET	voting	-	implement institution	implement institution
	contribution	$c_h = c_l = 0$	$c_h = c_l = 20$	$c_h = 20$, $c_l = 8$

to the voting behavior of the other players to vote in favor of the institution. If all other players also vote in favor of implementing the institution, the institution will be implemented and the player's preferred outcome is achieved. If, in contrast, at least one other player votes against implementing the institution, the institution will not be implemented and the VCM game will be played. However, the approving player is still equally well off as if he had voted against implementing the institution. Whenever $U^{INST} < U^{VCM}$, a player will vote against installing the institution. In our design, $U^{INST} > U^{VCM} = E = 20$ for all player types in treatments HOM-SYM, HOM-ASYM, HET-SYM and HET-ASYM.[12] Consequently, for all treatments, players with standard preferences are predicted to vote in favor of the respective institution. The institution will be implemented and players will contribute according to their individual obligation. To summarize, if players have standard preferences, unanimity voting on the formation of institutions is predicted to be a powerful tool to overcome the social dilemma of public good provision. This results holds irrespective of whether players are homogeneous or heterogeneous and whether a symmetric or an asymmetric institution is voted on.

Table 3 displays the behavioral predictions for players with social preferences in terms of inequality aversion (Fehr and Schmidt, 1999). If players have social preferences, there are multiple equilibria in treatment HOM-VCM.[13] The intuition is as follows: If all players are sufficiently averse to advantageous inequality (β sufficiently high)[14], they will exactly match the contribu-

[12]In treatment HOM-SYM $U^{SYM} = \gamma nE = 40$, in HET-SYM $U_l^{SYM} = 30$ and $U_h^{SYM} = 45$, in HET-ASYM $U_l^{ASYM} = U_h^{ASYM} = 36$, in HOM-ASYM, for an obligation of 8, $U_8^{ASYM} = 44$ and for an obligation of 20, $U_{20}^{ASYM} = 32$.

[13]The proof is provided in Fehr and Schmidt (1999).

[14]In the model of Fehr and Schmidt (1999), the parameter β captures the intensity of aversion to advantageous inequality, while the parameter α measures the degree of aversion to disadvantageous inequality.

tion level $c \in [0, E]$ of the other players to equalize payoffs. If players are not or only mildly averse to advantageous inequality (β low), the only equilibrium that remains is the one with zero contributions of all players. In treatment HET-VCM, the basic mechanism driving the existence of equilibria with positive contributions is the same. If all players are sufficiently averse towards earning more than others, they contribute positive amounts as soon as the other players contribute positive amounts to prevent an unequal payoff distribution. However, to achieve equal payoffs for all three players, the low type contributes less than the two high types.

In treatments HOM-SYM and HET-ASYM, assuming social instead of standard preferences does not change the predictions. In both cases, the proposed institution guarantees equality of payoffs while simultaneously maximizing utility of players who are sufficiently averse to unequal payoffs. Hence again, all players are predicted to vote in favor of the respective institution, it will be implemented, and players will contribute according to their obligation. In treatments HET-SYM and HOM-ASYM, however, predictions based on standard preferences and social preferences differ. In both treatments, players with standard preferences always support the formation of the institution as it offers a higher monetary payoff than the VCM and they do not suffer from unequal payoffs that arise from implementing the institution. In contrast, in treatment HET-SYM, low type players with social preferences who suffer sufficiently from being worse off than the high types (α sufficiently high), object to institution formation. They prefer a lower monetary payoff, but equal payoffs across players in the VCM, to a higher monetary payoff, but disutility from inequality due to the symmetric institution being in place. Consequently, low type players drive all rejections of the proposed symmetric institution. Similarly, in treatment HOM-ASYM, all players potentially have a motive for voting against the asymmetric institution that introduces inequality in payoffs: Players with an obligation of 8 tokens, if they are sufficiently averse to advantageous inequality, and players with an obligation of 20 tokens if they are sufficiently averse to disadvantageous inequality.[15]

[15] Preferences for efficiency, see, e.g., Charness and Rabin (2002), are an alternative explanation for rejecting an asymmetric institution. Efficiency seekers should reject institutions that do not induce full contributions in order to contribute more than they were obliged to with the institution being in place. Our results (see section 4.4) do not provide evidence for efficiency seeking as a predominant motive for rejections.

Table 3: Behavioral Predictions Based on Fehr-Schmidt Preferences

Institution Player type		VCM	SYM	ASYM
HOM	voting	-	implement institution	type $\bar{c} = 20$ rejects if α high, type $\bar{c} = 8$ rejects if β high
	contribution	(c,c,c), $c \in [0, 20]$ if $\beta_i > 1/3 \forall i$; (0,0,0) otherwise	$c = 20$	if reject: as in HOM-VCM otherwise: $c(\bar{c} = 20) = 20$, $c(\bar{c} = 8) = 8$
HET	voting	-	low type rejects if α_l high	implement institution
	contribution	$(c_h, c_h, c_l = 2/5c_h)$, $c_h \in [0, 20]$ if $\beta_h > 2/7$ and $\beta_l > 2/5$; (0,0,0) otherwise	if reject: as in HET-VCM otherwise: $c_h = c_l = 20$	$c_h = 20$, $c_l = 8$

4 Results

The results section is structured along five sets of predictions concerning differences in voting and contribution behavior across treatments. These predictions build on the theoretical results presented in Section 3 and derived in Appendix A. Moreover, we assume that at least some players are inequality averse to an extent that induces their behavior to deviate from the predictions based on standard preferences. Our questionnaire data[16] and actual behavior in the experiment provide evidence in favor of the assumption that many of our subjects are inequality averse.[17]

First, we will briefly present results in treatments HOM-VCM and HET-VCM that provide baseline scenarios for comparing whether unanimity voting on institutions increases efficiency. We proceed by discussing under which circumstances unanimity voting on symmetric or asymmetric institutions helps to increase public good provision. We thereby focus on predictions that are based on treatment comparisons in which changes in behavior can be attributed to a single change in setup. That means, we either compare treatments with different institutions, while keeping constant the composition of player types (HOM or HET) or we compare treatments with a different composition of player types, while keeping constant the nature of the institution to be voted on (SYM or ASYM).

Table 4 and Table 5 contain first descriptive results. Table 4 displays contributions averaged over all periods by treatment. Table 5 shows the share of affirmative votes and implementation rates averaged over all periods by treatment. Moreover, Figures 1 and 2 in Appendix C display the treatment-specific development of contributions and share of affirmative votes over time.

[16] In a non-incentivized, standard Dictator Game, 2/3 of our subjects donate positive amounts to the receiver. 22% of all subjects split 10 Euro equally. Positive levels of donations in a Dictator Game indicate aversion to advantageous inequality. Furthermore, we have information on responder behavior in a non-incentivized Ultimatum Game. For a pie size of 10 Euro, we use the strategy method to elicit the minimal acceptable offer. 77% of our subjects reject offers of 4 Euro or less, indicating aversion to disadvantageous inequality.

[17] The predictions are based on two further assumptions. First, when comparing two-stage treatments to the corresponding baseline VCMs, we assume that whenever an institution is rejected in the voting stage of a two-stage treatment, subjects play the equilibrium in the VCM of the contribution stage that the same group of subjects would play in the baseline VCM. Second, for each of the two treatments HOM-VCM and HET-VCM, all possible equilibria can be ranked according to efficiency on a continuous scale from 0 to 1. When comparing the two baseline VCMs in treatments HOM-VCM and HET-VCM, we assume that the same group of subjects would play the equilibrium of the same efficiency rank in treatment HOM-VCM and HET-VCM, e.g., a given group of subjects that chooses the most efficient equilibrium in treatment HOM-VCM, would also choose the most efficient equilibrium in treatment HET-VCM.

Table 4: Average Contributions by Treatment

Institution		VCM	SYM	ASYM
Player type				
HOM	overall	10.72	18.18	11.44
		(7.83)	(5.27)	(8.34)
	types $\bar{c} = 20$	–	–	12.12
		–	–	(8.79)
	types $\bar{c} = 8$	–	–	10.09
		–	–	(7.23)
HET	overall	8.05	14.21	13.85
		(6.56)	(7.79)	(7.20)
	high types	9.42	14.77	17.18
		(7.07)	(7.42)	(6.36)
	low types	5.33	13.08	7.21
		(4.33)	(8.38)	(2.90)

Standard deviations are in parantheses.

Table 5: Share of Affirmative Votes and Implementation Rate by Treatment

Institution		SYM	ASYM
Player type			
HOM	affirmative votes		
	overall	.95	.54
	types $\bar{c} = 20$	–	.48
	types $\bar{c} = 8$	–	.68
	implementation rate	.87	.27
HET	affirmative votes		
	overall	.84	.91
	high types	.96	.90
	low types	.60	.94
	implementation rate	.56	.77

4.1 Baseline Treatments: Homogeneous versus Heterogeneous Players in the VCM

Comparing behavioral predictions for treatments HOM-VCM and HET-VCM results in the following prediction:

Prediction 1:
Average contributions in treatment HET-VCM are lower than in treatment HOM-VCM.

On average, subjects contribute 10.7 out of 20 units in treatment HOM-VCM and 8.1 units in treatment HET-VCM (Mann-Whitney ranksum test (MWU), $p = 0.11$).[18] In line with prediction 1 and the findings of Fisher et al. (1995), contributions in a standard VCM tend to be lower with heterogeneous than with homogeneous agents.

Moreover, we observe that average contributions of low and high types differ in HET-VCM: while low type players contribute only 5.3 units, high type players contribute 9.4 units on average (MWU, $p < 0.01$). As a consequence, average payoffs for the two player types are similar, 26.8 and 28.7 units, respectively. Players of both types seem to intuitively strive for equal payoffs.

4.2 Unanimity Voting on the Symmetric Institution: Homogeneous versus Heterogeneous Players

We first consider the voting behavior of homogeneous players who are confronted with the decision whether to install the symmetric institution that obliges each player to contribute the efficient amount, 20 units. Overall, 95.2% of votes (914 out of 960 votes) are in favor of implementing the symmetric institution. As a result, in 86.6% of all cases, all three players of a group unanimously agree to implement the symmetric institution and it is indeed implemented.

Concerning contributions, our results are in line with prediction 2.

Prediction 2:
In treatment HOM-SYM, average contributions are (weakly) higher than in treatment HOM-VCM.

On average, subjects contribute significantly more in treatment HOM-SYM than in treatment HOM-VCM (18.2 instead of 10.7 units, MWU, $p < 0.01$). After some periods of initial learning

[18] Throughout the paper, we report two-sided p-values. Each matching group's average contribution is one independent observation.

efficiency is close to 100% (see also Figure 1 in Appendix C). To summarize, with homogeneous players, unanimity voting on the symmetric institution increases efficiency substantially.

Does unanimity voting on the efficient institution also yield sufficient support if players are heterogeneous, i.e., if the efficient institution introduces unequal payoffs? Prediction 3 summarizes our predictions for treatment HET-SYM.

Prediction 3:

1. In treatment HET-SYM, average contributions are (weakly) higher than in treatment HET-VCM.

2. In treatment HET-SYM, both implementation rate and average contributions are lower than in treatment HOM-SYM.

Again, we start by analyzing behavior in the voting stage. In treatment HET-SYM, the overall share of affirmative votes is lower than in treatment HOM-SYM, 83.9% instead of 95.2%. Heterogeneous players object the implementation of the efficient symmetric institution more often than homogeneous players. The difference in affirmative votes between treatment HOM-SYM and HET-SYM persists over time (see Figure 2 in Appendix C). Similarly, the overall implementation rate in treatment HET-SYM is 56.3%, substantially lower than in treatment HOM-SYM (86.6%). In line with the theoretical predictions for treatment HET-SYM, rejections of the institution are largely due to the voting behavior of low types. In our data, 95.9% of high types vote in favor of implementing the institution in treatment HET-SYM, but only 59.7% of low types do.[19]

As a consequence of the lower implementation rate and in line with prediction 3, average contributions are significantly lower in treatment HET-SYM than HOM-SYM: 14.2 instead of 18.2 (MWU, $p = 0.01$). However, average contributions in treatment HET-SYM are significantly higher than in the VCM with heterogeneous players (MWU, $p < 0.01$).

Overall, if players are heterogeneous rather than homogeneous in their marginal returns from the public good, unanimity voting on the efficient institution is a less powerful tool for

[19]More precisely, we expect those low types to vote against installing the institution who are sufficiently averse to disadvantageous inequality that arises if both types contribute the same amount to the public good, but benefit from it to a different extent. Recall that in the final questionnaire, we used the strategy method to elicit responders' minimal acceptable offer in a non-incentivized, 10 Euro Ultimatum Game. The minimal acceptable offer serves as a proxy for a player's aversion to disadvantageous inequality. For the low types, the correlation coefficient between voting in favor of the symmetric institution and the minimal acceptable offer (ranging from 0 to 5 Euro) is -0.11 with $p = 0.05$, i.e., low types who are more strongly averse to disadvantageous inequality tend to reject the institution more often.

increasing efficiency in public good provision. Still, compared to the standard public good game in which no institution is available, unanimity voting on the efficient institution increases efficiency substantially – even if players are heterogeneous.

4.3 Unanimity Voting on the Asymmetric Institution: Homogeneous versus Heterogeneous Players

A potential remedy to the frequent rejections of the symmetric institution by low type players is to design an asymmetric institution that ensures the maximum possible payoffs among the set of all equitable payoff allocations. Obviously, under the asymmetric institution, the low type players' obligation must be lower than under the symmetric institution. As a drawback, the implementation of the asymmetric institution results in a lower level of public good provision than the implementation of the symmetric institution. Prediction 4 summarizes our predictions for treatment comparisons.

Prediction 4:

1. In treatment HET-ASYM, the implementation rate is higher than in treatment HET-SYM. There is no unambiguous prediction whether average contributions are higher in treatment HET-ASYM or in treatment HET-SYM.
2. In treatment HET-ASYM, average contributions are (weakly) higher than in treatment HET-VCM.

Overall, 91.0% of players vote in favor of implementing the asymmetric institution which results in 77.2% successful implementations. In line with prediction 4, with heterogeneous players, the asymmetric institution that guarantees equal payoffs for both player types is more than 20% points more likely to be implemented than the symmetric one that induces the efficient outcome, but unequal payoffs across player types. The higher implementation rate is due to the substantially higher likelihood of low types to vote in favor of the asymmetric institution than the symmetric one: 94.1% instead of 59.7%. With 89.5%, the high types' share of affirmative votes for the asymmetric institution is roughly comparable to the share of affirmative votes for the symmetric institution (95.9%).

While implementation rates differ markedly for treatment HET-SYM and HET-ASYM, average contributions do not: 13.9 units in HET-ASYM compared to 14.2 units in HET-SYM (MWU, $p = 0.97$). There are two opposing effects that cancel each other out: while the higher imple-

mentation rate in HET-ASYM increases contributions, implementing the asymmetric institution instead of the symmetric one reduces contributions of the low types from 20 to 8 units. Compared to the benchmark VCM game with heterogeneous players, average contribution levels are significantly higher in treatment HET-ASYM than in treatment HET-VCM (MWU, $p < 0.01$).

Overall, designing institutions that address players' demand for equal benefits from institution formation is very successful in raising the implementation rate. In many contexts, a higher rate of institution formation could be considered beneficial per se, e.g., due to raising reliability of public good provision or by potentially triggering future institutionalized cooperation. However, increasing the implementation rate by voting on an asymmetric institution will always come at the cost of institutionalizing less than efficient levels of public good provision.

To rule out that the high implementation rate in HET-ASYM is not due the asymmetry in contributions per se, we now turn to treatment HOM-ASYM. Here, we can explore how the asymmetric institution performs if players are homogeneous, i.e., when it introduces binding rules concerning contributions to potentially increase efficiency, but those rules induce unequal payoffs across players.

Prediction 5:

1. There is no unambiguous prediction whether average contributions are higher in treatment HOM-ASYM or in treatment HOM-VCM.

2. In treatment HOM-ASYM, the implementation rate is lower and average contributions are (weakly) lower than in treatment HOM-SYM.

3. In treatment HOM-ASYM, the implementation rate is lower than in treatment HET-ASYM. There is no unambiguous prediction whether average contributions are higher in treatment HOM-ASYM or in treatment HET-ASYM.

Proposing an asymmetric institution to homogeneous players receives relatively low levels of support. The average share of affirmative votes ranges between 40 and 70% over time, resulting in an average implementation rate of only 26.7%. For players with an obligation of 8 units, the share of affirmative votes is 67.7%, while it is 20 percentage points lower for those with an obligation of 20. As our behavioral predictions point out, both types of players possibly have a motive to vote against the institution, namely aversion to advantageous inequality (for players with an obligation of 8 units) and aversion to disadvantageous inequality (for players with an obligation of 20 units).[20]

[20] Our data on behavior in a hypothetical Dictator Game and Ultimatum Game provide evidence for the former,

We have already shown that, with homogeneous players, proposing a symmetric institution helps to overcome the social dilemma of public good provision. This is not the case with an asymmetric institution. The average contributions in treatment HOM-ASYM are not significantly different from average contributions in treatment HOM-VCM (MWU, $p = 0.75$) and significantly lower than in treatment HOM-SYM (MWU, $p < 0.01$).

Finally, the asymmetric institution performs worse for homogeneous than for heterogeneous players, i.e., when it introduces inequality instead of addressing it. With homogeneous players, both the share of affirmative votes and the average contributions are lower (MWU, $p < 0.01$ for affirmative votes and $p = 0.04$ for contributions). This strongly suggests that the success of the asymmetric institution for heterogeneous agents is indeed due to addressing payoff inequalities between agents.

4.4 Contributions by Institution Formation Status

So far, we have analyzed average contributions in a given treatment, averaging over cases of successful institution formation and those of failure to form an institution. Our results document that, typically, unanimity voting on implementing institutions is a powerful tool to increase average contributions. We have not studied yet, however, how failure to implement the proposed institution affects contribution levels. If motives for objecting to institution formation differ across treatments, contribution levels in case of failed institution formation could also differ across treatments. For example, inequality aversion is a plausible motive for voting against institution formation in treatments HET-SYM and HOM-ASYM in which institutions induce unequal payoffs. In treatments HOM-ASYM and HET-ASYM, a preference for efficient levels of public good provision could drive rejections. Rejections of the institution are harder to rationalize in treatment HOM-SYM because implementation of the institution results in maximal and equal payoffs. Consequently, rejections could be due to, e.g., mistakes or pleasure from exerting (destructive) power. These motives could induce negative reciprocity, resulting in contribution levels well below the corresponding VCM. In contrast, efficiency seekers could reject an asymmetric institution aiming at contribution levels that exceed institutional obligations. Players who reject an institution due to inequality aversion have motives to contribute as in the baseline VCM whose

but not the latter motive. The correlation coefficient between voting in favor of the institution and the donated amount in the Dictator Game is -0.18, with $p < 0.01$, while the correlation coefficient between voting in favor and the minimal acceptable offer in the UG (ranging from 0 to 5) is small and not significantly different from zero.

Table 6: Average contributions after failed institution formation

Institution Player type	VCM	SYM	ASYM
HOM	10.72	6.43	9.78
	(7.83)	(6.90)	(8.56)
HET	8.05	6.76	6.59
	(6.56)	(6.33)	(7.12)

Standard deviations are in parantheses.

equilibria ensure equality of payoffs across players.

While we did not elicit subjects' individual beliefs about the preferences of players which rejected the institution, our data on average contributions in case of failed institution formation is still telling. Table 6 and Figure 3 in the Appendix show that, in the relatively rare case of institution failure (13%), average contribution levels in treatment HOM-SYM are indeed substantially below those of the corresponding VCM (6.4 instead of 10.7 units). In treatments HET-SYM and HOM-ASYM, average contributions are much closer to those of the corresponding baseline VCM which could indicate that rejections may largely be due to inequality aversion. Average contributions in treatment HET-ASYM do not exceed contributions of the baseline VCM as one would expect if efficiency seeking would be the predominant motive for rejections. Taken together, our results do not point at a large, "hidden cost" of failed institution formation, namely substantially and frequently reduced contributions in case of failed institution implementation (except for treatment HOM-SYM).

4.5 Do Payoffs of High and Low Types Differ?

In all treatments, there is a one-to-one relationship between contributions and average payoffs at the level of a matching group. However, for a given average contribution level, payoffs could still differ for low and high type players. Payoffs of high and low types are predicted not to differ in treatments HET-VCM and HET-ASYM. In treatment HET-SYM, however, payoffs are predicted to be lower for low types than for high types. Results in Table 7 are in line with these

predictions. In treatments HET-VCM and HET-ASYM, payoff differences between high and low types are small (about 2 and 0.4 units, respectively), while the payoff difference is about 9 units in treatment HET-SYM.

Table 7: Average Payoffs of High and Low Types (in Units)

Average payoff Treatment	High type	Low type
HET-VCM	28.70	26.75
HET-SYM	37.19	28.23
HET-ASYM	33.99	33.58

5 Conclusion

The paper at hand studied the process of institution formation in social dilemmas, in particular the role of heterogeneity among players i) in their benefits from cooperation and ii) in their institutional obligations. We found that the potential tension between efficiency and equality in payoffs, originating from these heterogeneities, strongly affected implementation rates of institutions. With heterogeneous players, aggregate implementation rates were significantly lower for institutions featuring equal rather than unequal obligations; and vice versa for homogeneous players — even though failed implementation usually implied severe cutbacks in monetary payoffs. Both with homogeneous and heterogeneous players, failed implementations arose primarily, but not exclusively, from rejections by the disadvantaged players that profited to a lesser degree from the implemented institution. Consequently, institutions which tailored obligations to players' specific heterogeneities were able to gather higher degrees of support. In fact, if benefits from institution formation were evenly distributed across players, we observed strikingly higher implementation and cooperation rates than what has typically been found in related studies that only require non-unanimous support for institutions to be implemented for all members (e.g., Kosfeld et al., 2009).

A potential reason for the latter finding is that, in contrast to other decision rules, unanimity voting entails a very strong notion of conditional cooperation. The veto right inherent in unanimity voting makes each player's cooperation decision contingent on the decision of all other

players involved. Consequently, the supporting players do not face the risk of being exploited by non-supporting players. On a similar note, no player will ever be governed by an institution that he did not support himself. Both, the notion of conditional cooperation and the retained sovereignty, make unanimity voting an attractive rule to settle on in the first place.

On the other hand, these advantages come at the cost of an increased likelihood of rejecting efficient institutions as well as potentially low levels of cooperation after a rejection has occurred. Already with three players, we saw that these problems exist. With larger groups, one might expect successful institution formation to be even more difficult, in particular if benefits from institution formation are not equally distributed across players. Moreover, our data suggest that voting against the institution is sometimes connected with the implicit costs of making subsequent cooperation more difficult. One might even imagine that rejecting players become the target of retaliation in other, seemingly unrelated, domains. Both threats might be bigger in large groups, simply because there are more players who might potentially opt against the institution and/or who might retaliate rejections. Yet, for groups deciding on the implementation of an institution that takes care of players' idiosyncrasies, these threats might instead strengthen the power of an unanimity rule. Furthermore, under institutions that lead to inequalities in payoffs, payoff differences might be less salient in large groups because they are harder to recognize — in particular if players do not compare themselves with everyone else in a large population, but rather choose a small reference group consisting of similar others. It would therefore be interesting to check in future studies whether the positive or negative effects dominate when group size is increased.

Follow-up studies might also investigate if aggregate behavioral patterns are affected by changes in other parameters of our design, like the marginal per capita return from cooperation or the exact content of the institution. We observed in our data on heterogeneous agents that, overall, the symmetric and asymmetric institution lead to similar average cooperation rates. This was due to two opposing effects that cancel each other out: while the higher implementation rate for the asymmetric institution generally increases cooperation, total obligations (and thus cooperation rates) are lower than when the efficient symmetric institution is implemented. Although this qualitative finding is not at the heart of our paper, it is still intriguing. Given the quantitative behavioral effects that we observe, one could imagine that the gap in average outcomes between symmetric and asymmetric institutions widens as the most efficient payoff-equalizing mechanism becomes more inferior to the efficient mechanism.

Along similar lines, natural next steps for future extensions also include more complex institutional arrangements. For example, redistribution might allay disadvantaged member's doubts about the implementation of efficient institutions for heterogeneous agents. The implementation of institutions with hierarchical structures, from simple leader-follower arrangements to multi-layered structures, yield the potential to increase implementation rates and cooperation, too (e.g., Gächter et al. (2010), Hamman et al. (2011), Falk and Kosfeld (2012)). Complementing these variations, one could also shed more light on the performance of different voting rules for implementing given institutions (e.g., Young (1995), Gillet et al. (2009), Austen-Smith and Feddersen (2006)). More generally, allowing for richer environments with competing institutions and voting rules opens up the possibility to learn even more about the type of institutions that *endogenously* arise within a group. Of course, in contrast to our approach, self-selection would make proper causal interpretation more difficult. Still, it would be a nice complement to the current research agenda: understanding what kind of institutions are created by groups, which voting rules are adopted for implementing these institutions, and how these institutions perform under a variety of circumstances.

Acknowledgements

Financial support from the German Research Foundation (DfG) through SFB-TR 15 is gratefully acknowledged.

References

Andreoni, J. and L. K. Gee (2012). Gun for hire: delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics 96*(11), 1036–1046.

Austen-Smith, D. and T. J. Feddersen (2006). Deliberation, preference uncertainty, and voting rules. *American Political Science Review 100*(2), 209–217.

Banks, J. S., C. R. Plott, and D. P. Porter (1988). An experimental analysis of unanimity in public goods provision mechanisms. *The Review of Economic Studies 55*(2), 301–322.

Bierbrauer, F. J. and M. F. Hellwig (2011). Mechanism design and voting for public-good provision. *MPI Collective Goods Preprint* (2011/31).

Birnberg, J. G., L. R. Pondy, and C. L. Davis (1970). Effect of three voting rules on resource allocation decisions. *Management Science 16*(6), 356–372.

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

Dannenberg, A., A. Lange, and B. Sturm (2010). On the formation of coalitions to provide public goods - experimental evidence from the lab. *National Bureau of Economic Research Working Paper Series No. 15967*.

Decker, T., A. Stiehler, and M. Strobel (2003). A comparison of punishment rules in repeated public good games an experimental study. *Journal of Conflict Resolution 47*(6), 751–772.

Falk, A. and M. Kosfeld (2012). It's all about connections: Evidence on network formation. *Review of Network Economics 11*(3), Article 2.

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

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

Fehr, E. and S. Gächter (2002). Altruistic punishment in humans. *Nature 415*, 137–140.

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

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

Fisher, J., R. M. Isaac, J. W. Schatzberg, and J. M. Walker (1995). Heterogenous demand for public goods: Behavior in the voluntary contributions mechanism. *Public Choice 85*(3-4), 249–266.

Gächter, S., D. Nosenzo, E. Renner, and M. Sefton (2010). Sequential vs. simultaneous contributions to public goods: Experimental evidence. *Journal of Public Economics 94*(7-8), 515–522.

Gerber, A., J. Neitzel, and P. C. Wichardt (2013). Minimum participation rules for the provision of public goods. *European Economic Review 64*, 209–222.

Gillet, J., A. Schram, and J. Sonnemans (2009). The tragedy of the commons revisited: The importance of group decision-making. *Journal of Public Economics 93*(5-6), 785–797.

Greiner, B. (2004). An online recruitment system for economic experiments. In K. Kremer and V. Macho (Eds.), *Forschung und Wissenschaftliches Rechnen*, Number 63 in GWDG Bericht. Gesellschaft für Wissenschaftliche Datenverarbeitung.

Gürerk, O., B. Irlenbusch, and B. Rockenbach (2006). The competitive advantage of sanctioning institutions. *Science 312*(5770), 108–11.

Hamman, J. R., R. Weber, and J. Woon (2011). An experimental investigation of electoral delegation and the provision of public goods. *American Journal of Political Science 55*(4), 738–752.

Hobbes, T. (1651). *Leviathan*. Meiner Verlag (1996), Hamburg.

Isaac, R. and J. Walker (1988). Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics 103*(1), 179–199.

Kamei, K., L. Putterman, and J.-R. Tyran (2014). State or nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods. *Experimental Economics, forthcoming*.

Kesternich, M., A. Lange, and B. Sturm (2014). The impact of burden sharing rules on the voluntary provision of public goods. *Journal of Economic Behavior & Organization 105*, 107–123.

Kosfeld, M., A. Okada, and A. Riedl (2009). Institution formation in public goods games. *American Economic Review 99*(4), 1335–1355.

Linardi, S. and M. A. McConnell (2011). No excuses for good behavior: Volunteering and the social environment. *Journal of Public Economics 95*(5), 445–454.

Maggi, G. and M. Morelli (2006). Self-enforcing voting in international organizations. *The American Economic Review 96*(4), 1137–1158.

Markussen, T., L. Putterman, and J.-R. Tyran (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *The Review of Economic Studies 81*(1), 301–324.

McEvoy, D. M., T. L. Cherry, and J. Stranlund (2011). The endogenous formation of coalitions to provide public goods: Theory and experimental evidence. *SSRN eLibrary*.

Ostrom, E., J. Walker, and R. Gardner (1992). Covenants with and without a sword: Self-governance is possible. *The American Political Science Review 86*(2), 404–417.

Putterman, L., J.-R. Tyran, and K. Kamei (2011). Public goods and voting on formal sanction schemes. *Journal of Public Economics 95*(9), 1213–1222.

Rauchdobler, J., R. Sausgruber, and J. Tyran (2010). Voting on thresholds for public goods: Experimental evidence. *FinanzArchiv: Public Finance Analysis 66*(1), 34–64.

Reuben, E. and A. Riedl (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior 77*(1), 122 – 137.

Rousseau, J. J. (1762). *Der Gesellschaftsvertrag*. Röder-Taschenbuch (1988), Köln.

Shavell, S. and A. M. Polinsky (2000). The economic theory of public enforcement of law. *Journal of Economic Literature 38*(1), 45–76.

Sutter, M., S. Haigner, and M. G. Kocher (2010). Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies 77*(4), 1540–1566.

Tyran, J.-R. and L. P. Feld (2006, March). Achieving compliance when legal sanctions are non-deterrent. *Scandinavian Journal of Economics 108*(1), 135–156.

Wicksell, K. (1964). A new principle of just taxation. In R. A. Musgrave and A. T. Peacock (Eds.), *Classics in the Theory of Public Finance*. London: Macmillan.

Young, P. (1995). Optimal voting rules. *Journal of Economic Perspectives 9*(1), 51–64.

A Model and Theoretical Predictions

A.1 Model

We study the following two-stage game in which players have perfect information on other players' preferences:

Voting stage: First, all players simultaneously and independently vote either in favor of or against adopting an institution. The institution specifies a contribution level that each player is obliged to contribute to the public good and introduces sanctions for deviant contribution levels. Sanctions are sufficiently severe to ensure that the prescribed contribution levels are indeed implemented.

Contribution stage: Second, all players simultaneously and independently choose their contribution level to the public good. If the institution has been implemented, players will contribute the amount specified by the institution. If the institution has not been implemented, there is no sanctioning mechanism and players play a standard public goods game (VCM).

In the contribution stage, players know how other players in their group voted in the voting stage. In the following, we demonstrate that rejecting the institution can increase utility in some treatments, while not in others. A multitude of equilibria exists. In order to keep the subsequent analysis tractable and short, when analyzing equilibria in which at least one player rejects, we focus on those equilibria in which rejecting the institution strictly increases the rejecting player's utility.[21] For each treatment, we will first characterize equilibria if players' utility functions coincide with the monetary payoff of the game, π_i, i.e., if players have standard preferences. We will then proceed by analyzing equilibria of the game if (some) players have social preferences, i.e., suffer from inequality in monetary payoffs (compare, among others, Fehr and Schmidt (1999)). Fehr and Schmidt (1999) assume that players compare their own monetary payoff with the monetary payoff of all other players. They introduce the following utility function:

$$U_i = \pi_i - \alpha_i \frac{1}{n-1} \sum_{j=1}^{n} max\{\pi_j - \pi_i; 0\} - \beta_i \frac{1}{n-1} \sum_{j=1}^{n} max\{\pi_i - \pi_j; 0\}$$

The first term represents the monetary payoff obtained in the game. The second term captures utility losses due to being worse off than other players. α_i measures the degree of individual envy.

[21] There also exist equilibria, in which players reject the institution although the resulting utilities are lower than in the state of successful institution formation: As soon as one player rejects, the decision of the other players does not affect institution formation under unanimity voting. Consequently, further equilibria exist in which at least two players reject the institution.

The last term denotes utility losses that players receive from being better off than other players. β_i is typically interpreted as a measure for the degree of compassion. Additionally, two important properties are assumed. First, $\alpha_i \geq \beta_i$ or, in words, envy is at least as strong as compassion. Second, $\beta_i < 1$, which prevents agents from "burning their own money" to achieve a more equal outcome. In our setup with heterogeneous benefits from the public good, players might vote against implementing an institution that obliges all players to contribute equally to the public good in order to avoid inequality. Hence, we consider the model of Fehr and Schmidt (1999) as a natural choice to derive predictions for our setup.

In the following, subscript l (h) stands for low (high) type, i.e., players with a low (high) MPCR. Given the focus of this paper, (only) the analysis of the treatments featuring heterogeneous players with social preferences focuses on the case of three players: one low type with low MPCR γ_l, two high types with high MPCR γ_h, and $\Delta\gamma = \gamma_h - \gamma_l < 1/2$. Additionally, the propositions presented in the text further specify results by setting $\gamma_l = 1/2$ and $\gamma_h = 3/4$, the parameters we have used in the experiment implementation.

A.2 Treatments $HOM - VCM$ and $HET - VCM$

The standard Voluntary Contribution Mechanism (VCM) is a one-stage game without voting on implementing an institution and without any sanctioning mechanism for low contributions.

Proposition 1 *If players are money-maximizers, they contribute $c_i = 0 \forall i$ in treatments $HOM - VCM$ and $HET - VCM$.*

Whenever $\frac{\partial \pi_i}{\partial c_i} = -1 + \gamma_i < 0$, the marginal individual cost of contributing to the public good exceeds the marginal individual benefit. Consequently, in any standard VCM game, a money-maximizing player will not contribute to the public good for all $\gamma_i < 1$. Condition $\gamma_i < 1$ is met by definition of the public goods game for all players in treatments $HOM - VCM$ ($\gamma = 2/3$) and $HET - VCM$ ($\gamma_l = 1/2$ and $\gamma_h = 3/4$).

Proposition 2 *Let us assume that players have social preferences.*
In treatment $HOM - VCM$, if $\gamma_i + \beta_i < 1$ for at least one player, there is a unique equilibrium in which all players contribute $c_i = 0$. If all players have $\gamma_i + \beta_i > 1$, other equilibria with $c_i > 0$ exist, in which all players contribute $c_i = c_j \in [0, E], \forall j \neq i$.
In treatment $HET - VCM$, if $\beta_h > 2/7$ for both high types and $\beta_l > 2/5$ for the low type,

equilibria with positive contributions exist in which $c_h \in [0, E]$ *and* $c_l = 2/5c_h$*. All players earn equal payoffs. Otherwise, there exists a unique equilibrium in which all players contribute* $c_i = 0$*.*

The proof of $HOM - VCM$ is provided in Fehr and Schmidt (1999). The intuition is as follows: If players are sufficiently averse to advantageous inequality (β sufficiently high), they are willing to exactly match the contribution levels of the other players to equalize payoffs. Using the parameters of our experiment, the proposition boils down to the result that equilibria with positive contribution levels only exist if $\beta > 1/3$ for all players.

In treatment $HET - VCM$, the basic mechanism that drives the existence of equilibria with positive contributions is the same as in the VCM with homogeneous players. If players are sufficiently averse towards earning more than others, they contribute positive amounts to prevent an unequal payoff distribution as soon as other players contribute a positive amount. To achieve an equal payoff distribution, the low type contributes less than the high types. In the following, we provide a formal analysis of the behavior of players with social preferences in treatment $HET - VCM$.

We start by analyzing the **behavior of the low type**.

Case 1: $\pi_l \leq \pi_1$ and $\pi_l \leq \pi_2$

Let us assume that the low type contributes such that her monetary payoff is not larger than the payoff of both high types (that are labelled by indices 1 and 2). Then, the utility function of the low type that is relevant for the marginal analysis is denoted by: $U_l = E - c_l + \gamma_l(c_l + c_1 + c_2) - \frac{\alpha_l}{2}(c_l - c_1 + \Delta\gamma(c_l + c_1 + c_2)) - \frac{\alpha_l}{2}(c_l - c_2 + \Delta\gamma(c_l + c_1 + c_2))$. The derivative with respect to c_l is given by $\frac{\partial U_l}{\partial c_l} = -1 + \gamma_l - \alpha_l(1 + \Delta\gamma)$ and will always be negative as $\gamma_l < 1$ and $\Delta\gamma \geq 0$. Hence the low type will never increase her contribution, but at least decrease her contribution until $\pi_l = \pi_1 \leq \pi_2$ or $\pi_l = \pi_2 \leq \pi_1$. In sum, the low type will never contribute such that her payoff will be lower than the payoffs of both high types.

Case 2: $\pi_1 < \pi_l < \pi_2$ or $\pi_2 < \pi_l < \pi_1$

If the low type's payoff is larger than the payoff of one high type, but still smaller than the other high type's payoff, the derivative of the utility function is given by $\frac{\partial U_l}{\partial c_l} = -1 + \gamma_l - 1/2(\alpha_l - \beta_l)(1 + \Delta\gamma)$. This derivative is strictly negative, as disadvantageous inequality is assumed to affect utility at least as strong as advantageous inequality ($\alpha_i \geq \beta_i$), $\gamma_l < 1$, and $\Delta\gamma \geq 0$. The low type will decrease her contribution until her payoff equals the payoff of the better off high type. Intuitively, by reducing her contribution the low type will increase her own monetary payoff and simultaneously decrease disutility from disadvantageous inequality at a faster rate than increasing

disutility from advantageous inequality.

Case 3: $\pi_l > \pi_1$ and $\pi_l > \pi_2$

Let us now assume that the payoff of the low type is strictly larger than the payoffs of both high types. The utility function that is relevant for the marginal analysis is now denoted by $U_l = E - c_l + \gamma_l(c_l + c_1 + c_2) - \frac{\beta_l}{2}(c_1 - c_l - \Delta\gamma(c_l + c_1 + c_2)) - \frac{\beta_l}{2}(c_2 - c_l - \Delta\gamma(c_l + c_1 + c_2))$. Thus, $\frac{\partial U_l}{\partial c_l} = -1 + \gamma_l + \beta_l(1 + \Delta\gamma)$, which is positive if $\beta_l > \frac{1-\gamma_l}{1+\Delta\gamma}$. If this condition is fulfilled, the low type will contribute in such a way that her payoff will equal the payoff of the high type with the lower contribution to the public good. If $\beta_l < \frac{1-\gamma_l}{1+\Delta\gamma}$, the low type does not contribute to the public good at all since she does not suffer sufficiently from advantageous inequality.

The next section analyzes **behavior of** one **high type**, player 1, given the actions of the other high type, player 2, and the low type l. Without loss of generality, we will only analyze the decisions of high type 1 who is representative for behavior of both high types.

Case 1: $\pi_1 < \pi_l$ and $\pi_1 < \pi_2$

If player 1 obtains the lowest monetary payoff, $U_1 = E - c_1 + \gamma_h(c_l + c_1 + c_2) - \frac{\alpha_1}{2}(c_1 - c_l - \Delta\gamma(c_l + c_1 + c_2)) - \frac{\alpha_l}{2}(c_1 - c_2)$. The derivative $\frac{\partial U_1}{\partial c_1} = -1 + \gamma_h - \alpha_1(1 - \frac{\Delta\gamma}{2})$ is always negative as $\gamma_h < 1$ and $\Delta\gamma \leq 1/2$. Thus, player 1 will never increase his contribution, but, in contrast, decrease it until his payoff at least equals the payoff of one other player. By reducing his contribution, player 1 can increase his monetary payoff and simultaneously decrease inequality.

Case 2: $\pi_l \leq \pi_1 < \pi_2$

Player 1 is worse off than the other high type, but weakly better off than the low type. As the analysis of the low type's behavior has shown, this case can never arise.

Case 3: $\pi_2 < \pi_1 < \pi_l$

Player 1 is better off than the other high type, but worse off than the low type. The utility function that is relevant for the marginal analysis is given by $U_1 = E - c_1 + \gamma_h(c_l + c_1 + c_2) - \frac{\alpha 1}{2}(c_1 - c_l - \Delta\gamma(c_l + c_1 + c_2)) - \frac{\beta_1}{2}(c_2 - c_1)$. Setting the derivative $\frac{\partial U_1}{\partial c_1} = -1 + \gamma_h - \frac{\alpha_1}{2}(1 - \Delta\gamma) + \frac{\beta_1}{2}$ larger than zero, results in the condition $\beta_1 > 2(1 - \gamma_h) + \alpha_1(1 - \Delta\gamma)$. If this condition is met, player 1 will match the contribution of the other high type no matter what the low type does. The low type may either choose her contribution to equalize payoffs of all three players or not contribute to the public good at all. In the following, equilibria that result in unequal payoffs will be called asymmetric.

With the parameters chosen in our experiment the condition for asymmetric equilibria is reduced to $\beta_1 > \frac{1}{2} + \frac{3}{4}\alpha_1$. This condition can never be satisfied. Consider the limiting case of $\alpha_1 \geq \beta_1$:

$\alpha_1 = \beta_1$. This results in $\frac{\beta_1}{4} > \frac{1}{2}$, which cannot hold as $\beta_i < 1$ is another assumption of the Fehr-Schmidt model.

If $\beta_1 < 2(1-\gamma_h) + \alpha_1(1-\Delta\gamma)$, player 1 will at least reduce his contribution until $\pi_2 < \pi_1 = \pi_l$.

Case 4: $\pi_1 > \pi_l$ and $\pi_1 > \pi_2$

If the payoff of player 1 is larger than the payoffs of the two other players, his utility function is now denoted by: $U_1 = E - c_1 + \gamma_h(c_l + c_1 + c_2) - \frac{\beta_1}{2}(c_l - c_1 + \Delta\gamma(c_l + c_1 + c_2)) - \frac{\beta_1}{2}(c_2 - c_1)$. The derivative $\frac{\partial U_1}{\partial c_1} = -1 + \gamma_h + \beta_1(1 - \frac{\Delta\gamma}{2})$ turns positive for $\beta_1 > \frac{1-\gamma_h}{1-\frac{1}{2}\Delta\gamma}$. This implies that for sufficiently large values of β_1, player 1 will increase his contribution to the public good until at least one other player obtains the same payoff as he does. Intuitively, a player 1 who is sufficiently averse to advantageous inequality will contribute in order to reduce inequality towards both other players. If $\beta_1 < \frac{1-\gamma_h}{1-\frac{1}{2}\Delta\gamma}$, player 1 will not contribute at all.

In the following, we summarize the resulting equilibria:

In treatment $HET-VCM$, if $\beta_h < \frac{1-\gamma_h}{1-\frac{1}{2}\Delta\gamma}$ for at least one high type player, there exists a unique equilibrium in which all players contribute $c_i = 0$.

If $\beta_h > \frac{1-\gamma_h}{1-\frac{1}{2}\Delta\gamma}$ for both high types and $\beta_l > \frac{1-\gamma_l}{1+\Delta\gamma}$ for the low type, equilibria with positive contributions exist with $c_h \in [0, E]$ and $c_l = c_h \frac{1-2(\gamma_h-\gamma_l)}{1+(\gamma_h-\gamma_l)}$ (symmetric equilibria).

If $\beta_l < \frac{1-\gamma_l}{1+\Delta\gamma}$ for the low type and $\beta_h > 2(1-\gamma_h) + \alpha_h(1-\Delta\gamma)$ for both high types, another class of equilibria with $c_i \geq 0$ exists, in which both high types contribute the same amount $c_h \in [0, E]$ and the low type contributes $c_l = 0$ (asymmetric equilibria). Proposition 2 summarizes results using the parametrization of our experiment.

A.3 Two-stage treatments with voting stage and contribution stage

In all two-stage treatments, players are assumed to apply backward induction. Let U^{INST} denote utility when the institution has received unanimous support and has been implemented, with $INST = SYM$ for the symmetric and $INST = ASYM$ for the asymmetric institution. In the contribution stage, players will compare the utility they receive with the respective institution being in place, U^{INST}, to U^{VCM}, the utility of the VCM that is played if the institution has not received unanimous support in the voting stage. Whenever $U^{INST} \geq U^{VCM}$, a player will vote in favor of implementing the institution. With unanimity voting, if all other players also vote in favor of implementing the proposed institution, the institution will be implemented and the player's preferred outcome is achieved. If, in contrast, at least one other player votes against implementing the institution, the institution will not be implemented and the VCM will be played.

However, the approving player is still equally well off as if he had voted against implementing the institution. Thus, unanimity voting ensures that it is always a best response to the voting behavior of the other players to vote in favor of the institution if $U^{INST} \geq U^{VCM}$. A player will never be hurt from voting for his preferred outcome no matter how the other players vote. Whenever $U^{INST} < U^{VCM}$, a player will vote against installing the institution.

A.3.1 Treatment $HOM - SYM$

In treatment $HOM - SYM$, homogeneous players vote on implementing the symmetric institution.

Proposition 3 *The following statements hold both for money-maximizing players and for players with social preferences. In treatment $HOM - SYM$, all players vote in favor of implementing the institution. The symmetric institution is always implemented and all players contribute according to the institutional rules, i.e., $c_i = E \forall i$.*

If players are homogeneous ($\gamma_i = \gamma$) and **money-maximizing**, they compare $U^{SYM} = \gamma nE$ to $U^{VCM} = E$ to decide on voting in favor of or against the symmetric institution that requires each player to contribute the efficient contribution level E. $\gamma nE > E$ if $\gamma > 1/n$, a condition that is always met by definition in a VCM game with homogeneous players. Consequently, with unanimity voting, all players will vote in favor of the symmetric institution.
In treatment $HOM - SYM$, assuming **social preferences** instead of pure money-maximizing does not change predictions. With homogeneous players, the symmetric institution guarantees equality of payoffs while simultaneously maximizing them. Hence again, all players are predicted to vote in favor of the symmetric institution. Formally, $U^{SYM} = \gamma nE \geq U^{VCM} = E - \hat{c} + \gamma n\hat{c}$, where $\hat{c}$ denotes contributions in equilibrium with $\hat{c} \in [0, E]$. As U^{VCM} is strictly increasing in $\hat{c}$ due to $n\gamma > 1$, the utility after successful implementation of the symmetric institution is always at least as large as the utility from the VCM. This analysis is equivalent to the one done by Gerber et al. (2011) for the 4 player case.

A.3.2 Treatment $HET - SYM$

In treatment $HET - SYM$, heterogeneous players vote on implementing the symmetric institution.

Proposition 4 *In treatment $HET - SYM$, money-maximizing players vote in favor of implementing the symmetric institution. The symmetric institution is always implemented and all players contribute according to the institutional rules, i.e., $c_i = E$, $i \in \{h, l\}$.*

If players are heterogeneous and **money-maximizing**, they compare $U^{SYM} = \gamma_i nE$ with $\gamma_i \in \gamma_l, \gamma_h$ to $U^{VCM} = E$ to decide on voting in favor of or against the symmetric institution that requires each player to contribute the efficient contribution level E. $U^{SYM} > U^{VCM}$ whenever $\gamma_i > 1/n$. Given the parametrization of our experiment ($\gamma_l = 1/2$, $\gamma_h = 3/4$, $n = 3$), this condition is met for both high and low type players. Consequently, all players vote in favor of the symmetric institution.

In treatment $HET - SYM$, predictions based on standard preferences and **social preferences** differ markedly. Players with standard preferences always support the formation of the symmetric institution as it offers a higher monetary payoff than the VCM and they do not suffer from inequality that arises from symmetric contributions of players with different MPCRs. In contrast, low type players with social preferences who suffer sufficiently from being worse off than the high types if the symmetric institution is implemented object to institution formation. They prefer a possibly lower payoff, but equal payoffs across players in the VCM to a higher monetary payoff, but disutility from inequality with the symmetric institution being in place.

Proposition 5 *High type players with social preferences will always vote in favor of installing the symmetric institution. In contrast, low type players with social preferences will reject the installation of the symmetric institution if they are sufficiently averse to disadvantageous inequality, more precisely, if $\alpha_l > \frac{2}{3} - \frac{4}{75}\hat{c_h}$, where $\hat{c_h}$ is the equilibrium contribution of high types in the VCM. If players have social preferences, the symmetric institution will not always be implemented.*

The proof of proposition 5 is provided below. We first analyze the **behavior of the low type**. If the symmetric institution is implemented, the low type's utility is $U_l^{SYM} = 3E(\gamma_l - \alpha_l \Delta\gamma)$. As has been shown in the previous analysis of treatment $HET - VCM$, if players have social preferences and heterogeneous MPCRs the VCM has both symmetric and asymmetric equilibria. Hence, U_l^{VCM} depends on the kind of equilibrium that is played in the VCM. If a symmetric equilibrium is played, the utility of the low type is $U_{l,sym}^{VCM} = E + c_h(\frac{2\gamma_h + \gamma_l - 1}{1 + \Delta\gamma})$, where c_h denotes the contribution level of the high types in the VCM. The low type will reject the symmetric institution, if $U_{l,sym}^{VCM} > U_l^{SYM}$, i.e., if $\alpha_l > \frac{3\gamma_l - 1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h + \gamma_l - 1}{1 + \Delta\gamma})$.

If an asymmetric equilibrium is played in the VCM, utility in the VCM is $U_{l,asym}^{VCM} = E + \gamma_l 2c_h - \beta_l(1 - 2\Delta\gamma)$. The critical threshold for rejecting the symmetric institution is given by $\alpha_l > \frac{3\gamma_l - 1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(2\gamma_l - \beta_l(1 - 2\Delta\gamma))$.

Next, we will analyze the voting **behavior of the high types**. Again, we must distinguish between symmetric and asymmetric equilibria being played in the VCM. If a symmetric equilibrium is played in the VCM, all players' payoffs are equal: $U_{h,sym}^{VCM} = U_{l,sym}^{VCM} = E + c_h(\frac{2\gamma_h + \gamma_l - 1}{1 + \Delta\gamma})$. If $U_{h,sym}^{VCM} > U_h^{SYM}$, the symmetric institution will be rejected. Setting $U_{h,sym}^{VCM} > U_h^{SYM} = 3E(\gamma_h - \frac{\beta_1}{2}\Delta\gamma)$, leads to the condition $\beta_h > \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{2c_h}{3E\Delta\gamma}(\frac{2\gamma_h + \gamma_l - 1}{1 + \Delta\gamma})$. This can be interpreted as follows: if the high types' sensitivity towards advantageous inequality and the contributions in the VCM are large enough, high types will vote against the symmetric institution to achieve an outcome with equal payoffs, rather than potentially higher, but unequal payoffs.

Íf an asymmetric equilibrium is played in the VCM, the utility of the high types is given by $U_{h,asym}^{VCM} = E + c_h(2\gamma_h - 1 - \frac{\alpha_h}{2}(1 - 2\Delta\gamma))$. Rearranging $U_{h,asym}^{VCM} > U_h^{SYM}$ leads to $\beta_h > \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{c_h}{E}(2\gamma_h - 1 - \frac{\alpha_1}{2}(1 - 2\Delta\gamma))$.

Let us summarize behavior of players with social preferences in treatment $HET - SYM$: If symmetric equilibria are played in the VCM, the low type votes against implementing the symmetric institution if $\alpha_l > \frac{3\gamma_l - 1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h + \gamma_l - 1}{1 + \Delta\gamma})$. The high types vote against implementing the symmetric institution if $\beta_h > \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{2c_h}{3E\Delta\gamma}(\frac{2\gamma_h + \gamma_l - 1}{1 + \Delta\gamma})$. If an asymmetric equilibrium is played in the VCM, the low type rejects the institution if $\alpha_l > \frac{3\gamma_l - 1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(2\gamma_l - \beta_l(1 - 2\Delta\gamma))$, while the high types reject it for $\beta_h > \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{c_h}{E}(2\gamma_h - 1 - \frac{\alpha_h}{2}(1 - 2\Delta\gamma))$. If the institution is not implemented, contribution levels are identical to those in the treatment $HET - VCM$. If the symmetric institution is implemented, all players contribute $c_i = E$, $i \in \{h, l\}$.

Using the parametrization of the experiment simplifies results drastically. Equilibria with asymmetric payoffs cannot arise. Low types will reject the symmetric institution if $\alpha_l > \frac{2}{3} - \frac{4}{75}c_h$. High types will reject the institution for $\beta_h > \frac{10}{3} - \frac{8}{75}c_h$. Since $\beta < 1$ by assumption of the Fehr-Schmidt model and $c_h \in [0, 20]$, this condition is never met and high types will never reject the institution.

A.3.3 Treatment $HET - ASYM$

In treatment $HET - ASYM$, heterogeneous agents vote on implementing the asymmetric institution.

Proposition 6 *The following statements hold both for money-maximizing players and for players with social preferences. In treatment $HET-ASYM$, all players vote in favor of implementing the asymmetric institution. The institution is always implemented and all players contribute according to the institutional rules, i.e., high types contribute $c_h = E$ and low types contribute c_l.*

If the asymmetric institution has been implemented the utility of **money-maximizing** low types is denoted by $U_l^{ASYM} = E - c_l + \gamma_l(n_1 E + n_2 c_l)$, the utility of money-maximizing high types by $U_h^{ASYM} = \gamma_h(n_1 E + n_2 c_l)$. Contribution levels in the asymmetric institution are designed to equalize payoffs across heterogeneous player types, i.e., the contribution level of the low types, c_l, is determined by $U_l^{ASYM} = U_h^{ASYM}$. Solving for c_l and restricting contributions to be non-negative results in $c_l = \max\{E\frac{1-n_1\Delta\gamma}{1+n_2\Delta\gamma}; 0\}$. Whenever $U_l^{ASYM} = U_h^{ASYM} > U^{VCM} = E$, players vote in favor of the asymmetric institution. Inserting c_l and rearranging $U_l^{ASYM} = U_h^{ASYM} > E$ leads to $\gamma_h n_1 + \gamma_l n_2 > 1$, which is the necessary condition for public good provision to be efficient, a condition that is met by definition of the public goods game.

Both high and low type players with **social preferences** will vote in favor of implementing the asymmetric institution. Implementing the asymmetric institution guarantees both player types the highest attainable payoff among all equilibrium payoffs of the VCM and does not induce payoff inequalities. This intuitive line of reasoning summarizes the part of the formal analysis provided below that is relevant for the parameters used in the experiment. In sum, in treatment $HET-ASYM$, we will show that the low type with social preferences will only reject the asymmetric institution if an asymmetric equilibrium is played in the VCM. Otherwise, the utility level with the asymmetric institution in place represents the highest attainable equilibrium utility level and the implementation of the asymmetric institution will always be supported.

If a symmetric equilibrium is played in the VCM, the **low type** will compare the utility obatined under the asymmetric institution, $U_l^{ASYM} = U_h^{ASYM} = \gamma_h E(\frac{3}{1+\Delta\gamma})$, to the utility level in a symmetric equilibrium of the VCM, $U_{l,sym}^{VCM} = E + c_h(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$. Simplifying $U_l^{ASYM} \geq U_{l,sym}^{VCM}$ results in the condition $E \geq c_h$ that is always met. Consequently, the low type will support the implementation of the asymmetric institution.
If an asymmetric equilibrium is played in the VCM, the low type will compare U_l^{ASYM} to $U_{l,asym}^{VCM} = E + \gamma_l 2c_h - \beta_l c_h(1 - 2\Delta\gamma)$. Setting $U_{l,asym}^{VCM} > U_l^{ASYM}$ and rearranging results in $\frac{c_h}{E} > \frac{2\gamma_h+\gamma_l-1}{(1+\Delta\gamma)(2\gamma_l-\beta_l)(1-2\Delta\gamma)}$, i.e., the low type will only reject the asymmetric institution if and only if the amount contributed to the public good in the asymmetric equilibrium of the VCM

is relatively large compared to the total endowment and if the conditions for the existence of an asymmetric equilibrium in the VCM are met, i.e., if $\beta_l < \frac{1-\gamma_l}{1+\Delta\gamma}$ for the low type and $\beta_h > 2(1-\gamma_h) + \alpha_h(1-\Delta\gamma)$ for both high types.

High types will always support the implementation of the asymmetric institution. First, the asymmetric institution guarantees them the highest attainable utility level in the VCM among all possible symmetric equilibria of the VCM. Second, also if an asymmetric equilibrium is played in the VCM, the high types' utility obtained under the asymmetric institution must at least be as large as the utility in every possible asymmetric equilibrium of the VCM, since the low type additionally contributes a non-negative amount to the public good and equal payoffs are ensured. Technically, let us compare the high types' utility from the asymmetric equilibrium in the VCM, $U^{VCM}_{h,asym} = E + c_h(2\gamma_h - 1 - \frac{\alpha_1}{2}(1-2\Delta\gamma))$ to the utility from the asymmetric institution $U^{ASYM}_h = E\gamma_h(\frac{3}{1+\Delta\gamma}) = E\gamma_h(2 + \frac{1-2\Delta\gamma}{1+\Delta\gamma})$. Setting $U^{VCM}_{h,asym} > U^{ASYM}_h$ results in $c_h(2\gamma_h - 1 - \frac{\alpha_1}{2}(1-2\Delta\gamma)) > E(\gamma_h(2 + \frac{1-2\Delta\gamma}{1+\Delta\gamma}) - 1)$. Since $E \geq c_h$, $-\frac{\alpha_1}{2}(1-2\Delta\gamma) > \gamma_h(\frac{1-2\Delta\gamma}{1+\Delta\gamma})$ which can be simplified to $-\frac{\alpha_1}{2} > \frac{\gamma_h}{1+\Delta\gamma}$ must hold for $U^{VCM}_{h,asym} > U^{ASYM}_h$ to be true. However, $-\frac{\alpha_1}{2} > \frac{\gamma_h}{1+\Delta\gamma}$ can never be true, since the left side of the inequality is negative, while the right one is positive. Consequently, high types will always vote in favor of implementing the asymmetric institution.

For the parameters used in the laboratory experiment, asymmetric equilibria in the VCM do not exist. The only source of rejecting the asymmetric institution is eliminated and all players are predicted to vote in favor of implementing the asymmetric institution.

A.3.4 Treatment $HOM - ASYM$

In treatment $HOM - ASYM$, homogeneous players vote on implementing the asymmetric institution.

Proposition 7 *For money-maximizing players, it is a weakly dominant strategy to vote in favor of implementing the asymmetric institution in treatment $HOM - ASYM$. The asymmetric institution is always implemented and all players contribute according to the institutional rules.*

The asymmetric institution obliges n_1 players to contribute their whole initial endowment E, while the other n_2 players are obliged to contribute only $\bar{c} < E$ with $n_1 + n_2 = n$. **Money-maximizing players** who are obliged to only contribute $\bar{c}$ will vote in favor of the asymmetric institution because it will increase their earnings: $U^{ASYM}_{\bar{c}} = E - \bar{c} + \gamma(n_1 E + n_2 \bar{c}) > E$, the

payoff in the VCM, since $\bar{c} < E$ and $\gamma(n_1 + n_2) > 1$. However, for players who are obliged to contribute E, the formation of the asymmetric institution does not pay off when the share of players with low contributions gets too large or these players' contribution level $\bar{c}$ gets too small. Their payoff from the asymmetric institution is denoted by $U_E^{ASYM} = \gamma(n_1 E + n_2 \bar{c})$. Only if $\gamma(n_1 E + n_2 \bar{c}) > E$, it is a weakly dominant strategy for all players to support the installment of the asymmetric institution. For the parametrization of our experiment ($\gamma = 2/3$, $n_1 = 2$, $n_2 = 1$, $E = 20$, and $\bar{c} = 8$), this is indeed the case.

Players with **social preferences** will reject the asymmetric institution if the inequality introduced by the asymmetric institution outweighs its monetary gains. The utility of the two players who contribute fully is denoted by $U_E^{ASYM} = \gamma(n_1 E + n_2 \bar{c}) - \alpha_i 1/2(E - \bar{c})$, the utility of the player who contributes $\bar{c}$ is given by $U_{\bar{c}}^{ASYM} = E - \bar{c} + \gamma(n_1 E + n_2 \bar{c}) - \beta_i(E - \bar{c})$. For three players with social preferences, $U^{VCM} = E - \hat{c} + \gamma(n_1 + n_2)\hat{c}$, where $\hat{c}$ denotes the equilibrium contribution to the public good in the VCM. Comparing U_E^{ASYM} to U^{VCM} shows that the players contributing fully will reject the asymmetric institution if $\alpha_E > 2\frac{(\hat{c}-E)+\gamma(n_1 E+n_2\bar{c}-(n_1+n_2)\hat{c})}{E-\bar{c}}$, while the player contributing $\bar{c}$ will reject it if $\beta_{\bar{c}} > \frac{(\hat{c}-\bar{c})+\gamma(n_1 E+n_2\bar{c}-(n_1+n_2)\hat{c})}{E-\bar{c}}$. Inserting the experimental parameters, the two conditions simplify to $\alpha_E > 2 - \frac{\hat{c}}{6}$ and $\beta_{\bar{c}} > 2 - \frac{\hat{c}}{12}$. These inequalities also show that the asymmetric institution is more attractive if equilibrium contributions in the VCM $\hat{c}$ are low.

Proposition 8 *Players with social preferences who are obliged to contribute fully will vote against the asymmetric institution if $\alpha_{\bar{c}} > 2 - \frac{\hat{c}}{6}$, while players who are obliged to contribute $\bar{c}$ will vote against the asymmetric institution if $\beta_E > 2 - \frac{\hat{c}}{12}$. Hence, if homogeneous players have social preferences, the asymmetric institution will not always be implemented.*

B Translated Instructions

The instructions below are translations of the German instructions for treatment $HET-ASYM$. Instructions for the other treatments were as similar as possible except for the necessary adjustments concerning the composition of types (in treatments with homogeneous players), the level of obligations (in treatments with the symmetric institution), and the omittance of the first stage in the baseline VCM treatments.

General explanations to the participants

You are now participating in an economic experiment. If you read the following explanations carefully, you will be able to earn a considerable amount of money – depending on your decisions and those of the other participants. Thus it is very important to read these instructions carefully and to understand them.

During the experiment, it is absolutely prohibited to communicate with the other participants. If you have any questions, please ask us: please raise your hand and we will come to your seat. If you violate this rule, you will be dismissed from the experiment and forfeit all payments.

How much money you will receive after the experiment depends on your decisions and those of the other participants. During the experiment, payoffs will be calculated in Taler instead of Euro. Your total income will be calculated in Taler first. The total amount of Taler that you have accumulated during the experiment will be converted into Euro and paid to you in cash at the end of the experiment. The exchange rate from Taler to Euro is as follows:

$$40 \text{ Taler} = 1 \text{ Euro}$$

The experiment consists of exactly one part. This part is divided into **20 periods**. At the beginning of the experiment you are randomly assigned to a group of three. Thus, there are two other participants in your group. In each group of three, there are **two participants of type A** and **one participant of type B** (the difference between type A and type B will be explained in detail shortly). Whether you are of type A or of type B is determined randomly. **In all periods your type remains the same, just as the types of the other participants in your group remain the same**. You will be interacting with the same two participants in all periods. Neither

during, nor after the experiment will you receive any information about the identities of the other participants in your group.

Each period is divided into three stages:

1. In the **second stage** you have to decide on how many Taler you contribute to a project and how many Taler you keep for yourself.

2. In the **first stage** you can decide if you want to commit yourself and the other participants in your group to certain contributions to the project in stage 2. Only if **all** participants decide in stage 1 to commit all participants in your group to certain contributions to the project, the contributions will actually be fixed. If not all participants decide to fix the contributions, then you and the other participants in your group will be able to choose any contribution level in the second stage.

3. In the **third stage** you get to know the contributions of all participants in your group to the project in stage 2 and the payoffs of all participants in your group in this period.

Detailed information about the course of a period

At the beginning of each period every participant receives **20 Taler**. In each period you have to decide on how to use these 20 Taler. You can contribute Taler to a **project** or put them on a **private account**. Every Taler that you don't contribute to the project is automatically put on your private account.

Income from your private account:

For each Taler you put on your private account, you earn exactly one Taler. For example, if you put 20 Taler on your private account (thus contributing zero Taler to the project), you would earn 20 Taler from your private account. If, e.g., you would put 2 Taler on your private account (thus contributing 18 Taler to the project), your income from the private account would be 2 Taler. Nobody but you receives Taler from your private account.

Income from the project:

For each Taler that you or another participant in your group contributes to the project, you (and each other participant in your group) earn a certain number of Taler. Each participant's income from the project depends on his or her type and is determined as follows:

*Type A's income from the project = $\frac{3}{4}$ * sum of all contributions to the project*

*Type B's income from the project = $\frac{1}{2}$ * sum of all contributions to the project*

Example 1: The sum of contributions from all participants to the project is 12 Taler (e.g. if you and the two other participants contribute 4 Taler each, or if one of the three participants contributes 12 Taler and the two other participants contribute 0 Taler). Then the two participants in your group who are of type A each receive an income of $\frac{3}{4}$ * 12 = 9 Taler from the project, and the participant in your group who is of type B receives an income of $\frac{1}{2}$ * 12 = 6 from the project.

Example 2: The sum of contributions from all participants to the project is 36 Taler. Then the two participants in your group who are of type A each receive an income of $\frac{3}{4}$ * 36 = 27 Taler from the project, and the participant in your group who is of type B receives an income of $\frac{1}{2}$ * 36 = 18 from the project.

Income at the end of a period:

Your income at the end of a period is the sum of your income from your private account and your income from the project:

Type A:
Income from the private account (20 – contribution to the project)
*+ Income from the project ($\frac{3}{4}$ * sum of contributions to the project)*

= Income at the end of the period

Type B:
Income from the private account (20 – contribution to the project)
*+ Income from the project ($\frac{1}{2}$ * sum of contributions to the project)*

= Income at the end of the period

Let us illustrate how your income at the end of a period is calculated using two examples:

Example 1: Assume that you are of type A and contribute 16 Taler to the project, just as the other two participants. The sum of contributions is then 16 + 16 + 16 = 48 Taler. Your income in this example would be

4 Taler from the private account + $\frac{3}{4}$ * 48 Taler from the project = 4 + 36 = 40 Taler

Example 2: Assume that you are of type A and contribute 0 Taler to the project, while the other two participants contribute 16 Taler each. The sum of contributions is then 16 + 16 + 0 = 32 Taler. Thus, your income would be

$$\underline{20 \text{ Taler}} \text{ from the private account} + \underline{\frac{3}{4} * 32 \text{ Taler}} \text{ from the project} = 20 + 24 = \underline{44 \text{ Taler}}$$

The first stage

In the **first stage** you can decide whether you want to commit yourself and the other participants in your group to a certain contribution to the project in the second stage. All participants decide simultaneously. Only if **all** participants in your group decide to commit themselves and the other participants to certain contributions, are the contributions in stage 1 actually fixed. In this case contributions will be fixed as follows:

Type A: *Contribution of 20 Taler to the project*

Type B: *Contribution of 8 Taler to the project*

If **not all** participants decide to fix the contributions, you and the other participants in your group can freely contribute any number of your 20 Taler to the project in the second stage.

The second stage

At the beginning of the second stage you get to know how each participant in your group decided in the first stage.

If in the first stage all participants decided to fix the contributions in the second stage, then in the second stage you have to contribute the corresponding amount. Thus, if you are of type A you have to enter a contribution of 20 Taler and if you are of type B you have to enter a contribution of 8 Taler. Other inputs are not possible and will automatically be adjusted by the computer program.

In this case the period income of the participants of type A is $\frac{3}{4} * 48 = 36$ Taler each and the period income of the participant of type B is $12 + \frac{1}{2} * 48 = 36$ Taler.

If in the first stage not all participants decided to fix the contributions in the second stage, then in the second stage all participants can freely choose any integer contribution between 0 and 20 to the project (0, 1, 2, . . . , 19, 20).

In this case your period income is computed as indicated above:

Type A: 20 – your contribution to the project + $\frac{3}{4}$ * (sum of all contributions to the project in your group)

Type B: 20 – your contribution to the project + $\frac{1}{2}$ * (sum of all contributions to the project in your group)

The third stage

In the third stage you get to know the contributions to the project by all participants in your group, as well as their period income. Furthermore, you will again see how each participant in your group decided in the first stage.

Then the current period ends and the next period begins with the same participants. Your type and the types of the other participants remain the same. All participants can then again decide in the first stage whether they want to fix contributions in the second stage. Again, the second stage follows and finally the third stage.

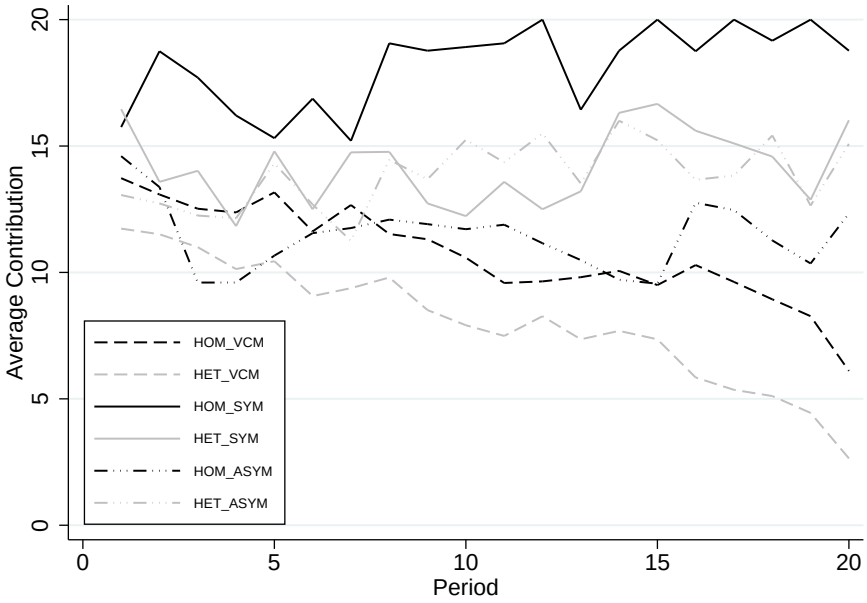
End of the experiment and payment

The experiment ends after 20 periods. Subsequently, we will ask you to answer a few general questions on the computer. Your answers to these questions have no influence on how much money you earn in the experiment. When all participants have filled out the questionnaire, payments will be made. Your total income from the 20 periods will be converted into Euro and paid to you in cash.

Do you have any questions? If so, please raise your hand.

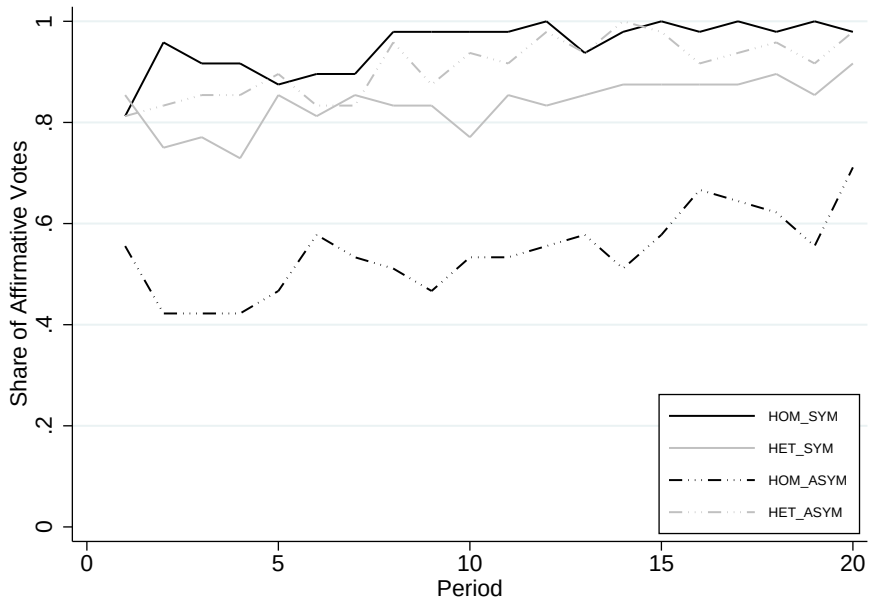
C Figures

Figure 1: Development of Average Contributions over Time



Note: In treatments $HOM-VCM$ and $HET-VCM$, average contributions decrease over time ($HOM-VCM$: Spearman's Rho $r=-0.27$, $p<0.01$ and $HET-VCM$: $r=-0.47$, $p<0.01$). In treatments $HOM-SYM$ and $HET-ASYM$, average contributions increase over time ($HOM-SYM$: $r=+0.27$, $p<0.01$, $HET-ASYM$: $r=+0.21$, $p<0.01$). In treatments $HET-SYM$ and $HOM-ASYM$, time trends in contributions are not significant ($HET-SYM$: $r=+0.08$, $p=0.16$, $HOM-ASYM$: $r=-0.03$, $p=0.64$).

Figure 2: Share of Affirmative Votes over Time



Note: In all four two-stage treatments, the share of affirmative votes increases over time ($HOM-SYM$: Spearman's Rho $r = +0.76$, $HET-SYM$: $r = +0.75$, $HOM-ASYM$: $r = +0.73$, $HET-ASYM$: $r = +0.74$, all $p < 0.001$).

Figure 3: Contributions in Case of No Institution over Time