Essays on Public Good Provision: Fair Contribution Rules and Institution Formation

Jakob Neitzel

Dissertation Thesis submitted for the Degree of

Doctor Rerum Politicarum

Department of Economics, University of Hamburg

June 5, 2014

Thesis Committee

Chairman:	Prof. Dr. Dr. Lydia Mechtenberg
First Examiner:	Prof. Dr. Anke Gerber
Second Examiner:	Prof. Dr. Andreas Lange

The disputation was held on October 9, 2014.

Acknowledgments

I first and foremost thank my supervisor Prof. Dr. Anke Gerber for giving me the opportunity to write this dissertation, as well as her guidance and support throughout the process of its completion. Moreover, I also thank all my colleagues at the Department of Economics, especially Leonie Baumann, Johannes Ebert, Dominik Hattrup, Menusch Khadjavi, Wiebke Probst, Lauri Sääksvuori and Nico Singer for comments, company and good times. I am grateful to Angelika Höllendörfer for every bit of administrative assistance she has given me over the last years.

I am thankful to the Department of Economics for providing the financial means to attend multiple conferences and workshops. I would also like to thank Olaf Bock and the staff of the WiSo research laboratory for their assistance in generating the experimental data.

The research compiled to this dissertation has greatly profited from comments and suggestions by Arno Apffelstaedt, Leonie Baumann, Hans Carlsson, Tim Cason, Anke Gerber, Fabian Herweg, Håkan J. Holm, Mathias Kifmann, Charles Kolstad, Michael Kosfeld, Andreas Lange, Sandra Ludwig, Gerd Mühlheusser, Axel Ockenfels, Arno Riedl, Bettina Rockenbach, Lauri Sääksvuori, Klaus Schmidt, Miriam Schütte, Ferdinand von Siemens, Nico Singer, Achim Voß, Gari Walkowitz, Achim Wambach, Hans-Peter Weikard, Philipp C. Wichardt and a couple of anonymous referees, as well as seminar participants at the University of Zurich, the University of Cologne, the University of Frankfurt, the 19th EAERE conference in Prague, the 2013 Florence Workshop for Behavioral Economics, the International Workshop on Mechanism Design and the Environment in Edinburgh, the 20th EAERE conference in Toulouse, the 2013 meeting of the Verein für Socialpolik in Düsseldorf and the 8th Nordic Conference on Behavioral and Experimental Economics in Stockholm.

My special thanks go to my fiancée, Christin Wolmeyer, to my parents, Britt-Marie Henriksson and Ulrich Neitzel, and to my sister, Annika Neitzel, for their enduring support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

60 years ago, Samuelson (1954) first introduced the problem of private provision of public goods into economic theory. However, even after all these years the question of how to circumvent inefficient provision is still not fully answered. Numerous research has shown that humankind is not as self-centered as economists have assumed for a long time; nevertheless, the conflict of individual and collective interest leads to a social dilemma that offers no easy solution.

In this dissertation, I approach the social dilemma in private provision of public goods from two different directions. In the first part of the dissertation, I investigate the impact of fair contribution rules on public good provision in societies which are heterogeneous in wealth. In the second part, I study how endogenous institution formation can help to overcome the public goods problem in the absence of a governing body setting the rules.

I extract three key findings from the three chapters on the impact of fair contribution rules. First and foremost, in the presence of wealth inequality, people apply normative rules of fair contributions in a self-serving way. As a consequence, the sequence of contribution solicitation affects the total level of public good provision, where the highest level is achieved if the wealthier individuals are asked to contribute first. Second, an increase in wealth inequality does not alter people's perceptions of fairness, but may erode existing revealed contribution norms, resulting in lower public good provision. Third, the erosion of contribution norms can be attributed to the existence of a social norm that serves as a lower bound for the contribution decision.

The second part of the dissertation explores institution formation in the context of international environmental agreements. The key insights here are that stricter minimum participation rules are more efficient in the provision of public goods, as they leave no room for free-riding behavior. Furthermore, I show that the size of stable coalitions in the global emissions game with asymmetric countries primarily depends on the number of countries suffering from high damages from pollution. Coalitions can become large, if only few countries of this type join.

A public good is defined as a commodity for which the use of one consumer does not prevent it to be used by another consumer (nonrival) and no consumer can be excluded from its use (nonexclusive) (Varian, 1992; Mas-Colell et al., 1995, among many others). If every consumer only considers her own benefits from the provision of this good, then she does not account for the fact that the provision yields positive externalities on all other consumers. Consequentially, the private provision of the public good will be inefficiently small. As a result, economists refer to the free-rider problem when individuals relish on the provision of a good while not providing it themselves (Groves and Ledyard, 1977).

Many alternative concepts have been developed in order to get a hold of the free-rider problem and to increase the efficiency of private public good provision. In most of these concepts an institution regulates the society and provides individuals with the incentive to contribute the efficient amount to the public good. The incentive is provided either through centralized sanctioning (Clarke, 1971; Groves, 1973; Falkinger, 1996) or through peer punishment and reward systems (Fehr and Gächter, 2000; Masclet et al., 2003). However, non of these concepts offers a comprehensive solution. Centralized sanctioning only works in an environment where a government has the power to punish deviation from the efficient contribution level (although, there have been approaches to invent sophisticated mechanisms to provide incentives even in the absence of a strong institution (Gerber and Wichardt, 2009, 2013)). Peer punishment has the drawback that some actors need to sacrifice part of their own payoff in order to punish others, thus again creating a public goods problem.

It is by now widely acknowledged that people are not as narrow-minded as they are assumed to be in many models. The use of experimental methods has shown that there is a considerable number of individuals who are willing to sacrifice their own short-term payoff for the collective long-term benefit (Fehr and Gächter, 2000). Actual provision levels observed in public goods experiments are well above the pessimistic theoretical predictions using standard preferences (see Ledyard (1995) and Chaudhuri (2011) for excellent surveys of the literature on public goods experiments). Building on these experimental results, new models of cooperative behavior have been developed using inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), reciprocity (Charness and Rabin, 2002; Falk and Fischbacher, 2006), or the existence of social norms (López-Pérez, 2008) as drivers of human cooperation.

In this dissertation, I focus mainly on conditional cooperation as the superordinate phenomenon of the listed reasons for cooperative behavior. Conditional cooperation can be described as the willingness to contribute an amount to the public good which is significantly higher than the individual optimum, if the other actors also do so. In situations where the individuals differ from each other in certain characteristics, the notions of fairness and distributive justice come into play. Fairness principles have been discussed by philosophers for many thousand years, dating back as far as Aristotle and his *Nicomachean Ethics*. In general, one may distinguish between egalitarian approaches, in which everyone is supposed to be treated equal, regardless of individual characteristics, and proportional approaches, in which differences in individual characteristics should be taken into account (Konow, 2003).

One possible and self-evident source of heterogeneity between agents in the private provision of a public good is inequality in wealth. Fairness considerations, in line with the egalitarian or the proportional approach, lead to the question whether it is fair to cooperate based on equal absolute contributions or on equal relative contributions. Hence, agents have two prominent fair contribution rules to choose from.

Chapter 2, *Normative Conflict and Cooperation in Sequential Social Dilemmas*, represents the starting point for the inquiries into the use of fair contribution rules in sequential public good provision with wealth inequality. I investigate an economic environment where monetary payoff is determined by a linear public goods game, agents differ from each other in initial levels of wealth and voluntary contributions are made sequentially in two stages. Similar frameworks are also used in Chapters 3 and 4.

The goal of Chapter 2 is to offer guidance for organizations collecting contributions. I investigate how the sequence of contributions and the use of different fair contribution rules affect the total level of contributions. I use experimental and survey data to show that, within this framework, people hold distinct normative views of fair contributions. The experimental evidence supports the predictions that individuals employ fair contribution rules in a self-serving fashion. Rich players predominantly cooperate on the basis of equal absolute contributions, while poor players prefer cooperation on the basis of equal relative contributions. As a result, both the proposed model and the experimental evidence suggest that voluntary public good provision is maximized if all rich individuals contribute first. Furthermore, I find proof that this sequential structure also decreases income inequality more than any other sequence. Yet, it does not diminish the income of rich players but only increases the income of poor individuals. These findings suggest that changing the succession of decisions can result into Pareto-improvements in the final outcome of the game.

In the light of growing wealth disparities in almost all countries in the world, the subsequent Chapter 3, *Wealth Inequality, Perceptions of Fair Behavior and Cooperation*, deals with the impact of increasing endowment asymmetry in the experimental set-up used in Chapter 2. Herein, I aim to understand how increasing wealth disparities affect our perceptions of fair behavior and how this translates into revealed contribution norms. The questionnaire data suggest that perceptions of fairness do not change with increasing wealth inequality, and that contributions proportional to endowment are considered fair by the absolute majority of people. I find notable behavior in the experimental data: poor individuals switch their revealed contribution norm from equal relative contributions to equal absolute contributions. Thereby, poor players depart both from the behavior considered to be most appropriate by the external observers from the questionnaire, as well as from the self-serving interpretation of fair contribution rules. Moreover, coordination on one contribution norm seems to be more difficult under larger wealth disparity and leads to lower overall provision of the public good.

The results of Chapter 3 provide the motivation for Chapter 4, *Wealth Inequality and Fair Contribution Rules in Sequential Public Good Provision*. Chapter 4 aims at providing an explanation for the change in revealed contribution norms under increasing wealth inequality. Based on the intuition that people face unobservable limits in their cooperation choice, I build a model which incorporates a social norm serving as an absolute lower bound on contributions. Using this model, I show that the switch in the revealed contribution norm by the poor players can be attributed to the interplay of the self-serving use of fair contribution rules and the social norm. Moreover, the model can correctly predict the behavior observed in the lab, as well as offer some additional insight on the effects of increasing wealth inequality. I find that increasing wealth inequality indeed lowers the private provision of public goods.

One could interpret the results such that increasing disparities in wealth shift the strategic advantage from the second-mover to the rich individuals, independent of the sequence of play. While it is difficult to derive policy implications from the small-scale model of Chapter 4 and the results from Chapter 3, in the light of social unrest around the world, it requires only little imagination to translocate the discrepancy between what people consider fair and what is essentially implemented to real life problems. These results are at least thought-provoking.

Having analyzed general behavioral aspects of cooperation between individuals in the first part of my dissertation, I turn to a more specific problem in the second part consisting of Chapters 5 and 6. While many public goods problems can be solved by a government dictating an institution that remedies the incentives for free-riding, this is not possible on the international scale. Environmental protection and the reduction of greenhouse gas emissions are global public goods for which the punishment power of existing global institutions is too weak to force countries to participate in provision.

Therefore, other endogenous measures have to be considered.

One prominent measure is the signing of international environmental agreements to curb greenhouse gas emissions and its most famed example: the Kyoto protocol. However, these agreements face another problem: they need to be self-enforcing, as early noted by Barrett (1994). The large theoretical literature on self-enforcing environmental agreements yields the pessimistic result that only very small coalitions fulfill this property (Carraro and Siniscalco, 1993; Finus and Rundshagen, 2003). I contribute to this literature in two ways. First, I experimentally investigate how the introduction of a minimum participation requirement for the endogenous formation of an institution can increase the level of public good provision. Second, I extend the literature on stable coalitions by allowing for heterogeneity between countries.

Chapter 5 marks the beginning of the investigation of institution formation for the provision of public goods. *Minimum Participation Rules for the Provision of Public Goods* investigates how the implementation and the overall efficiency of an institution is affected by the requirement of a minimal number of signees. In particular, I address the question whether there is a tradeoff between the effectiveness of an institution and the probability of signing. One could think that less strict participation requirements facilitate the formation of an institution, while stricter rules generate higher contributions once the institution is established. The game theoretic analysis cannot answer this question, as both types of minimum participation requirements also yield equilibria in which no institution is formed. Using experimental data, I show that stricter minimum participation rules increase average provision levels because they remedy the free-riding incentive on both the first level (contributions to public good) and the second level (institution formation) of the social dilemma. That is, all rules yield the same number of institutions, but the stricter rules produce higher contributions, because non-members of the institution free-ride on the provision of members in the less strict case. A careful interpretation of the results suggests that international environmental agreements should aim at being ratified by all countries before entering into force.

The dissertation ends with Chapter 6, *The Stability of Coalitions when Countries are Heterogeneous.* Therein, I depart from the linear public goods game used in all previous chapters. I extend the non-cooperative approach of self-enforcing international environmental agreements originated by Barrett (1994) to a model which accounts for four types of countries. Countries can either generate high or low benefits from own emissions and at the same time either suffer a lot or little from global emissions. The analysis of the model yields that large coalitions can be stable if only one or two countries which suffer a lot from emissions enter the coalition. The larger the disparity between countries with high costs and low costs from emissions, the larger is the maximally stable coalition. The different levels of benefit only have an effect on the stable coalition size if two countries with high costs are members. Hence, I can not deliver much more optimistic results than other previous studies on the topic (cf. Barrett, 1994; Fuentes-Albero and Rubio, 2010; Pavlova and de Zeeuw, 2013), but rather provide evidence that less sophisticated models suffice for the analysis of stable coalitions.

With this dissertation, I make significant contributions to the economic literature on public good provision. I show how the self-serving use of fair contribution rules affects the efficiency of alternative sequential contribution mechanisms and how increasing wealth inequality can lead to the erosion of existing contribution norms. I propose a model that explains the subjects' behavior observed in the lab. Furthermore, I provide evidence that a stricter minimum participation rule yields more efficient outcomes for the endogenous formation of institutions. Last, with the to date most generalized model of heterogenous countries in the global emissions game, I demonstrate that self-enforcing international environmental agreements are very unlikely to comprise all countries in the world.

Chapter 2

Normative Conflict and Cooperation in Sequential Social Dilemmas[1]

Abstract We investigate how conflicting normative views of fair contribution rules can be used to design sequential contribution mechanisms to foster cooperation. We show, using survey and experimental data, that individuals hold well-defined yet widely diverse normative views of fair contribution rules. We use the information about the conflicting normative views of fair contributions to model cooperative behavior in sequential collective action problems in the presence of wealth inequality. Our model predicts that a sequential mechanism which solicits contributions first from wealthy actors generates greater public good provision and narrows wealth inequality more than any alternative sequential mechanism. Our experimental data show that the mechanism with wealthy first-movers generates greater contributions and narrows wealth inequality more than the alternative mechanisms, as predicted. Our results suggest how altering the sequential order of contributions in heterogeneous populations may affect public good provision and help organizations to increase the total value of solicited contributions.

Keywords Sequential Public Good Provision, Equity, Heterogeneous Wealth, Normative Conflict, Norms
JEL Classification C92, D63, H41

[1] This chapter is co-authored by Lauri Sääksvuori.

2.1 Introduction

A conflict between individual and collective interest is a pervasive feature of human social organization. This conflict often occurs in the presence of resource and wealth inequality. Team members in schools and work places have different amounts of time, talent and skills available for their joint undertakings. Farmers cultivating soil and raising livestock regularly face collective environmental hazards and severe weather phenomena, while their accumulated wealth and opportunities to reckon with prevailing crises often substantially differ. At international level, countries differ in their capabilities and incentives to contribute to greenhouse gas reduction and engage in abatement activities designed to curb the effects of global climate change. Likewise, the presence of wealth inequality is apparent in many collective decisions of every-day life varying from jointly purchased gifts to contributions to charitable organizations.

When individuals and organizations derive equal benefits from their joint undertakings, but differ in their personal characteristics and resources to contribute, determining the appropriate level of contributions may become difficult and give rise to conflicting normative views of fair contribution rules. The existence of multiple plausible rules for fair contributions may in turn lead to tensioned relationships and destructive normative conflicts between the decision makers. In contrast to the prevailing view, which has emphasized the destructive nature of normative conflict, this paper investigates how conflicting normative views of fair contribution rules can be used to design voluntary contribution mechanisms to foster cooperation and increase the amount of solicited contributions in heterogeneous populations.

The existing economic literature on public good provision predominantly assumes that contributions are solicited simultaneously. Yet, in practice, the nature of multiple public goods and the characteristics of numerous fundraising campaigns suggest that contributions are often made in a sequential order. For instance, the European Union has regularly made an advance commitment to increase its emission reduction before the next round of post-Kyoto negotiations on greenhouse gas emissions, if other major emitting countries accept to contribute a reasonable share to the emission targets. Observations from fundraising campaigns show that the campaign organizers regularly announce prior contributions to the cause. Likewise, persons solicited for contributions often ask themselves about the previously donated amounts.

This paper shows, using survey and experimental data, that uninvolved and involved individuals hold well-defined yet widely diverse normative views of fair contribution rules in the presence of wealth inequality. We use the information about the conflicting normative views of fairness to model cooperative behavior in sequential collective action problems. We embed the normative conflict into a utility framework by assuming that individuals derive utility from behaving in concord with normatively appealing contribution rules. Using this framework, we pay particular attention to the fact that normative views of fair contribution rules can systematically vary across different individuals. This enables us to generate precise and testable propositions about cooperative behavior in alternative sequential contribution mechanisms. More precisely, our model predicts that a sequential mechanism which solicits contributions first from wealthy agents generates greater public good provision and narrows wealth inequality more than any alternative sequential mechanism.

We implement a laboratory experiment which renders it possible to test our predictions in a controlled environment. In the experiment, we enable conflicting normative views of fair contribution rules by introducing a real-effort tournament and rewarding the best performers within each group with greater initial endowments from which individuals can contribute to the public good. The results from the experiment provide partial support for the hypothesis that changes in the sequential order of contributions may affect public good provision in the presence of wealth heterogeneity. We observe that the highest contributions are provided in the treatment where contributions are solicited first from wealthy actors. We find that the mechanism with wealthy first-movers narrows wealth inequality more than the alternative mechanisms, as predicted.

By combining incentivized experimental data with non-incentivized survey data we show the importance of monetary incentives for the emergence of normative conflict. In other words, we show that conflicting normative views of fair contribution rules are more likely to emerge when individuals have their own money at stake. This has broader implications for studies that aim to elicit social norms and normative views of fairness to generate precise behavioral propositions a priori. In particular, the difference between the opinions of impartial external observers and involved decision makers concerning fair contribution rules suggest that theoretical models and empirical investigations based on survey data may underestimate the prevalence and consequences of conflicting normative views of fair contribution rules.[2]

[2] A study by Krupka and Weber (2013) develops an incentivized method to identify social norms in an uninvolved set of individuals with coordination games. Their study shows that elicited social norms are often consistent with the behavior of a new sample of participants. The proposed mechanism appears to be a very promising method to circumvent many problems associated to non-incentivized questionnaire studies of social behavior. However, it is important to note that social norms may differ from normatively fair contribution rules which describe how one ought to behave in a particular situation. It is also not clear as yet how well-suited the method is for identifying conflicting norms within a single set of individuals.

2.2 Existing Literature

Economic science has responded to the pervasive challenge of free-riding in collective action problems by suggesting numerous sophisticated mechanisms to enhance cooperation and generate efficient allocations to public good (Clarke, 1971; Groves, 1973; Groves and Ledyard, 1977; Walker, 1981; Moore and Repullo, 1988; Falkinger, 1996). A common characteristic of these mechanisms is that they assume narrowly self-interested economic agents and require fairly strong institutional arrangements to implement centralized sanctions or transfer payments to generate socially efficient contributions. However, these theoretically optimal schemes are rarely implemented in reality and offer only limited opportunities to develop new practical solutions to foster cooperation in naturally-occurring situations. In practice, several different types of public goods have to be provided in the absence of strong institutions and coercive power to punish violators.

The economics literature on public good provision in sequential move games complements the literature on theoretically optimal provision mechanisms. An early theoretical model for voluntary contributions to public goods was provided by Bergstrom et al. (1986) who show that the fund-raiser prefers not to announce past contributions to future donors. Bagnoli and Lipman (1989) study sequential contribution mechanisms which condition public good provision on private contributions and show that such mechanisms implement the social optimum if the sum of private contributions is sufficiently high. Varian (1994a) describes a two-stage mechanism which implements efficient allocations as a subgame-perfect equilibrium. In a similar vein, Varian (1994b) studies public good provision in a sequential game with quasi-linear utility functions and shows that simultaneous contributions generate greater public goods provision than sequential contributions if the sequential order enables the first-mover to free-ride on the subsequent contributions of others. However, Andreoni (1998) shows that announcing binding commitments to contribute may increase the total value of donations if this enables donors to coordinate on positive provision levels. Bracha et al. (2011) provide experimental evidence in support of the proposition that early contributions may increase the total value of donations if the fixed costs of providing public goods are reasonably high. Vesterlund (2003) studies the impact of announcing contributions in environments where there is uncertainty about the quality of the soliciting organization and shows that the initial contributions can serve as credible signals about the quality of the organization. Potters et al. (2005) show that sequential move structures may result in larger public good provision if the followers mimic a responsible leader. They also report that, when given an opportunity to choose between sequential and simultaneous move structures, donors predominantly choose to contribute sequentially. List and Lucking-Reiley (2002) conduct a field experiment to test the impact of various amounts of seed money and report that increasing the amount of announced contributions resulted in a manifold increase in contributions by the general public.

Besides the large theoretical interest and experimental literature directly testing the empirical relevance of proposed models of public good provision, a substantial experimental effort has been devoted to understand the empirical nature of free-riding. There are a growing number of stylized facts describing human behavior in public good games. The individually optimal null provision hypothesis is typically rejected and groups attain better outcomes than foreseen by theories based on narrowly self-interested motivations (Ledyard, 1995). Positive contributions to public good are frequent even after dozens of repetitions within the same group, but typically decline in time towards the equilibrium (Gächter et al., 2008). Likewise, the importance of social norms and decentralized norm enforcement for sustaining positive contributions has been demonstrated in numerous economic experiments where individuals derive equal benefits from the public good (Fehr and Gächter, 2000; Rege and Telle, 2004; Chaudhuri et al., 2006).

It is likely that observed behavior patterns from homogeneous groups, where equal contributions create a natural and intuitively appealing norm, do not readily generalize to environments characterized by resource heterogeneity. Hence, researchers have lately begun to investigate the emergence and enforcement of contribution norms in heterogeneous groups. Data from surveys and experimental investigations show that people in heterogeneous populations have well-defined but differing normative views of fair contribution rules (Reuben and Riedl, 2013). Consequently, in such situations, determining the most appealing and appropriate level of contributions becomes more difficult and may give rise to unpredictable and destructive normative conflicts about how one ought to agree upon contribution levels under resource heterogeneity.

The emergence of normative rules and fairness principles is gaining growing attention as an important phenomenon in economic decision-making.[3] There are three major observations. First, there exists a plurality of normatively appealing rules which may lead to conflicting normative views of fair contributions (Cappelen et al., 2007; Nikiforakis

[3] Literature often points out that normative contribution rules differ from social norms. According to a definition by Bicchieri (2006), as cited in Nikiforakis et al. (2012), unlike normative rules the establishment of a social norm requires a sufficient number of individuals who know what the rule is and conform to this rule, provided that they believe that (a.) a sufficient number of others conforms to the rule, and (b.) a sufficient number of others expects the individual to conform. In many occasions where a social norm emerges, these requirements are supplemented by willingness to punish those who do not conform. According to this definition, a normative contribution rule, which suggests how one ought to behave in a given situation, differs from a social norm, unless the aforementioned conditions are satisfied.

et al., 2012; Winter et al., 2012). Second, normative views of fair contribution rules often relate to the principles of *equity* and *equality* (Konow, 2003; Reuben and Riedl, 2013). Third, people exhibit both unconscious (Babcock et al., 1996) and calculated distortions (Dahl and Ransom, 1999; Konow, 2001) in fairness principles which often leads to self-serving use of equity norms. Consequently, Lange et al. (2010) show that the self-interested use of equity arguments can lead to distortions with large economic costs when negotiating about important economic and social issues such as international environmental agreements. Our paper contributes this line of research by showing how conflicting equity principles can be used to design voluntary contribution mechanisms to foster cooperation and increase the amount of solicited contributions in heterogeneous populations.

Finally, our experiment uses an analogous real-effort procedure to determine wealth differences between participants as experiments that have studied the effects of social status on market behavior (Ball et al., 2001) and charitable giving (Kumru and Vesterlund, 2010). In particular, Kumru and Vesterlund (2010) show that low-status followers are likely to mimic donations by high-status leaders, leading to larger overall contributions when individuals of high status contribute before those of low status. In contrast to studies examining the effects of social status on economic behavior, our experiment does not directly induce differences in interpersonal status. Moreover, our results show that differences in social status are unlikely to explain the dynamics of normative conflict and variation between the sequential contribution mechanisms in our experiment. First, we find that the lower endowed participants do not mimic the behavior of high endowed participants, but both types of participants have a self-serving interpretation of desirable contribution rules. Second, in an accompanying paper we study the role of increasing wealth inequality in sequential contribution games and find that the impact of increasing wealth heterogeneity on sequential public good provision goes against the status-based explanation in our data set (Neitzel and Sääksvuori, 2014b).

2.3 Model and Propositions

We study the private provision of a public good in a population of $n \geq 2$ players where players differ in their capacity to contribute due to wealth heterogeneity. There are two different types of players: rich players with an endowment w_R and poor players with an endowment w_P, where $w_R > w_P$. We model public good provision such that after the contribution decision player i receives income determined by

$$\pi_i(c_i, c_{-i}) = w_i - c_i + \alpha(c_i + c_{-i}), \tag{PG}$$

where $\frac{1}{n} < \alpha < 1$ is the return on investment, $w_i \in \{w_R, w_P\}$ is player i's endowment, c_i is his or her own contribution to the public good and $c_{-i} = \sum_{k \neq i} c_k$ are the contributions of all players except i.

We assume that players value both monetary income and adherence to a fair contribution rule when deciding about the public good provision. We posit that wealth heterogeneity between actors raises divergent views about fair contributions. In particular, we are interested in analyzing the co-existence of two prominent fairness rules that relate contributions to initial wealth. *Equality* suggests the equalization of *absolute* contributions with no necessary link to individual characteristics such as capacity to contribute. By contrast, *equity* links contributions to individual characteristics such as wealth in a proportional manner and stipulates the equalization of *relative* contributions to public projects.

We assume that individuals apply alternative contribution norms in a self-serving manner. In other words, we assume that individuals choose a normative principle and a contribution level which yields the highest utility for the given player type and observed contributions of the other players. More formally, we assume that player i maximizes the following utility function when contributing to the public good:

$$\begin{aligned} \max_{c_i, k} \; & u_i(c_i, c_{-i}, k) = \pi_i(c_i, c_{-i}) - \frac{\beta}{2}\left(c_i - m(k, c_{-i})\right)^2 \\ s.t. \; & c_i \in [0, w_i] \\ & k \in \{0, 1\} \end{aligned}$$

where the parameter $\beta \geq 0$ determines the weight that is attached to deviations from the applied norm $m(k, c_{-i})$. The norm depends on the applied contribution norm k and the contributions of the other players. It is defined by

$$m(k, c_{-i}) = \left(\frac{1}{n-1}\right)^{1-k} \left(\frac{w_i}{w_{-i}}\right)^{k} c_{-i}$$

where $w_{-i} = \sum_{k \neq i} w_k$ denotes the sum of other players' initial endowments.[4] Hence, if $k = 0$, player i is said to apply the equality norm, since $m(0, c_{-i}) = \frac{1}{n-1} c_{-i}$. In this case,

[4] The use of a quadratic loss function is a common assumption in economics and social science literature. This particular utility function has been used in a similar fashion by Cappelen et al. (2007).

the player only uses the absolute contributions of other players as a reference point with no link to initial endowments. By contrast, when $k = 1$, the player applies the equity norm and accounts for the initial endowments ($m(1, c_{-i}) = \frac{w_i}{w_{-i}} c_{-i}$).

Our primary interest pertains to the impact of conflicting normative views of fair contribution rules in simultaneous and sequential move contribution games. In a simultaneous contribution game, all players act simultaneously and are regarded as first-movers. In a sequential contribution game (SEQ), we separate players into first- and second-movers.[5] In conjunction with wealth heterogeneity, this leads to three different possible combinations of first- and second-movers: (a.) all rich players contribute first, (b.) all poor players contribute first, or (c.) rich and poor players contribute in a mixed order such that there is at least one first- and second-mover of each type.

We first study the selection of the preferred normative principles of second-movers and state that

Proposition 2.1 *Rich players prefer the equality norm, whereas poor players prefer the equity norm.*

The intuition behind the proposition is straightforward. For rich players, adherence to the equality norm requires lower contributions than adherence to the equity norm ($\frac{c_{-i}}{n-1} \leq \frac{c_{-i}}{w_{-i}} w_i$, since $w_i > \frac{w_{-i}}{n-1}$). For poor players, the opposite applies and the equity norm stipulates lower contributions than the equality principle ($\frac{c_{-i}}{w_{-i}} w_i \leq \frac{c_{-i}}{n-1}$, since $w_i < \frac{w_{-i}}{n-1}$). A complete proof of the proposition is provided in the Appendix 2.A.

Thus, in the following analysis of alternative sequential mechanisms, we assume that all second-movers contribute according to their preferred normative principle based on the observed contributions of first-movers. Rich second-movers use the average absolute amount contributed by first-movers as their normative principle, while poor second-movers use the average amount contributed by first-movers relative to the initial endowments of first-movers as their normative principle. Formally, $c_i = \frac{c_{fm}}{\#fm}$, if player i is a second-mover and has an endowment w_R, and $c_i = \frac{c_{fm}}{w_{fm}} w_i$, if player i is a second-mover and has endowment w_P. c_{fm} denotes the sum of first-mover contributions, $\#fm$ the number of first-movers and w_{fm} is the sum of endowments of first-movers.[6]

[5] An obvious alternative for this structure would be to let all players move sequentially. However, such a game puts the main emphasis on individual first-movers as leaders, which is not of our interest here. Using the two-stage structure and groups of simultaneously moving players at each stage has the advantage of focussing on the different types of actors rather than on individuals.

[6] The exact solution of the utility maximization problem requires players to undercut the prescribed contributions by the fraction $\frac{1-\alpha}{\beta}$. However, to simplify, in the following we assume that β is large enough such that deviations from the preferred normative principle are negligible.

Obviously, the same reasoning is not applicable for first-mover contribution behavior, as contribution decisions can not be based on any observed contributions by other players. Therefore, we hypothesize that first-movers employ a heuristic to determine their contribution level. As a result, we assume that all first-movers contribute an equal proportion, $0 \leq x \leq 1$, of their endowment to the public good.[7] In other words, we assume that the rich and poor first-movers contribute the same relative amount to the public good.

Using these assumptions for first- and second-mover contribution behavior we derive the following proposition concerning public good provision across the distinct sequential contribution mechanisms:

Proposition 2.2 *SEQ with rich first-movers generates greater public good provision than alternative sequential move mechanisms.*

The proof of the proposition is driven by the fact that the equality norm dictates rich second-movers to contribute xw_P, while rich first-mover contributions equal xw_R. Thus, since $w_R > w_P$, rich players obviously contribute less as second-movers than as first-movers. In contrast, the contributions made by poor players do not differ when acting either as a first- or second-mover and amount to xw_P. Therefore, the more rich first-movers there are, the higher the public good provision. A complete proof of Proposition 2.2 can be found in Appendix 2.A. This result also triggers the next proposition regarding income inequality between rich and poor:

Proposition 2.3 *SEQ with rich first-movers diminishes wealth inequality between rich and poor players more than alternative sequential move mechanisms.*

The intuition behind the proposition is self-evident. Wealth inequality diminishes whenever rich individuals contribute larger absolute amounts than poor individuals. By definition, this only occurs in sequential mechanisms with rich first-movers. Thus, the mechanism with most rich first-movers induces the lowest inequality. The proof is omitted.

[7] Abundant experimental research on voluntary public good provision has shown that the average relative contributions are fairly stable across various experimental designs at the beginning of the interaction. If players interact repeatedly over a finite number of rounds, their contributions often start out from a range between 40% and 60% of the social optimum (Chaudhuri, 2011). The empirical evidence provided in this paper is in concordance with the previous observations.

2.4 Experiment and Questionnaire

2.4.1 Experimental design and procedure

We implement a laboratory experiment which renders it possible to study public good provision across alternative sequential contribution mechanisms in a controlled environment. The experiment consists of two parts. The first part involves a real-effort task. This task is used to determine which individuals have high initial endowments and which individuals have low initial endowments when contributing to the public good in the second part of the experiment. We use the Encryption task (Erkal et al., 2011) which proceeds as follows. Participants are divided into groups of four and given an encryption table which assigns a number to each letter of the alphabet. During the next ten minutes, participants are presented words which need to be encrypted by substituting the letter with numbers using the encryption table. The words are presented in a predetermined sequence and are the same for all participants. A subject cannot proceed to the next word until the word has been correctly encrypted. After the real-effort task, the two group members with the highest number of encrypted words receive a high initial endowment, w_R, while the two other members receive a low initial endowment, w_P. Ties for the second place are broken at random.

We employ a real-effort task with tournament incentives to facilitate the emergence of conflicting normative views of fair contributions. We expect that earned endowments decrease the normative appeal of the equal earnings principle and increase the salience of equal contributions as an alternative normative principle under wealth heterogeneity. Participants stay in the same group of four throughout the experiment and know in the second part of the experiment if they were among the two most productive individuals in the group in the first part. However, no information about the exact number of encrypted words is revealed to the participants after the first part. After completing the first part, participants receive new instructions concerning the second part.[8] The second part involves four different treatments.[9] Common to all treatments is the linear public goods game (PG) which is played in groups of four for 15 consecutive periods. In our experiment, we set α to be 0.4.

We investigate public goods provision under alternative sequential contribution mechanisms. In sequential move games, we have two first-movers who contribute simultane-

[8] An English translation of the originally German instructions are attached in Appendix 2.C.

[9] This paper is accompanied by a related paper where we study how increasing wealth heterogeneity may lead to the erosion of revealed contribution norms and hamper the efficiency of sequential contribution mechanisms (Neitzel and Sääksvuori, 2014b). The accompanying paper uses in part the same data as presented in this paper.

ously and two second-movers who contribute simultaneously after observing the contributions made by the first-movers. This gives rise to three possible combinations of first movers: two rich players (RR), two poor players (PP) and one rich and poor player each (RP). In addition, we conduct a treatment (SIM) with simultaneous move structure. In each treatment, rich players are assigned an endowment of 25 points, while poor players are assigned an endowment of 15 points.

In addition to the contribution decisions, we elicit participants' beliefs about the behavior of other group members in the current period. Participants are requested to predict the contribution of the same participant type as they are and the average contribution of two other types of participants. Participants are paid a small reward based on the accuracy of their estimates. In all treatments, participants are informed at the end of each period about the total contribution in their group, the contribution of each individual and whether the individual making a specific contribution is low or high endowed participant. In order to track individual behavior during the interaction, every participant is assigned a unique identification letter A, B, C, D and the individual contributions are listed in the same order such that for the two first movers are assigned letters A and B, while for the two second-movers we use letters C and D.

The experiment was conducted at the experimental laboratory of the School of Business, Economics and Social Sciences at the University of Hamburg. We used z-tree (Fischbacher, 2007) for programming and ORSEE (Greiner, 2004) for recruiting. A total of 192 subjects participated in the experiment in 8 different sessions. The vast majority of 108 female (Age Mean: 24.0, Std: 2.55, Min: 20, Max: 34) and 84 male (Age Mean: 25.2, Std: 3.89, Min: 19, Max: 39) subjects were undergraduate students representing a wide range of different disciplines. Upon arriving at the laboratory, participants received written instructions and were randomly assigned to their cubicles preventing communication and visual interaction. Instructions were read publicly by a member of the research team. Subjects then took a post instruction quiz and were not allowed to continue until all answers were correct.

At the end of the experiment, one of the 15 rounds was chosen at random to determine the earnings from the public goods game. Likewise, one of the 15 rounds was chosen at random to determine the earnings from the prediction task. The random draws were performed publicly by two different participants who draw a card from a deck of cards numbered from 1 to 15. The sessions lasted approximately 75 minutes including instructions, post instruction quiz, demographic questionnaire and payment procedure. Earnings per participant were on average 11.22 €.

2.4.2 Questionnaire Study

In addition to the data from the experiment, we elicit normatively appealing rules of behavior in the public goods game among impartial external observers. We conduct this online questionnaire study among uninvolved university students to compare the incidence of alternative normative views of fair contribution rules among involved and uninvolved individuals. First, this enables us to investigate how normative views of fair contribution norms change when people have their own money at stake. Second, by collecting data from involved experimental participants and impartial external observers, we are able to study if there are systematical differences in the behavior of rich and poor individuals when contrasting the choice data with the normative views of external observes.

Reuben and Riedl (2013) conduct a related questionnaire study concerning the normatively appealing rules of behavior in simultaneous public goods games among uninvolved individuals. Nikiforakis et al. (2012) elicit normatively appealing rules of behavior among their experimental participants after conducting the experiment. Our questionnaire study follows in outline the procedure of Reuben and Riedl (2013). The contribution of our questionnaire study is twofold. First, our data complements previous studies by ascertaining how contextual differences between our study and other studies translate into differences in normative considerations.[10] Second, obtaining direct evidence about the normative views of fair contribution rules among uninvolved and involved individuals enables us to study the applicability of our theoretical model in various possible decision environments. In this respect, our empirical results lead us to conjecture that the model developed in this paper provides more accurate predictions about the performance of alternative sequential contribution mechanisms in situations where participants have their own money at stake.[11]

We conducted an online questionnaire among uninvolved individuals using the same

[10] Preceding laboratory experiments have shown that minor contextual features of a choice environment may lead to substantially different choices and outcomes. A recent study by Krupka and Weber (2013) develops an incentivized method to identify social norms in experimental studies and demonstrates that contextual differences in behavior are often consistent with varying social norms.

[11] Reuben and Riedl (2013) discuss various advantages and disadvantages of questionnaire studies over incentivized elicitation of normative views. In the context of this study, it is important to note that the combination of a questionnaire study and incentivized experiment serve multiple purposes that would have been difficult to achieve through the incentivized elicitation of normative second-order expectations. We want to test if the *revealed behavioral norms* of rich and poor individuals systematically differ from the *normative considerations of impartial observers*. Our results strongly suggest that there are systematical differences between the normative views of external observers and revealed behavioral norms. Hence, the comparison of these two types of data enables us to highlight the differences that the data collection procedure can cause when motivating theoretical models of normative behavior with empirical observations.

pool of subjects than in our experimental study. Participants of the questionnaire study began by reading the experimental instructions used in our experiment and were told that they would be answering normative questions concerning fair behavior in various decision situations described in the experimental instructions. The experimental instructions given to the impartial external observers included a section describing the real-effort task used in the experiment. Thus, the external observers were aware that the starting endowments in the experiment were generated through participation in a real-effort tournament. To ensure comprehensive understanding of the decisions situation, participants in the questionnaire study had to answer a series of control questions before submitting their opinions about fair behavior.

Each participant of the questionnaire study answered the same set of questions.[12] We varied the sequence of questions between the participants to control for possible order effects. The questionnaire contained two different types of questions. First, we elicited an unconditional contribution norm by requesting the respondents to answer the following question: "From the viewpoint of a neutral external observer, what is in your opinion a fair contribution to the group account by a group member who has an endowment of 25 (15) EMUs?". Second, we elicited a conditional contribution norm by requesting the respondents to answer the following question: "From the viewpoint of neutral external observer, what is in your opinion a fair contribution to the group account by a group member who has an endowment of 25 (15) EMUs, if a group member with 15 (25) EMUs has contributed 9 EMUs to the group account?" The fixed contribution was set to be 9 in all questions eliciting the conditional contribution rule to disentangle alternative fairness principles from each other in the simplest manner.

For the analysis of the questionnaire data we define two other contribution rules in addition to the equality and equity rules. We refer to an efficiency rule when an external observer finds the efficiency maximizing contribution to be the normatively most appealing contribution rule. Since the collective welfare increases in concordance with the contributions to the group account, the efficiency rule prescribes maximum contributions independent of individual wealth or the contributions of other participants. In addition to previously discussed normatively appealing rules, we define a rule that prescribes equal payoffs between the rich and poor subjects. The moral appeal of this rule is to negate the differences in earnings due to unequal endowments. In other words, the

[12] The full questionnaire included questions that are not analyzed in this paper. These questions are similar to the questions reported here but vary the degree of inequality between the rich and poor participants. The questionnaire also included a section eliciting normative views of fair contribution rules in case the group accounts of two different groups are directly compared with each other. Analysis of the questionnaire data that is not included here can be found in Neitzel and Sääksvuori (2014b). See Appendix 2.C for the full questionnaire.

Table 2.1: Contributions corresponding to normative principles

Other type contributes 9	Equity	Equality	Efficiency	Equal Payoff
Rich should contribute	15	9	25	19
Poor should contribute	5.4	9	15	0

Notes: Observe that the strict normative principle of equal payoffs would require poor players to contribute -1. However, negative contributions are not possible in our questionnaire and experimental studies.

equality of payoffs rule suggest that individuals are ought to be treated equal independent of different performances in the real-effort task. Concretely, the equality of payoffs rule prescribes that rich individuals should contribute a greater relative share of their endowment than the poor individuals.

It is clear that the defined rules of normatively appealing behavior do not converge to a single focal contribution in heterogeneous groups. Consequently, there are a variety of normatively appealing rules of behavior in our public good game which makes it interesting to see if the impartial external observers prefer certain rules over other possible rules. Unfortunately, there is no integer contribution that leads to integer contributions for all defined contribution rules. Thus, we have chosen to set the fixed contribution for conditional contributions to be 9 to separate alternative normative principles from each other and minimize the number of non-integer contributions that match the defined rules of normatively appealing behavior. The general introduction to the questionnaire study and the precise wording of each question can be found in Appendix 2.C.

The questionnaire study was conducted as an online survey among voluntary participants using the same pool of subjects as in our experimental study. A total of 100 subjects participated in the questionnaire study. Most of the 40 female (Age Mean: 24.5 Std: 3.1, Min: 20, Max: 32) and 60 male (Age Mean: 25.8, Std: 5.65, Min: 17, Max: 55) subjects were undergraduate students representing a wide range of different disciplines. To increase the response rate, we randomly selected two participants who received 100 € for completing the questionnaire. Participants of the questionnaire study were able to submit their answers in their own pace independent of the other participants.

2.5 Results

In this section, we first characterize the normatively appealing rules of behavior among the impartial external observers. We interpret these observations as normative fairness principles that fairness minded individuals ought to follow in the experiment. We demonstrate the emergence of well defined, but conflicting normative views of fair contributions. Second, we contrast these normative fairness principles with the actual observations from the experimental data and examine how the assumptions of our theoretical model correspond with observed data. Finally, we test the empirical validity of our theory by investigating the efficiency of varying sequential contribution mechanisms for the financing of public good.

2.5.1 Normative rules of behavior in questionnaire and experimental data

Table 2.2: Fraction of answers corresponding to normative fairness principles, n=100

	Equity	Equality	Efficiency	Equal Payoff	Other
Unconditional	.24	.03	.12	.15	.46
Conditional Rich	.36	.02	.05	.10	.47
Conditional Poor	.45	.01	.04	.05	.45

Notes: For Conditional Poor, Equity comprises all answers of 5 as the rounded value of the exact answer 5.4. Equal Payoff comprises all answers of zero as the nearest possible value of -1.

Table 2.2 sets the stage for our analysis by presenting the fraction of answers that coincides with the defined normative fairness principles in unconditional and conditional decision situations of the questionnaire study.[13] Herein it is important to note that the efficiency norm coincides with the equity and equal payoff norms in unconditional decision situations, whereas the conditional decision situation enables to clearly differentiate between all defined normative principles. Table 2.2 shows that the majority of answers regarding the normatively most appealing behavior coincide with the defined normative principles. Moreover, we observe that the majority of external observers find the equity norm with equal contributions proportional to endowments to be the most appealing behavioral strategy. We test whether the impartial external observers suggest that poor and rich individuals ought to follow different normative principles. We find that the

[13] Here, we use strict categorization rules to quantify the fraction of answers corresponding to normative fairness principles. In Appendix 2.B we present an alternative categorization which allows for small decision errors. The qualitative nature of the results presented here is robust to decision errors. However, if we allow a decision error of ± 1, the share of unclassified contribution rules decreases substantially and we are able to classify 90 percent of the individuals using the four alternative normative fairness principles.

fractions of answers corresponding to normative fairness principles do not differ depending on the initial wealth of individuals (Fisher's exact test based on the categorization of conditional fairness norms between the rich and poor individuals as presented in table 2, p=0.557). Taken together, our results show that the impartial observers find that the normatively most appealing rules of conduct are independent of individual wealth.

Result 2.1 *The most appealing contribution rule coincides with the normative principle related to equity. Impartial external observers find the most appealing rules of conduct to be independent of individual wealth.*

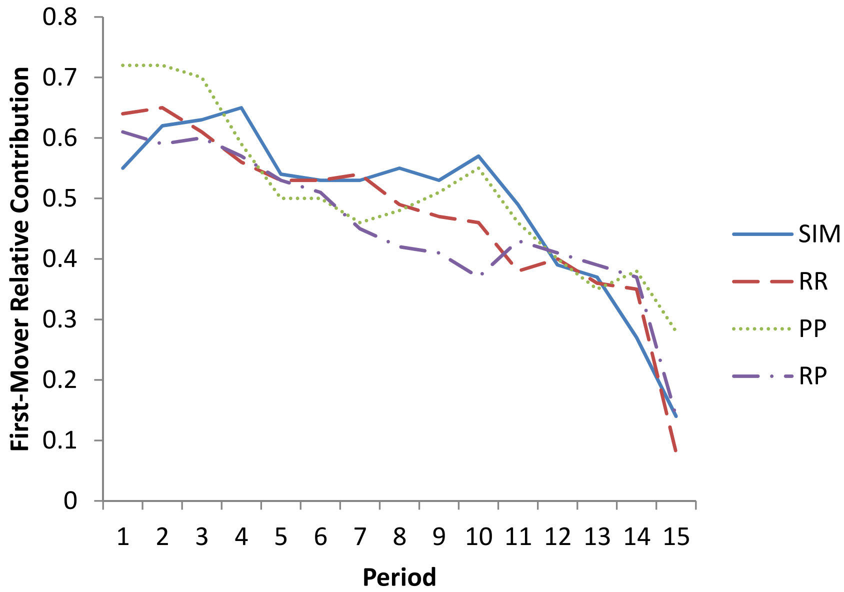


Figure 2.1: Development of first-mover relative contributions, Periods 1-15

We turn next to the experimental data and begin our investigation with first-mover contributions. Figure 2.1 illustrates the evolution of first-mover relative contributions by treatment in periods 1-15. We conclude from Figure 2.1 and accompanying regression models in Table 2.3 that there are no differences in first-mover average relative contributions across treatments (Average relative contributions in SIM are 49% of the initial endowment, in RR 47%, in PP 51% and in RP 45%).[14] As in abundant other studies on voluntary public good provision, we observe a strong decline in relative first-movers contributions over time in all treatments ($p < 0.01$).

[14] We include control variables here and in all following regression analyses. Control variables include age, a dummy variable for gender, a dummy variable for native German speaker, a dummy variable for West-German high school education and a continuous variable for the working hours per week.

Table 2.3: First-mover relative contributions

Relative Contribution		
SIM	0.041	(0.072)
PP	-0.011	(0.090)
RP	0.010	(0.087)
Period	-0.028***	(0.003)
Constant	0.610***	(0.155)
Controls	Yes	
Observations	1665	
χ^2	194.3	
$Prob > \chi^2$	0.000	

Notes: Random-effects GLS regressions. RR is baseline. Std. errors in parentheses are adjusted for individual clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Thus, the behavior observed in the experiment confirms the assumptions of our model. Both rich and poor players contribute the same relative amount from their endowment if they are first-movers. We summarize as follows.

Result 2.2 *There are no differences in first-mover relative contributions between the treatments.*

We continue to investigate how the normative fairness principles compare with actual observations from the experimental data. Figure 2.2 presents the applied contribution norms by plotting the kernel densities of average contribution ratios by rich and poor second-movers together with data from the questionnaire study. For each rich second-mover, we compute a contribution ratio based on subject's own contribution and the contributions made by the poor first-movers in the same group. For each poor second-mover, we compute the same contribution ratio based on subjects' own contribution and the contributions made by the rich first-movers in the same group. The ratios are thereafter averaged over the 15 periods to obtain one contribution ratio per subject. We interpret these ratios to represent the preferred normative principle of the subject. For instance, a contribution ratio that equals one implies that the subject follows the equality norm, whereas a contribution ratio that equals $\frac{25}{15} = \frac{5}{3} \approx 1.66$ implies that the subject follows the equity norm. Figure 2.2(a) plots the distribution of contribution ratios that the questionnaire participants find to be the most appealing contributions for rich individuals and the actual contributions ratios of rich individuals in the sequential

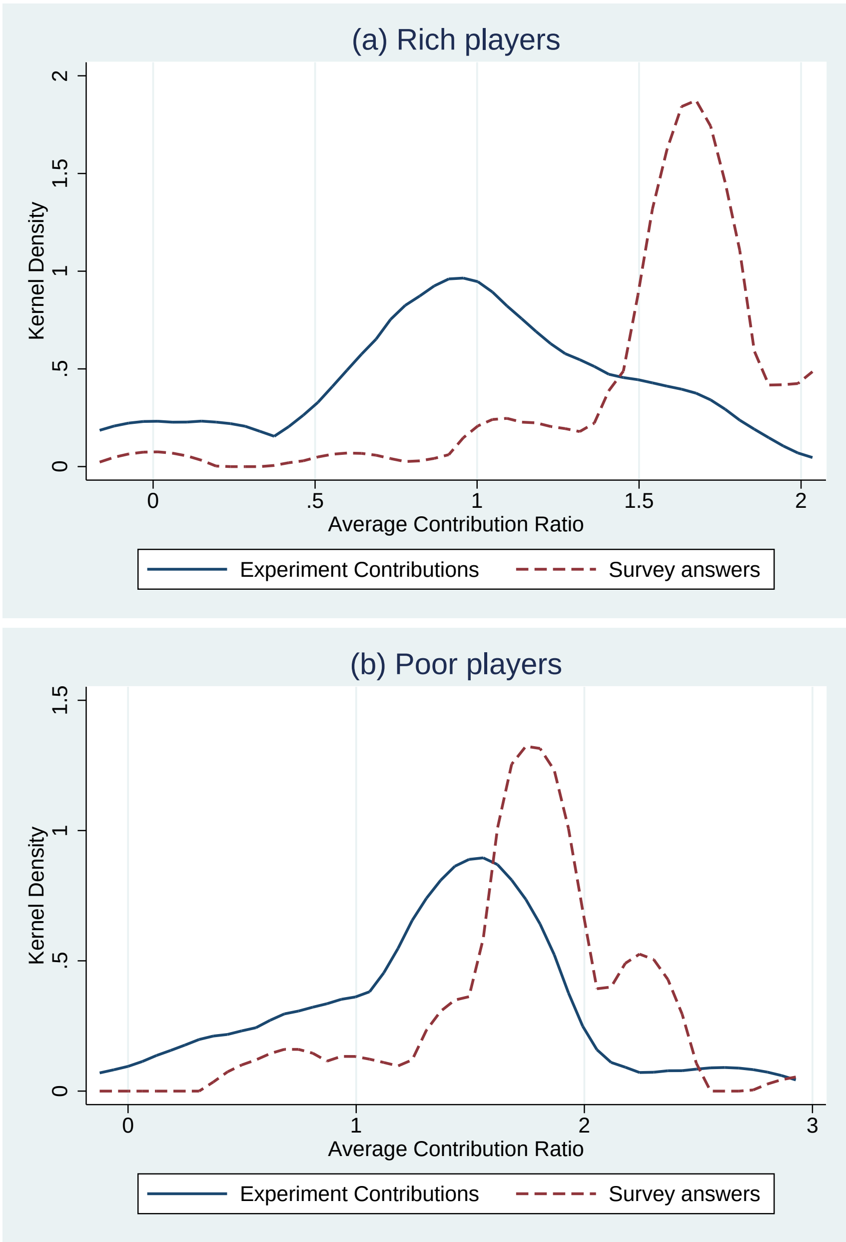


Figure 2.2: Kernel densities of average contribution ratios using the Epanechnikov kernel function. Bandwidths are calculated as to minimize the mean integrated squared error for an underlying Gaussian density.

treatments. Figure 2.2(b) plots the distribution of contribution ratios that the questionnaire participants find to be the most appealing contributions for poor individuals and the actual contributions ratios of poor individuals in the sequential treatments.

We make two observations from Figures 2.2(a) and (b). First, plotted distributions have multiple peaks. In particular, plotted density distributions have peaks at 1 and $\frac{5}{3}$ which correspond to the normative principles related to equality and equity. These observation suggests that normative views of fair contribution rules play a prominent role in the behavior of second-movers. Second, the preferred normative principle differs between player types. A two-sample Kolmogorov-Smirnov test for group level observations shows that the differences between the player types are significant ($D = 0.4167$, $p = 0.024$). The behavior of poor individuals coincides with the equity norm which is also the most appealing contribution for poor individuals from the viewpoint of impartial external observers. By contrast, the behavior of rich individuals coincides with the equality norm. Consequently, the behavior of rich individuals does not only deviate from the behavior of poor individuals but also from the normative principle preferred by the impartial external observers (a two sample K-S test confirms the observation, $D = 0.7467$, $p = 0.000$). The comparison of experimental and questionnaire data suggest that the differences between rich and poor individuals in incentivized contribution decisions do not coincide with the opinion of impartial external observers concerning fair contribution rules. Taken together, these observations suggest that rich individuals have a self-serving interpretation of desirable contribution rules which gives rise to conflicting normative views of fair contributions.

We summarize the observations as follows. First, in line with the assumptions of our theoretical model, rich and poor first-movers contribute the same relative amount to the public good in all treatments. Second, the revealed fairness principle differs across the rich and poor second-movers. The behavior of poor individuals coincides with the equity principle, whereas the behavior of rich individuals coincides with the equality principle. These observations suggest that self-serving interpretation of desirable contribution rules gives rise to conflicting normative views of fair contributions between the rich and poor individuals.

Result 2.3 *Self-serving interpretation of desirable contribution rules leads to conflicting normative views of fair contributions.*

Table 2.4: Summary statistics

		SIM	RR	PP	RP
Contributions	All	9.73	9.90	8.10	8.36
		(6.17)	(7.00)	(7.05)	(6.94)
	Rich	11.74	11.72	8.61	9.46
		(8.54)	(8.93)	(8.58)	(8.49)
	Poor	7.71	8.08	7.60	7.25
		(4.98)	(5.67)	(6.02)	(5.99)
Income	Inequality	5.97	6.36	8.99	7.79
		(6.56)	(5.26)	(4.55)	(4.84)
	Rich	28.82	29.12	29.36	28.91
		(3.17)	(3.21)	(3.45)	(3.51)
	Poor	22.85	22.76	20.36	21.12
		(6.24)	(6.24)	(5.84)	(5.83)

Notes: Average contributions and average income by treatment and type. Standard deviations using group averages as observation units in parentheses.

2.5.2 Public good provision and income inequality between treatments

Our theoretical analysis provides clear predictions about the relative performance of sequential contribution mechanisms in the presence of heterogeneous wealth distributions. In the following, we investigate the impact of alternative contribution mechanisms on public good provision and income inequality using our experimental data. If contributions are lower in treatments PR and PP than they are in RR and SIM, we have substantial evidence to support the proposition that the order of solicited contributions affects public good provision.

Table 2.4 shows the average contributions to public good across all periods by treatment and initial endowment as well as the average difference in earnings between the rich and poor individuals by treatment. Table 2.4 reveals that the order of solicited total contributions between treatments coincides with the predictions of our model. We find that the contribution levels are the highest in treatment RR (average contribution per player 9.9) followed by SIM (9.7), RP (8.4) and PP (8.1). Figure 2.3 depicts the evolution of average contributions over time in all treatments. When aggregating the total contributions over all periods and performing pairwise Mann-Whitney tests at the group level (n=12), we do not find any significant differences in total contributions between the treatments (see Table 2.5).

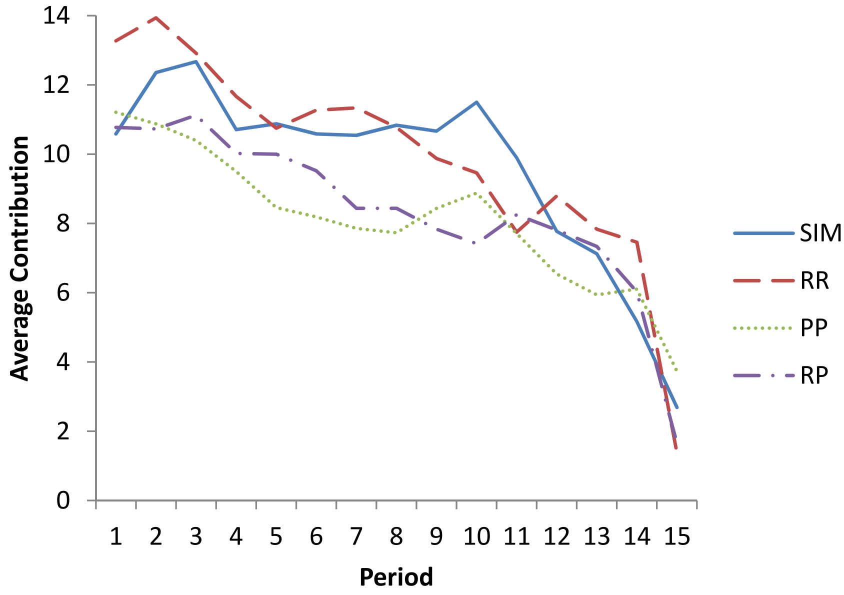


Figure 2.3: Development of average contributions, Periods 1-15

Table 2.5: Pairwise Mann-Whitney tests for contributions

		RR-SIM	RR-PP	RR-RP	SIM-PP	SIM-RP	PP-RP
All	z	0	0.866	0.636	1.039	0.520	-0.115
	p	1	0.387	0.525	0.299	0.603	0.908
Rich	z	-0.173	1.386	0.606	1.212	0.924	-0.289
	p	0.863	0.166	0.544	0.225	0.356	0.773
Poor	z	0.115	0.231	0.578	0.115	0	0.462
	p	0.908	0.817	0.564	0.908	1	0.644

Notes: In all cells the p-value is based on group level observations, n=12. *All* denotes total average contributions. *Rich* denotes average contributions by rich players. *Poor* denotes average contributions by poor players.* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

By contrast, regression models in Table 2.6 provide partial statistical support for the empirical relevance of our theoretical predictions. In Table 2.6, models (1), (3) and (5) use the treatment RR as a baseline and regress the average contributions in other treatments against this baseline treatment. We observe that the total contributions in treatment PP are significantly lower than in the treatment RR. Our theoretical model predicts no differences in total contributions between treatments RR and SIM. We do not observe statistically significant differences between these two treatments. Consequently, we combine the observations from these two treatments in models (2), (4) and (6) to have an unambiguous and theoretically motivated baseline to test the predictive power of our model. Model (2) compares average contributions between the treatments and shows that the total contributions are significantly lower in treatment PP than in the baseline. Models (4) and (6) add separate time trends for each treatment. Model (6) includes the same variables as model (4), but clusters the standard errors at group level. We observe that the treatment PP in all models exhibits significantly lower total contributions than the treatments RR and SIM. Finally, we estimate separate regression models for rich and poor individuals. In concordance with our previous analysis, we observe that the contributions of rich individuals are significantly lower in treatment PP than in treatment RR. We do not observe significant differences in contributions between the treatments among the poor individuals. Taken together, these observations lend partial empirical support for the theoretical proposition that altering the sequential order of solicited contributions in the presence of wealth heterogeneity may affect the total value of provided public good. In particular, altering the sequential order of contributions mainly affects the contributions of rich individuals.

Result 2.4 *We find partial empirical support for the proposition that a sequential mechanism with rich first-movers generates greater public good contributions than alternative sequential mechanisms.*

Table 2.6: Contribution regressions

	Overall average							
Contribution	(1)	(2)	(3)	(4)	(5)	(6)	Rich	Poor
SIM	0.473		-0.396		-0.396		-0.739	-0.808
	(1.249)		(1.673)		(2.335)		(3.276)	(1.948)
PP	-2.116	-2.337**	-3.953***	-3.754***	-3.953*	-3.754**	-5.677**	-2.575
	(1.302)	(1.097)	(1.446)	(1.215)	(2.275)	(1.882)	(2.866)	(1.941)
RP	-1.302	-1.525	-2.587	-2.390*	-2.587	-2.390	-3.796	-1.750
	(1.274)	(1.074)	(1.574)	(1.376)	(2.516)	(2.182)	(3.345)	(1.984)
Period			-0.660***	-0.607***	-0.660***	-0.607***	-0.745***	-0.566***
			(0.0755)	(0.0627)	(0.109)	(0.0952)	(0.144)	(0.102)
SIM × Period			0.109		0.109		0.107	0.115
			(0.126)		(0.191)		(0.258)	(0.152)
PP × Period			0.230**	0.177*	0.230	0.177	0.313	0.138
			(0.103)	(0.0935)	(0.155)	(0.145)	(0.196)	(0.146)
RP × Period			0.161	0.108	0.161	0.108	0.133	0.181
			(0.101)	(0.0917)	(0.146)	(0.136)	(0.195)	(0.147)
Constant	8.917***	9.110***	14.20***	13.97***	14.20***	13.97***	19.58**	12.21***
	(2.507)	(2.415)	(2.551)	(2.485)	(2.939)	(2.648)	(7.736)	(3.230)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2700	2700	2700	2700	2700	2700	1395	1305
χ^2	24.97	24.34	254.5	245.1	195.6	178.3	125.8	113.4
$Prob > \chi^2$	0.0016	0.0010	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Random-effects GLS regressions. RR is baseline in (1), (3), (5), Rich and Poor. RR/SIM is baseline in (2), (4) and (6). Std. errors in parentheses adjusted for individual clusters in (1), (2), (3) and (4), and for group clusters in (5), (6), Rich and Poor. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.7: Pairwise Mann-Whitney tests for inequality and income

		RR-SIM	RR-PP	RR-RP	SIM-PP	SIM-RP	PP-RP
Inequality	z	-0.029	-1.791	-0.981	-1.964	-0.924	1.733
	p	0.977	0.073*	0.326	0.050**	0.356	0.083*
Rich	z	0.173	0	0.115	-0.462	-0.173	0.462
	p	0.863	1	0.908	0.644	0.863	0.644
Poor	z	-0.058	1.386	0.693	1.242	0.981	-0.462
	p	0.954	0.166	0.488	0.214	0.326	0.644

Notes: In all cells the p-value is based on group level observations, n=12. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

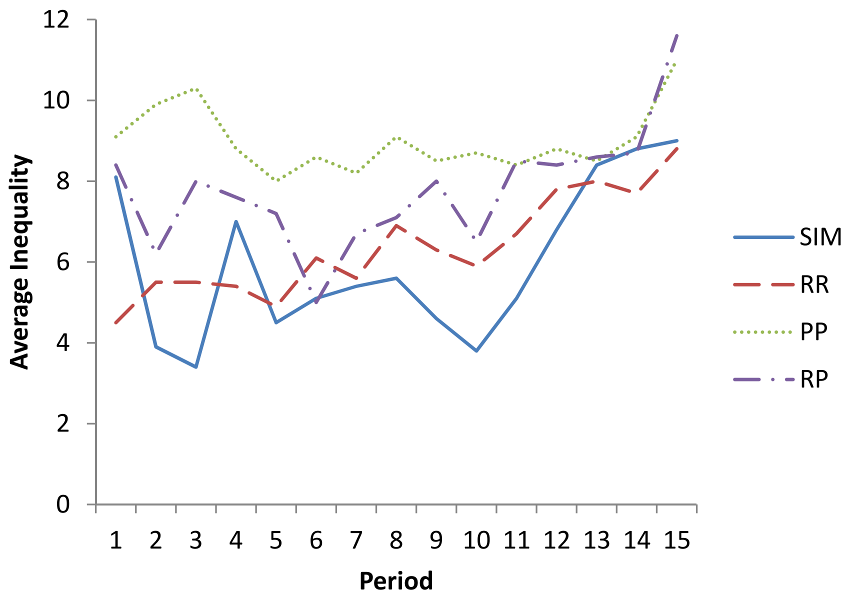


Figure 2.4: Development of income inequality, Periods 1-15

Common to all treatments is that inequality between the rich and poor individuals may change through positive contributions to the public good. Table 2.4 shows the average difference in income between the rich and poor individuals by treatment and Figure 2.4 illustrates the evolution of the income inequality over time. We observe that the income inequality is highest in treatment PP (average difference in earnings between the rich and poor individuals is 9.0) followed by the treatments RP (7.8), RR (6.4) and SIM (6). At the same time, the initial inequality between players diminishes in all treatments due to higher absolute contributions by rich individuals. Pairwise Mann-Whitney tests in Table 2.7 confirm that the inequality is larger in PP than in all other treatments.

Table 2.8 provides additional support for these observations.[15] The first two models (1, 2) use the difference in average income between rich and poor players in each group as the dependent variable. These regressions indicate that income inequality is significantly higher in treatments PP and PR than in treatment RR.

Our theoretical model predicts no differences in income inequality between treatments RR and SIM. We do not observe statistically significant differences between these two treatments. Consequently, we combine the observations from these two treatments in models (3) and (4) to have an unambiguous and theoretically motivated baseline to test the predictive power of our theoretical proposition germane to income inequality between treatments. We observe that the income inequality is larger in treatments PP and PR than in the combined baseline consisting of treatments RR and SIM. Finally, we estimate separate regression models for rich and poor individuals to study the earnings of different player types between treatments. We follow the same specification strategy as before and estimate models where we use either the treatment RR or the combination of treatments RR and SIM as a baseline treatment. We observe that the income of rich individuals does not differ between treatments (models 5 and 6). At same time, we find that the income of rich individuals is significantly lower in treatment PP than in treatment RR (model 7). We observe that this finding is robust to the pooling of treatments RR and SIM (model 8). In sum, these observations lend empirical support to our proposition that the order of solicited contributions in heterogeneous populations affects the ex-post income inequality between the rich and poor individuals.

Result 2.5 *A sequential mechanism with rich first-movers narrows wealth inequality more than any alternative sequential mechanism.*

[15] All regression models included in Table 2.8 cluster standard errors at the individual level. We study the robustness of the presented models and estimate all models with standard errors clustered at the group level. We find that all the main results presented in Table 2.8 are robust to clustering the standard errors at group level. The alternative regressions are attached in Appendix 2.B.

Table 2.8: Income inequality and income by type

	Income inequality				Income Rich		Income Poor	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SIM	-0.534	-0.237			-0.330		-0.545	
	(0.865)	(1.037)			(1.507)		(1.714)	
PP	2.983***	5.018***	3.233***	5.125***	0.262	0.424	-5.375***	-5.116***
	(0.764)	(0.924)	(0.675)	(0.875)	(1.237)	(1.094)	(1.676)	(1.543)
RP	1.521**	2.038**	1.772***	2.147**	-0.517	-0.354	-3.127*	-2.868*
	(0.715)	(1.028)	(0.638)	(0.992)	(1.246)	(1.110)	(1.818)	(1.680)
Period		0.255***		0.237***	-0.244***	-0.221***	-0.534***	-0.496***
		(0.0544)		(0.0443)	(0.0876)	(0.0618)	(0.0771)	(0.0715)
SIM × Period		-0.0371			0.0478		0.0808	
		(0.0893)			(0.124)		(0.145)	
PP × Period		-0.254***		-0.237***	0.00837	-0.0150	0.283***	0.245**
		(0.0693)		(0.0617)	(0.117)	(0.0989)	(0.103)	(0.0988)
RP × Period		-0.0647		-0.0468	0.0625	0.0391	0.168	0.129
		(0.0876)		(0.0817)	(0.110)	(0.0910)	(0.117)	(0.114)
Constant	5.192***	3.150*	4.974***	3.076*	24.94***	24.78***	33.18***	32.92***
	(1.868)	(1.874)	(1.861)	(1.847)	(2.675)	(2.559)	(3.678)	(3.541)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2700	2700	2700	2700	1395	1395	1305	1305
χ^2	37.49	97.01	36.74	88.26	50.06	49.80	130.8	118.6
$Prob > \chi^2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Random-effects GLS regressions. RR is baseline in (1), (2), (5) and (7). RR/SIM is baseline in (3), (4), (6) and (8). Std. errors in parentheses adjusted for individual clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.6 Conclusion

Numerous complex transfer payments schemes and coercive institutional solutions have been designed to overcome the problem of free-riding and achieve socially optimal levels of public good contributions. However, these theoretically optimal mechanisms have led to few successful applications to foster human cooperation in naturally-occurring situations. In practice, several different types of public goods have to be provided in the absence of strong institutions and coercive power to punish violators. Heterogeneous distribution of wealth among the actors is omnipresent and plays an apparent role in numerous problems of collective action in the absence of coercive institutions. In particular, the plurality of normatively appealing fair contribution rules in heterogeneous populations may cause unpredictable and destructive normative conflicts which impede the attainment of economically and socially efficient solutions. In contrast to the prevailing view, which emphasizes the destructive nature of conflicting normative ideals, this paper shows how conflicting equity principles can be used to design sequential contribution mechanisms to foster human cooperation.

The goal of this paper is to provide theoretical and empirical evidence on the role of conflicting normative principles for the financing of public goods through sequential contributions. We show the coexistence of equality and equity principles in heterogeneous populations using both questionnaire and experimental data. By combining these two sources of data, we show that the self-serving interpretation of desirable contribution rules leads to conflicting normative views of fair contributions. Based on these conflicting normative views of fair contributions, we are able to show how small changes to the sequential order of contributions may affect the total value of public good provision in heterogeneous groups. In particular, our model proposes that a sequential contribution mechanism which solicits contributions first from wealthy actors generates greater public good provision and narrows wealth inequality more than alternative mechanisms. Our experimental data lends partial empirical support for our propositions concerning the performance of alternative sequential mechanism. We observe that the highest contributions are provided in a treatment where contributions are solicited first from the wealthy individuals. At the same time, we find that the mechanism with wealthy first-movers narrows wealth inequality more than alternative mechanism, as predicted.

Our results may give guidance to various charitable corporations organizing public fund raising campaigns. We have provided both theoretical and empirical evidence showing how altering the sequential order of contributions may affect the total value of solicited contributions. However, much more evidence is needed before we can begin to make forceful arguments that the observed regularities readily generalize to naturally-occurring situations and characterize human behavior in utterly complex economic decisions. At the same time, the need to observe initial wealth distributions among the actors may hinder the applicability of proposed mechanism in naturally-occurring situations.

The robustness of our results can be analyzed in multiple new settings where the behavioral assumptions of the theory cannot necessarily be guaranteed to hold. One particularly promising approach to provide robustness checks for our laboratory results involves field studies in naturally-occurring environments. In particular, various fund-raising campaigns in pre-existing social groups are likely to provide opportunities for research in an environment where people have information about the initial wealth distribution among the potential donors. In a similar vein, increasing interest to disclose the income and assets of public sector officials in many countries is likely to create opportunities to develop new designs to evaluate the robustness of our experimental results in naturally-occurring settings. Finally, the public disclosure of personal income tax filings in many Scandinavian countries may create a natural test bed for future studies evaluating the proposition that conflicting normative views can be used to design contribution mechanisms to increase public good provision.

Appendix

2.A Proofs

Proof of Proposition 2.2

The optimal contribution level for maximizing utility function u_i result to

$$c_i(k, c_{-i}) = m(k, c_{-i}) - \frac{1-\alpha}{\beta}.$$

Since $k \in \{0, 1\}$, we only need to compare the utilities that player i gets from applying the two norms. The utility amounts to

$$u_i(\frac{c_{-i}}{n-1}, c_{-i}, 0) = w_i + c_{-i}(\alpha + (\alpha - 1)\frac{1}{n-1}) + \frac{(1-\alpha)^2}{2\beta}$$

when $k = 0$ and to

$$u_i(\frac{w_i}{w_{-i}}c_{-i}, c_{-i}, 1) = w_i + c_{-i}(\alpha + (\alpha - 1)\frac{w_i}{w_{-i}}) + \frac{(1-\alpha)^2}{2\beta},$$

respectively. It is then straightforward to show that

$$u_i(\frac{c_{-i}}{n-1}, c_{-i}, 0) > u_i(\frac{w_i}{w_{-i}}c_{-i}, c_{-i}, 1) \Leftrightarrow w_i > \frac{w_{-i}}{n-1}$$

and vice versa

$$u_i(\frac{c_{-i}}{n-1}, c_{-i}, 0) < u_i(\frac{w_i}{w_{-i}}c_{-i}, c_{-i}, 1) \Leftrightarrow w_i < \frac{w_{-i}}{n-1}.$$

Hence, if the endowment of player i is larger than the average endowments of the other players, she applies the equality norm. If the endowment of i is lower than the average, then the equity norm is applied.

Proof of Proposition 2.3

Suppose that there are h rich players and $l = n-h$ poor players. Poor players choose the same contribution level xw_l in all mechanisms, since rich players contribute xw_R as first-movers. However, rich players contribute at most $\frac{(h-1)xw_h+xw_l}{h} = x\frac{(h-1)w_h+w_l}{h} < xw_h$ as second-movers. Hence, the allocation of rich players on first- and second-movers triggers differences in provision levels across mechanisms and contributions by the rich are hxw_h in the rich first mechanism, $(h-1)xw_h + x\frac{(h-1)w_h+w_l}{h}$ in the mixed mechanism with one rich second-movers, etc., and hxw_P in the poor first mechanism. Thus, SEQ with rich first-movers generates the highest provision level and SEQ with poor first-movers the lowest provision level.

2.B Supplementary Tables

2.B.1 Alternative Categorization of Questionnaire Data

Table 2.9: Fraction of answers corresponding to normative fairness principles when accounting for calculation errors, n=100

	Equity	Equality	Efficiency	Equal Payoff	Other
Unconditional	.62	.04	.12	.15	.07
Conditional Rich	.56	.07	.05	.17	.15
Conditional Poor	.75	.04	.04	.09	.08

2.B.2 Additional Regressions for Income Inequality and Income by Type

Table 2.10: Income inequality and income by type

	Income inequality				Income Rich		Income Poor	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
SIM	-0.534	-0.237			-0.330		-0.545	
	(1.679)	(2.007)			(1.392)		(1.985)	
PP	2.983**	5.018***	3.233**	5.125***	0.262	0.424	-5.375**	-5.116***
	(1.450)	(1.789)	(1.274)	(1.686)	(1.253)	(1.008)	(2.099)	(1.959)
RP	1.521	2.038	1.772	2.147	-0.517	-0.354	-3.127	-2.868
	(1.378)	(2.031)	(1.227)	(1.956)	(1.328)	(1.101)	(2.274)	(2.136)
Period		0.255**		0.237***	-0.244***	-0.221***	-0.534***	-0.496***
		(0.109)		(0.0880)	(0.0864)	(0.0528)	(0.0933)	(0.0915)
SIM $\times$ Period		-0.0371			0.0478		0.0808	
		(0.177)			(0.105)		(0.186)	
PP $\times$ Period		-0.254*		-0.237*	0.00837	-0.0150	0.283**	0.245**
		(0.139)		(0.123)	(0.115)	(0.0930)	(0.123)	(0.121)
RP $\times$ Period		-0.0647		-0.0468	0.0625	0.0391	0.168	0.129
		(0.176)		(0.163)	(0.117)	(0.0945)	(0.146)	(0.145)
Constant	5.192***	3.150	4.974**	3.076	24.94***	24.78***	33.18***	32.92***
	(1.999)	(2.014)	(1.964)	(1.904)	(2.804)	(2.566)	(3.735)	(3.440)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2700	2700	2700	2700	1395	1395	1305	1305
χ^2	19.97	41.54	17.82	33.55	63.78	58.94	99.09	87.68
$Prob > \chi^2$	0.011	0.000	0.013	0.000	0.000	0.000	0.000	0.000

2.C Experimental Instructions and Questionnaire

2.C.1 Experimental Instructions

Instructions

Welcome and thank you for participating in this experiment. By carefully reading the instructions and depending on your decisions in the experiment you can earn a considerable amount of money. Your earnings will be privately paid to you in cash at the end of the experiment. The payment will be done in private, such that no other participant will know the amount of your earnings.

Please read these instructions carefully, they are solely for your private information. From now on, we ask you to remain seated and stop all communication with the other participants of the experiment. Please switch off your mobile phone and remove all non-necessary things from your seat. If you have any questions at any time during the course of this experiment, please raise your hand and a member of the experimenter team will privately assist you. It is important that you follow these rules as any violation will lead to exclusion from the experiment and all payments.

During the experiment all decisions and transfers are made in Experimental Monetary Units (EMUs). At the end of the experiment, your income will be calculated in EMUs and converted to Euro at the following rate:

$$1 \text{ EMUs} = 0.40 \text{ Euro}$$

This experiment consists of two stages. You will complete the first stage before learning more about the decisions during the second stage. At the beginning of the first stage you will be randomly matched with three other participants in this room. That is, you will be part of a group of four people.

STAGE 1

All group members are given the same task. You will be presented with a number of words on your computer screen and your task will be to encode these words by substituting the letters with numbers using a conversion table. presented on your computer screen. The same table is made available also on the bottom of this page. For all participants the sequential order of the presented words is the same.

Example: You are given the word SCHNITZEL. The letters in the Table show that S=1, C=7, H=2, N=18, I=8, T=12, Z=13, E=11, L=19.

Your performance during the task will determine your endowment size during the second stage. Your endowment will depend on the number of points you and the other participants in your group have after the encoding period of 10 minutes. The two persons with the highest numbers of encoded words will receive an endowment of 25 EMUs. The two persons with the lowest numbers of encoded words will receive an endowment of 15 EMUs. That is, if you have encoded more words than at least two other members of your group, you will receive an endowment of 25 EMUs. If you have encoded less words than at least two other members of your group, you will receive an endowment of 15 EMUs. You will not be told the exact number of words that the other group members have encoded and your exact rank within your group. The other group members will also get no information about the number of words you encoded. If two or more individuals have encoded the same number of words, the computer will randomly determine the ranking of the tied persons. Each person will have the same probability of being ranked above the other group members with the same amount of points.

A	B	C	D	E	F	G	H	I	J	K	L	M
8	12	14	10	9	6	24	22	7	5	11	3	18
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	21	16	23	2	13	19	25	4	26	17	20	15

STAGE 2 [for treatment RR]

The second stage of the experiment consists of 15 consecutive decision periods. In all 15 periods you will make your decisions in the same group of four people that you were matched with at the beginning of the experiment. That means that you will be interaction with the same three people throughout the experiment.

As a result from part 1 of the experiment, at the beginning of each round there are two members in your group who are endowed with 25 EMUs and two members who are endowed with 15 EMUs. Your task is to decide how many EMUs from your endowment you want to keep in your private account and how many EMUs you want to contribute to a group accounts. Each unit not allocated to the group account will be automatically remain in your private account.

The members of your group will not all decide simultaneously about the allocation of endowments to the private and group account. More precisely, there will be two early deciders and two late deciders which each decide simultaneously. The late deciders will be informed about the decision of the early deciders, before making their decisions. All participants however will be informed about the decisions of all other group members at the end of the period. For that reason, before the first period every group member will be assigned an identity A or B, and C or D, respectively. The two early deciders both have an endowment of 25 EMUs and identities A or B and the two late deciders have an endowment of 15 EMUs and identity C or D. That is, if you earned an endowment of 25 EMUs in part 1 of the experiment, you are an early decider. If you earned an endowment of 15 EMUs, you are a late decider.

Your income from the private account

Each unit allocated to your private account earns you one EMU. No other member in your group benefits from transfers to your private account.

Income from private account = Your endowment - Your contribution to group account

Your income from the group account

For each EMU you contribute to the group account, you will earn 0.4 EMUs. Each of the other three people in your group will also earn 0.4 EMUs. Thus, the contribution of 1 EMU to the group account yields a total of 1.6 EMUs for all of group members together. Your earnings from the group account are based on total number of EMUs contributed by all members in your group. Each group member will earn equally from the contributed amount. Thus, you earn both from your own contribution as well as from the contributions of all other group members. The income from the group account will therefore for all group members be determined by

Income from group account = Sum of contributions to group account $\cdot 0.4$

Your total income

Your total earnings result per period is the sum of your income from your private account and the income from the group account.

Total income = Earnings from your private account + Earnings from the group account

BONUS STAGE

In addition to the actual allocation decision, we ask you to separately estimate the average contribution by the other three members of your group. You will have to make these estimations before the actual decisions are made.

You will be asked to estimate the contributions of group members with endowments 15 EMUs and 25 EMUs separately. That is, if you have an endowment of 25 EMUs (15 EMUs), you estimate the contribution of the other group member with an endowment of 25 EMUs (15 EMUs) and the average contribution of the two group members with an endowment of 15 EMUs (25 EMUs). Please give your estimations in integers, as - if necessary - the average contribution will be rounded to the nearest integer. You will be rewarded for the accuracy of your estimates.

(I.) If your estimate exactly corresponds with the actual (average) allocation, you will earn 2 EMUs.

(II.) If your estimate deviates one EMU from the actual (average) allocation, you will earn 1 additional EMU.

(III.) If your estimate deviates two or more EMUs from the actual (average allocation), you will not earn any additional EMUs.

Hence, per period you can additionally earn a total of 4 EMUs by correct estimations.

Your total earnings at the end of the experiment

For your earnings at the end of the experiment one out of the 15 periods will be randomly chosen as relevant for payment for both total income and bonus. That is, the total income of one period will be paid out at the end and therefore each period may be relevant for your earnings. The same is true for the bonus payments. The payment relevant period may be the same for income and bonus, but it does not have to be the same. Your total earnings therefore comprise of

Your total income from the randomly drawn period (converted into Euro)

+ Your earnings from the bonus stage in the randomly drawn period (converted into Euro)

= Your total earnings from the experiment

2.C.2 Questionnaire

Welcome! Thank you for participating in this questionnaire. Answering all questions will take about 10 minutes. If you complete this questionnaire, you will have the chance to win 100 Euro. Among 100 participants you have the chance to two times 100 Euro. Please enter your e-mail address at the end of the questionnaire, so that we can contact you if you have been drawn as a winner.

This questionnaire consists of two parts. In the first part we ask you to make yourself acquainted with a decision making problem from an economic experiment. You will get the experimental instructions, which you please read carefully. However, you will not yourself take part in this experiment but act as a neutral external observer. The questions that you are asked to respond to will be explained more closely in part 2 of the questionnaire.

PART 1

Please carefully read the following instructions for the proceedings of an economic experimental and answer the subsequent control questions. The control questions are designed as to help you to understand the decision problem and to make sure that you have understood the instructions correctly. If you have found the correct answer for a question, please enter it in the corresponding box. As soon as all questions have been answered correctly, you may proceed with part 2 of this questionnaire.

Experimental Instructions

This experiment is design such that the other participants will not learn anything about your decisions and earnings from the experiment. You will make all your decisions anonymously at your computer. During the experiment, all amounts will be displayed in Experimental Monetary Units (EMUs). At the end of the experiment, your income will be calculated in EMUs and converted to Euro at the following rate:

$$1 \text{ EMUs} = 0.40 \text{ Euro}$$

This experiment consists of two stages. You will complete the first stage before learning more about the decisions during the second stage. At the beginning of the first stage you will be randomly matched with three other participants in this room. That is, you will be part of a group of four people.

Stage 1

All group members are given the same task. You will be presented with a number of words on your computer screen and your task will be to encode these words by substituting the letters with numbers using a conversion table. presented on your computer screen. The same table is made available also on the bottom of this page. For all participants the sequential order of the presented words is the same.

Example: You are given the word SCHNITZEL. The letters in the Table show that S=1, C=7, H=2, N=18, I=8, T=12, Z=13, E=11, L=19.

Your performance during the task will determine your endowment size during the second stage. Your endowment will depend on the number of points you and the other participants in your group have after the encoding period of 10 minutes. The two persons with the highest numbers of encoded words will receive a higher endowment than the two persons with the lowest numbers of encoded words. That is, if you have encoded more words than at least two other members of your group, you will receive a high endowment. If you have encoded less words than at least two other members of your group, you will receive a low endowment. You will not be told the exact number of words that the other group members have encoded and your exact rank within your group. The other group members will also get no information about the number of words you encoded. If two or more individuals have encoded the same number of words, the computer will randomly determine the ranking of the tied persons. Each person will have the same probability of being ranked above the other group members with the same amount of points.

A	B	C	D	E	F	G	H	I	J	K	L	M
8	12	14	10	9	6	24	22	7	5	11	3	18
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	21	16	23	2	13	19	25	4	26	17	20	15

Stage 2

As a result from part 1 of the experiment there are two members in your group who have a high endowment and two members who have a low endowment. Your task is to decide how many EMUs from your endowment you want to keep in your private account and how many EMUs you want to contribute to a group accounts. Each unit not allocated to the group account will be automatically remain in your private account.

Your income from the private account

Each unit allocated to your private account earns you one EMU. No other member in your group benefits from transfers to your private account.

Income from private account = Your endowment - Your contribution to group account

Your income from the group account

For each EMU you contribute to the group account, you will earn 0.4 EMUs. Each of the other three people in your group will also earn 0.4 EMUs. Thus, the contribution of 1 EMU to the group account yields a total of 1.6 EMUs for all of group members together. Your earnings from the group account are based on total number of EMUs contributed by all members in your group. Each group member will earn equally from the contributed amount. Thus, you earn both from your own contribution as well as from the contributions of all other group members. The income from the group account will therefore for all group members be determined by

Income from group account = Sum of contributions to group account $\cdot 0.4$

Your total income

Your total earnings result per period is the sum of your income from your private account and the income from the group account.

Total income = Earnings from your private account + Earnings from the group account

PART 2

Scenario 25-15

In this scenario the two better encoders from stage 1 receive in stage 2 an endowment of 25 EMUs and the two worse encoders an endowment of 15 EMUs. We ask you to give your personal opinion regarding fair behavior from the viewpoint of a neutral external observer in the economic experiment from which you have read the instructions. With fair behavior, we mean behavior that may be characterizes as "socially appropriate" or "ethically correct".

[Screen 1] From the viewpoint of a neutral external observer, what is in your opinion a fair contribution to the group account by a group member who has an endowment of 25 EMUs? From the viewpoint of a neutral external observer, what is in your opinion a fair contribution to the group account by a group member who has an endowment of 15 EMUs?

[Screen 2a] A group member with endowment 15 EMUs has contributed 9 EMUs to the group account. From the viewpoint of a neutral external observer, in this case what is in your opinion a fair contribution to the group account by a group member who has an endowment of 25 EMUs?

[Screen 2b] A group member with endowment 25 EMUs has contributed 9 EMUs to the group account. From the viewpoint of a neutral external observer, in this case what is in your opinion a fair contribution to the group account by a group member who has an endowment of 15 EMUs?

Scenario 30-10

In this scenario the two better encoders from stage 1 receive in stage 2 an endowment of 30 EMUs and the two worse encoders an endowment of 10 EMUs. We ask you to give your personal opinion regarding fair behavior from the viewpoint of a neutral external observer in the economic experiment from which you have read the instructions. With fair behavior, we mean behavior that may be characterizes as "socially appropriate" or "ethically correct".

[Screen 1] From the viewpoint of a neutral external observer, what is in your opinion a fair contribution to the group account by a group member who has an endowment of 30 EMUs? From the viewpoint of a neutral external observer, what is in your opinion a fair contribution to the group account by a group member who has an endowment of 10 EMUs?

[Screen 2a] A group member with endowment 10 EMUs has contributed 9 EMUs to the group account. From the viewpoint of a neutral external observer, in this case what is in your opinion a fair contribution to the group account by a group member who has an endowment of 30 EMUs?

[Screen 2b] A group member with endowment 30 EMUs has contributed 9 EMUs to the group account. From the viewpoint of a neutral external observer, in this case what is in your opinion a fair contribution to the group account by a group member who has an endowment of 10 EMUs?

Scenario Competition

In this scenario the two better encoders from stage 1 receive in stage 2 an endowment of 25 EMUs and the two worse encoders an endowment of 15 EMUs. Assume additionally, that the decision making problem is changed as follows: The total amount of EMUs on the group account is now compared with the group account of another group. If the sum of contributed EMUs of one group is larger than the sum of the other group, then the group with the larger group account wins the difference to the losing group. At the same time, the group with the smaller sum on the group account loses the difference to the winning group. Gains and losses from this group account comparison are equally distributed among group members.

For example: The sum of EMUs in the group account amounts to 60 EMUs in group A and to 20 EMUs in group B. The differences is thus 40 EMUs. This difference is now equally distributed among the group members of group A. All group members of group A receive 10 EMUs additionally to their income from their private and group accounts and all group members of group B lose 10 EMUs from their income from private and group accounts.

We ask you to give your personal opinion regarding fair behavior from the viewpoint of a neutral external observer in the economic experiment from which you have read the instructions. With fair behavior, we mean behavior that may be characterizes as "socially appropriate" or "ethically correct".

[Screen 1] From the viewpoint of a neutral external observer, what is in your opinion a fair contribution to the group account by a group member who has an endowment of 25 EMUs? From the viewpoint of a neutral external observer, what is in your opinion a fair contribution to the group account by a group member who has an endowment of 15 EMUs?

Chapter 3

Wealth Inequality, Perceptions of Normatively Fair Behavior and Cooperation[1]

Abstract This paper investigates the impact of increasing wealth inequality on people's perceptions of fairness and willingness to cooperate. We combine experimental and survey data to distinguish people's normative perceptions of fairness from their revealed preferences and measure their consistency under different wealth distributions. We show that increasing wealth inequality does not alter people's normative perceptions of fairness, but may erode revealed contribution norms. Increasing wealth inequality shifts the revealed contribution norms of less wealthy individuals from equal relative contributions to equal absolute contributions. As a result, total cooperativeness decreases and the sequence of collected public good contributions becomes largely irrelevant in populations with large wealth inequality.

Keywords Heterogeneous Wealth, Fairness Perceptions, Sequential Public Good Provision
JEL Classification C92, D63, H41

[1] This chapter is co-authored by Lauri Sääksvuori.

3.1 Introduction

The distribution of wealth between individuals and social groups is a source of a seemingly eternal debate among scientists and the public. The debate often ponders whether the observed disparities in wealth are morally, politically and economically justified. However, at least equally relevant are questions about the potential behavioral and societal consequences of wealth disparities. Lately, the intellectual and political discussions about the fair distribution of wealth have greatly profited from the collection of historical data documenting the development of wealth and income distributions in the long run. The data show that the inequality of wealth has been constantly increasing since the 1970s across the rich countries (Piketty, 2014).

It appears therefore desirable to understand how the changing levels of wealth inequality shape our perceptions of fairness and affect our willingness to cooperate in evidently heterogeneous social groups and societies. This paper investigates the impact of increasing wealth inequality on people's perceptions of fairness and willingness to cooperate. We combine experimental and survey data to distinguish people's normative perceptions of fairness from their revealed preferences for fairness under different wealth distributions.

The economic debate about the consequences of wealth disparities on people's willingness to cooperate has for a long time relied on theoretical investigations and empirical studies using naturally-occurring data. For example, Meltzer and Richard (1981) show in an early theoretical study that larger income inequality may result to greater public good provision, if the mean income at the same time rises relative to the income of the median voter. In a similar vein, Warr (1983) and Bergstrom et al. (1986) show that the total supply of public goods is largely unaffected by small redistributions of wealth. However, Alesina and Drazen (1991) propose that the existence of heterogeneous interest groups in the society may lead to large distortions in the public good supply. In the empirical literature, Alesina and la Ferrara (2000) use survey data on group membership and data on U.S. localities to show that inequality decreases the level of privately provided public goods in local communities.

In contrast to the earlier theoretical and empirical attempts to study the effects of various wealth distributions on human cooperativeness, this paper uses survey and experimental data to quantify the impact of increasing wealth inequality on people's perceptions of fairness and willingness to cooperate. Our controlled experiment and accompanying internet survey enables us to quantify the opinion of impartial observers and measure participants' revealed preferences for fairness under different wealth distributions, while all other factors are kept constant. Furthermore, by combining experimental and questionnaire data we can measure the consistency between people's perception of fairness and revealed preferences under different wealth distributions.

The experimental literature on the impact of income heterogeneity on public good contributions remains inconclusive. Ostrom et al. (1994), van Dijk et al. (2002) and Cherry et al. (2005) report evidence showing that contributions to public goods are lower in groups with heterogeneous endowments than in groups with homogenous endowments. In contrast, Chan et al. (1996) and Chan et al. (1999) report that income heterogeneity among individuals increases voluntary contributions. Sadrieh and Verbon (2006) provide evidence that neither the extent nor the skew of inequality affects cooperativeness. In a similar vein, Hofmeyr et al. (2007) find that randomly induced wealth heterogeneity affects neither the total contributions to public good nor the relative shares of individual contributions. Buckley and Croson (2006) observe that less wealthy individuals contribute the same absolute amount as the more wealthy individuals. More closely related to our investigation is the study by Keser et al. (2011), who distinguish between weakly and strongly asymmetric initial endowments and show that the total contributions are significantly lower in the strongly asymmetric treatment.

Our study diverges from the previous literature in multiple important ways. First, we induce wealth heterogeneity by introducing a real-effort tournament and rewarding the best performers with greater initial endowments. This practice, arguably, enhances the prominence of conflicting normative views of fair contributions compared to a random allocation of initial wealth. Second, we elicit people's perceptions of fairness and willingness to contribute always in the presence of wealth inequality, enabling us to examine the effect of increasing inequality in groups with multiple plausible contribution norms. Consequently, we circumvent the problem that the notions of *equality* (equal absolute contributions) and *equity* (equal relative contributions) overlap in homogenous wealth distributions. Third, we study people's normative fairness perceptions and revealed preferences in a sequential game. This practice enables us to examine how individuals endowed with different wealth levels condition their behavior on the contributions of other individuals. Finally, the sequential order of decisions allows us to provide causal evidence about the consequences of increasing wealth inequality in a large class of economic activities with sequential decision structures.

Due to the sequential decision structure, our investigation connects to a growing literature studying the effects of leadership in collective action. Several recent studies have shown that a leader who sets a good example may increase the level of cooperativeness (Moxnes and van der Heijden, 2003; Güth et al., 2007; Rivas and Sutter, 2011). More relevant to our study is the observation that a good leader can spur the rest of the group to increase their contributions even in heterogeneous populations (Levati et al., 2007). Finally, this paper relates to an accompanying paper where we investigate how conflicting normative views of fairness can be used to design alternative sequential contribution mechanisms to foster cooperation (Neitzel and Sääksvuori, 2014a). Therein, we show that a sequential contribution mechanism which solicits contributions first from wealthy actors generates greater public good provision than any alternative sequential mechanism.

The plurality of normatively appealing fairness principles has been recognized as an important phenomenon of economic decision-making (Konow, 2003; Cappelen et al., 2007; Lange et al., 2010). Nikiforakis et al. (2012) show that the existence of multiple plausible contribution norms may lead to a sequence of mutually retaliatory sanctions. Reuben and Riedl (2013) combine experimental and questionnaire data and show the co-existence of efficiency, equality and equity norms in a public goods game with punishment opportunities.

In this paper we find that the principle of equity dominates people's normative perceptions of fairness independent of the underlying wealth inequality. At the same time, increasing wealth inequality leads to larger deviations of revealed contribution norms from the normative principles preferred by the impartial observers. In particular, adherence to the equity principle diminishes among the less wealthy individuals as the inequality increases. This erosion of revealed contribution norms leads to smaller public good provision as the inequality increases.

3.2 Data Collection

3.2.1 Experimental Design and Procedure

We use the same experimental set-up as in Neitzel and Sääksvuori (2014a). Therefore, we only provide a brief overview of the experimental design.[2]

At the beginning of each session, the individuals are assigned into groups of four and stay in the same group for the entire experiment. The experiment itself is divided into two parts. In the first part individuals compete in a real-effort tournament task. We use the Encryption task introduced by Erkal et al. (2011) in order to determine the endowments for the sequential public goods game in the second part. In each group the two participants who encrypt most words receive a high endowment w_R, in following

[2] An English translation of the used instructions can be found in Appendix 2.C of Chapter 2. The instructions were originally written in German.

Table 3.1: Experimental treatments

	WEAK wealth inequality $w_R = 25$ & $w_P = 15$	STRONG wealth inequality $w_R = 30$ & $w_P = 10$
Rich first-movers	RR-W	RR-S
Poor first-movers	PP-W	PP-S

referred to as rich players, and the other two participants receive a lower endowment w_P, hereafter called poor players.[3]

In the second part of the experiment the participants play a sequential public goods game in which there are two first-movers who contribute simultaneously and two second-movers who contribute simultaneously after having observed the contributions made by the first-movers. We use the standard configuration for a four-player linear public goods game in which monetary payoff of player i is determined by

$$\pi_i(c_i) = w_i - c_i + 0.4 \cdot C,$$

where w_i is i's endowment, c_i is i's contribution and C is the sum of contributions by all four players.

Our experimental design consists of four treatments, varying the sequence of play and the degree of wealth inequality. We implemented two treatments in which the two rich individuals contribute first and two treatments in which the two poor individuals contribute first. Table 3.1 provides the overview of the 2×2 experimental design.

We keep overall wealth constant at $W = w_R + w_P = 40$ in all treatments. In two WEAK treatments rich players receive endowments of 25 points and poor players endowments of 15 points. In two STRONG treatments the endowments are 30 points and 10 points for rich and poor players, respectively. The game is played for 15 consecutive periods in all treatments. After each period, subjects are informed about the total contribution in their group, the contribution of each individual and whether the individual making a specific contribution is rich or poor.

The experiment was conducted at the experimental laboratory of the School of Business, Economics and Social Sciences at the University of Hamburg. We used z-tree (Fischbacher, 2007) for programming and ORSEE (Greiner, 2004) and H-Root (Bock et al., 2012) for recruiting. We ran in total 8 different sessions with 192 subjects. The vast majority of 88 female (Age Mean: 24.3, Std: 2.98, Min: 18, Max: 38) and 104 male

[3] See Chapter 2 for a brief discussion of why we use a real-effort instead of a lottery for the allocation of initial endowments.

(Age Mean: 25.2, Std: 4.19, Min: 19, Max: 38) subjects were undergraduate students from various disciplines, but predominantly from social sciences. At the beginning of each session, participants were randomly assigned to their seats. To ensure common knowledge, the written instructions were distributed and read publicly by a member of the research team before subjects had to answer a set of control questions. Participants were not allowed to continue with the actual experiment until all control questions were answered correctly.

At the end of the experiment, one of the 15 periods was chosen at random to determine the earnings from the public goods game. The sessions lasted approximately 75 minutes including instructions, post instruction quiz, demographic questionnaire and payment procedure. Earnings per participant were on average 13.43 €.[4]

3.2.2 Questionnaire Study

The questionnaire was conducted online using a subjects pool with equal demographic characteristics as in the experiment. No subject was however allowed to participate in both studies. The questionnaire started with the same instructions and with the same control questions as the experiment. In particular, the external observers were informed about the real-effort task and the origin of unequal endowments for the public goods game. Before reading the instructions, participants were only told that they would be asked to answer questions regarding fair behavior in various situations.

Each external observer answered the same questions for three different scenarios, whereof the two scenarios of interest for this study only differed from each other in the distribution of wealth. The third scenario with weak wealth inequality and competition between groups is not analyzed in this paper. The order of the scenarios' appearance on the screen and the order of questions within each scenario was randomized to control for possible order effects. In each scenario, the observers were asked to answer two types of questions. The first type elicited unconditional contribution norms by asking for fair contributions by both rich and poor players simultaneously. The second type elicited conditional contribution norms by asking for fair contributions by rich and poor players

[4] Total earnings include revenues from a task in which participants were asked to predict the contributions of the other members in their group. Like for the earnings from the public goods game, one of the 15 periods was chosen at random to determine the earnings from the prediction task. Total earnings per participant differ significantly between WEAK and STRONG. In the WEAK treatments participants earned on average 11.37 €, while the earnings were on average 15.49 € in STRONG treatments. This large difference arises from an extra show-up fee of 5 € we needed to pay in order to fulfil the regulations of the lab. However, the participants were not informed of the show-up fee in advance. Correcting for the 5 € show-up fee, the average earnings were 10.49 € in STRONG treatments and overall average earnings from the game were 10.93 €.

separately, given a fixed contribution of the respective other type. A German translation of the general introduction to the questionnaire study and the precise wording of each question can be found in Appendix 2.C of Chapter 2.

A total of 100 subjects participated in the questionnaire study. There were 40 female (Age Mean: 24.5 Std: 3.1, Min: 20, Max: 32) and 60 male (Age Mean: 25.8, Std: 5.65, Min: 17, Max: 55) mostly undergraduate students from different disciplines. Two randomly drawn subjects were each awarded with 100 € for completing the questionnaire.

3.3 Normative Principles of Fairness

In this study we mainly investigate the prevalence of two particularly prominent fairness norms that relate individual contributions to initial wealth. The principle of *equality* suggests the equalization of absolute contributions with no link to initial wealth. By contrast, the principle of *equity* links contributions to initial wealth in a proportional manner. Hence, the principle of equity stipulates the equalization of relative contributions to common projects.[5] To enhance the comprehensiveness of our investigation about people's perceptions of normatively fair behavior, we additionally include the principles of *efficiency* and *equal payoff* in our analysis. We refer to the efficiency principle when an impartial observer finds the efficiency maximizing contribution to be the fairest possible contribution. The principle of equal payoff rule refers to an opinion which states that a contribution aiming to equalize final payoffs is the fairest of all possible contributions. In this case, a rich individual should contribute more than a poor individuals and completely close the wealth gap between due to initial differences in endowments.

3.4 Processing of Questionnaire Data

Table 3.2 shows the applied categorization rules for the questionnaire data. *Both* refers to the questions in which subjects were asked to state unconditional contribution norms for both player types simultaneously. *Rich* and *Poor* refer to the questions in which subjects were asked to state the conditional contribution norms for rich and poor players. We chose the contribution of the hypothetical first-mover to be 9 in all questions eliciting conditional contribution norms. Thereby we can separate the alternative fairness principles while at the same time being able to compare perceptions across both wealth distributions. The reference contribution of 9 was chosen as to maximize the number of

[5] We have previously theoretically shown that wealthier individuals prefer the equality principle, whereas poor individuals prefer the equity principle (Neitzel and Sääksvuori, 2014a).

Table 3.2: Strict categorization rules

	WEAK			STRONG		
	Both	Rich	Poor	Both	Rich	Poor
Equity	$c_R = \frac{5}{3}c_P$	15	5	$c_R = 3c_P$	27	3
Equality	$c_R = c_P$	9	9	$c_R = c_P$	9	9
Efficiency	$c_R + c_P = 40$	25	15	$c_R + c_P = 40$	30	10
Equal Payoff	$c_R = c_P + 10$	19	0	$c_R = c_P + 20$	29	0

Notes: Reference contribution was set to 9 in all questions eliciting conditional contribution norms. For Equity, we use rounded values for the classification of answers.

fairness principles' contributions being integers. This practice enabled subjects to match the alternative contribution principles in the simplest manner.

Note that answers prescribing full contributions by both types in the unconditional questions may be classified into the three alternative categories: *equity*, *efficiency* and *equal payoff*. In such case, the answer was allocated to *efficiency* as the most restrictive fair contribution rule.[6] Furthermore, we use the rounded values to classify answers to *equity* for the unconditional contribution norms, whenever applicable. In similar vein, we classified all answers of 5 to *equity* as the rounded value of the exact answer 5.4 for the conditional contribution norm of poor players under weak wealth inequality. *Equal payoff* comprises all answers of zero as the nearest possible value of -1 WEAK and of -11 in STRONG for poor players' conditional contribution norms. We argue that these practices provide the most conservative distribution of perceived fairness principles.

To allocate a larger share of survey answers to the four fairness principles, we undertake an alternative categorization. Therein, we allow for small decision errors of ± 1. The exact categorization rules with small decision errors are shown in Appendix 3.A. In accordance with the previous practice to provide the most conservative distribution of perceived fairness principles, we always classify an answer to the most restrictive rule whenever two rules conflict. For strong wealth inequality, we account for the fact that contributions by rich individuals required for *efficiency* and *equal payoffs* are not sufficiently distinct to allow for errors. The same practice in necessary for poor player contributions required for *efficiency* and *equality*. In these cases, the allocation was done as in the strict categorization.

[6] Here and in the following, by *most restrictive fair contribution rule* we denote the rule that comprises least combinations of contributions.

3.5 Results

3.5.1 Impact of Increasing Wealth Inequality on the Perceptions of Fair Contributions

Table 3.3: Fraction of answers corresponding to normative fairness principles, n=100

(a) Strict categorization

	WEAK			STRONG		
	Both	Rich	Poor	Both	Rich	Poor
Equity	.24	.36	.45	.36	.32	.52
Equality	.03	.02	.01	.01	.02	.04
Efficiency	.12	.05	.04	.11	.06	.01
Equal Payoff	.15	.10	.05	.03	.09	.12
Other	.46	.47	.45	.49	.51	.31

(b) Categorization allowing for small errors

	WEAK			STRONG		
	Both	Rich	Poor	Both	Rich	Poor
Equity	.62	.56	.75	.60	.34	.75
Equality	.04	.07	.04	.01	.06	.04
Efficiency	.12	.05	.04	.11	.06	.01
Equal Payoff	.15	.17	.09	.03	.13	.16
Other	.07	.15	.08	.25	.41	.04

Table 3.3 summarizes the impartial observers' opinions about fair contributions both under weak and strong wealth inequality. Table 3.3(a) presents the fraction of answers corresponding to the normative contribution principles using the strict categorization rules defined in the previous section, while Table 3.3(b) displays the classification according to the categorization allowing for small decision errors.

Using the strict categorization, we observe that in all questions the simple majority of classifiable opinions coincides with the equity norm. By the use of the less restrictive categorization rules, the number of unclassified answers declines substantially in five out of six questions. The vast majority of additionally allocated answers are classified to *equity*. As a result of the looser categorization, the normative principle of equity is considered to be the fair contribution rule by the absolute majority of impartial observers in all cases but for rich players in STRONG. We therefore find forceful evidence that equity is perceived to be the most appropriate contribution rule for almost all sequences and wealth disparities.

We use the categorization allowing for small errors to test whether an increase in wealth inequality has an impact on the distribution of perceived fair contribution rules. For the elicited unconditional contribution norms, we find that the fractions of answers differ between WEAK and STRONG (Fisher's exact test, $p = 0.000$). This result is driven by differences in the number of classified answers into *equal payoff* and the number of unclassified answers. If we combine *equal payoff* and *other* to one category, the difference vanishes (Fisher's exact test, $p = 0.496$).

For the conditional contribution norms, we find disparate results. For poor players, we do not find any differences in the distribution of most appealing contribution rules (Fisher's exact test, $p = 0.296$), while there is a significant difference for rich players (Fisher's exact test, $p = 0.001$). This finding in line with the previous observation that the classification for rich players in STRONG differs from all other distributions, in particular in the number of unclassified answers. We take a closer look at the data and find that the difference is due to a large number of observers stating 25 as the fair contribution. We do not find any significant differences between rich player conditional contribution norm distributions, if we allocate these answers to *equity* (Fisher's exact test, $p = 0.645$). Overall, we find sufficient evidence to state that the level of wealth inequality has no or only very limited impact on the perceived fair contributions.

We summarize the analysis of the questionnaire data as follows.

Result 3.1 *Opinions of the impartial observers coincide with equity and are independent of the level of wealth inequality among the group members.*

3.5.2 Impact of Increasing Wealth Inequality on Revealed Contribution Norms

We continue by analyzing the observed behavior of second-movers in the experimental treatments. To that respect, we identify the normative rule guiding individual contributions for each player. Therefore, we compute for each period and each subject the ratio of own contribution and average first-mover contributions. To ensure comparability between rich and poor players, we always keep the rich players' contributions in the numerator and the poor players' contributions in the denominator. That is, a poor second-mover's contribution ratio is obtained by the average contribution of the two rich first-movers divided by her own contribution. A rich second-mover's contribution ratio is obtained by her own contribution divided by the average contribution of the two poor first-movers. We thereafter average these ratios over the 15 periods to get one contribution ratio per subject. We interpret the averaged contribution ratio as the

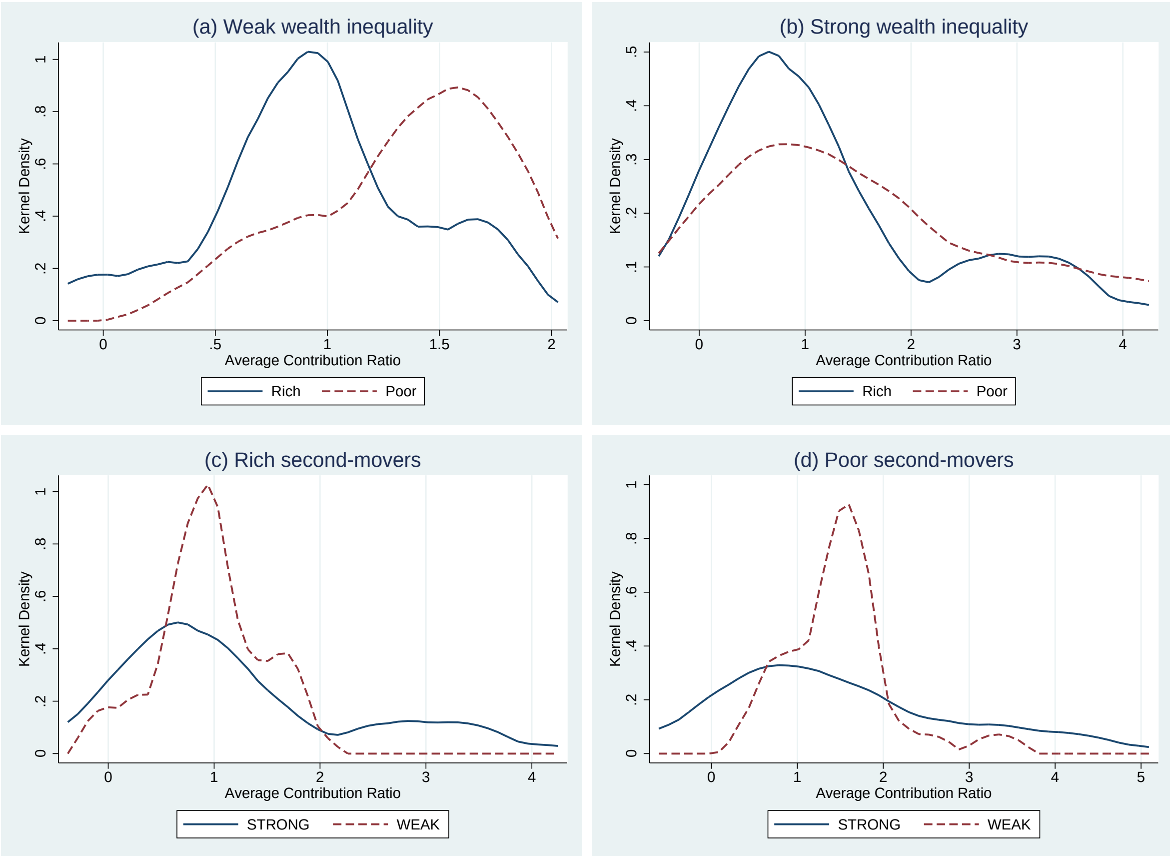


Figure 3.1: Kernel densities of average contribution ratios using the Epanechnikov kernel function. Bandwidths are calculated as to minimize the mean integrated squared error for an underlying Gaussian density.

revealed contribution norm of each subject. Specifically, a contribution ratio that equals one implies that the subject follows the equality norm. A contribution ratio that equals $\frac{25}{15} \approx 1.66$ in WEAK or $\frac{30}{10} = 3$ in STRONG, implies that the subject follows the equity norm.

Figure 3.1 presents the impact of increasing wealth inequality on the revealed contribution norms of rich and poor second-movers by plotting the kernel densities of average contribution ratios. In Figures 3.1(a) and 3.1(b), we contrast the player types' revealed contribution norms separately for WEAK and STRONG treatments. In Figures 3.1(c) and 3.1(d), we then plot the same kernel densities separately for the two player types to assess the impact of wealth inequality on the types' behavior.

We make two observations from Figures 3.1(a) and 3.1(b). First, the two types reveal diverse contribution norms under weak wealth inequality. While the peak of the kernel density for the rich players corresponds with the equality norm, the kernel density for poor players suggests the application of the equity norm, as expected from the self-serving use of fairness principles.[7] A two-sample Kolmogorov-Smirnov test for group level observations confirms that the differences in the revealed contribution norms for the two player types are significant ($D = 0.667$, $p = 0.005$). Second, under strong wealth inequality the revealed contribution norms align. We note that the peaks of both kernel densities correspond with the equality norm. We test this observation with a two-sample Kolmogorov-Smirnov test and can not reject the null hypothesis that the distribution functions are equal ($D = 0.333$, $p = 0.256$).

The evidence thus suggests that poor players alter the revealed contribution norm as wealth inequality increases. More specifically, while the behavior of the rich players seems to be driven by the self-serving interpretation of fair contribution rules in all situations, poor players under strong wealth inequality change their revealed contribution norm to the disadvantageous option. In particular, in STRONG treatments neither of the player types predominantly applies the normative principle that is considered to be fair by the impartial observers.

Figures 3.1(c) and 3.1(d) add supplementary evidence that increasing wealth inequality affects revealed contribution norms. We observe that, both for rich and for poor players, the specificity of revealed contribution norms is lower in STRONG treatments. The kernel densities for WEAK exhibit sharp in- and declines around the peaks, implying low variances. By contrast, the distributions for STRONG treatments are subject to higher variance. We compute the coefficient of variation for all treatments in order to compare the unambiguity of revealed contribution norms. We find that the coefficient is lower in WEAK (RR-W: $v = 40.6$; PP-W: $v = 50.2$) than in STRONG (RR-S: $v = 85.1$; PP-S: $v = 90.2$). We conclude that increasing wealth inequality leads to less pronounced contribution norms for both player types. This indicates that larger wealth inequality hampers the players ability to coordinate on one normative principle of fair contributions.

We sum up the main insights. First, increasing wealth inequality does not have an effect on the predominantly revealed contribution norm by rich individuals. Under both levels of initial endowments, the self-serving interpretation of fair contributions prevails and opinions of impartial observers have no impact on decisions. Second, increasing wealth inequality prompt poor individuals to change the revealed contribution norm from equity to equality. That is, while under weak wealth inequality the revealed contribution norms of the poor players align with the opinions of the impartial observers, under strong wealth inequality they diverge from both the self-serving interpretation as well as the

[7] The peaks of both distributions are slightly lower than the exact ratio proposed by the normative principle. This is consistent with previous observations that subjects undercut contributions when they are conditionally cooperating (cf. Fischbacher et al., 2001).

impartial observers' opinions. Third, the increase in wealth inequality increases the ambiguousness of revealed contribution norms for both player types.

Result 3.2 *Increasing wealth inequality leads to larger deviations from the normative principles preferred by the impartial observers. In particular, adherence to the equity principle diminishes among the (poor) individuals as the inequality increases.*

3.5.3 Impact of Increasing Wealth Inequality on Public Good Provision

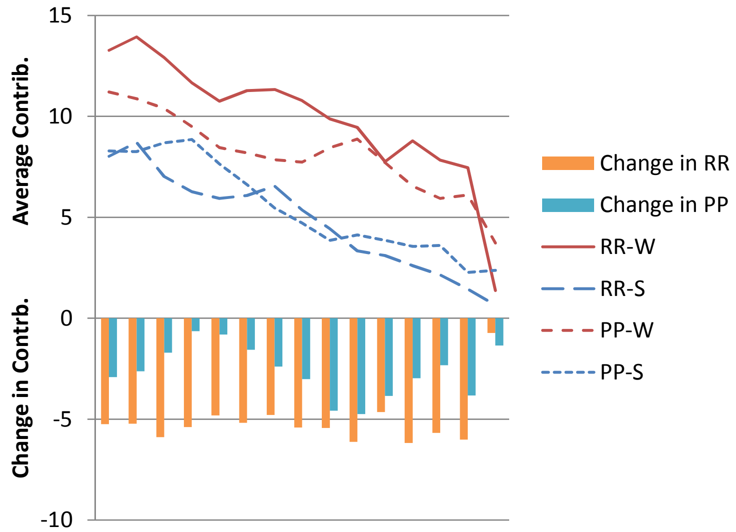


Figure 3.2: Development of average contributions per treatment and the impact of increased wealth heterogeneity, Periods 1-15

Last, we investigate how increasing wealth inequality impacts overall public good provision. Figure 3.2 illustrates the evolution of average contributions over time across treatments, and the difference in contributions between WEAK and STRONG treatments for both sequential structures. We make two observations. First, the increase in wealth inequality has a negative impact on public good provision. Contributions are higher in RR-W (9.9) than in RR-S (4.8), and higher in PP-W (8.1) than in PP-S (5.5). A Mann-Whitney test performed on the aggregated average group contribution levels (n=24) confirms that the drop in contributions between weak wealth inequality (9.0) and strong wealth inequality (5.1) is a significant ($z = 2.330$, $p = 0.02$). Second, increasing wealth inequality has unequally large effects on the alternative sequential structures. Even though both yield lower contributions in STRONG, Mann-Whitney tests at the

Table 3.4: Effects of sequential structure on contributions

Contribution	WEAK		STRONG	
PP	-3.520**	(1.457)	0.126	(1.446)
Period	-0.660***	(0.076)	-0.551***	(0.085)
PP × Period	0.230**	(0.103)	0.058	(0.110)
Constant	23.08***	(5.149)	12.67***	(4.473)
Controls	Yes		Yes	
Observations	1365		1350	
χ^2	155.5		105.0	
$Prob > \chi^2$	0.000		0.000	

Notes: Random-effects GLS regressions. RR-W is baseline in WEAK, RR-S is baseline in STRONG. Std. errors in parentheses are adjusted for individual clusters. Controls include age, a dummy variable for gender, a dummy variable for native German speaker, a dummy variable for West-German high school education and a continuous variable for the working hours per week. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

group level (n=12) show that only the effect for RR is statistically significant ($z = 2.483$, $p = 0.01$), while the effect for PP is not ($z = 0.924$, $p = 0.36$).

Result 3.3 *Increasing wealth inequality decreases voluntary public good contributions if solicited first from rich players. Increasing wealth inequality has no significant effect on the voluntary contributions if poor players contribute first.*

Next, we compare the relative performance of sequential mechanisms under increasing wealth inequality. Figure 3.2 indicates that RR generates higher contributions than PP when wealth inequality is low. This effect seems to vanish for larger wealth inequality. Mean contributions are even higher in treatment PP-S than in treatment RR-S. When performing group level Mann-Whitney tests we can not reject the null hypothesis that the samples are drawn from equal distributions. The effects are neither statistically significant for WEAK ($z = 0.520$, $p = 0.603$) nor STRONG ($z = -0.115$, $p = 0.908$). However, the regressions in Table 3.4 support the observation that there is an efficiency gain when rich players contribute first in WEAK, and no effect in STRONG.

In Figure 3.3 we present the evolution of income inequality over time for each treatment. We define income inequality as the difference in earnings between the rich and poor individuals in each group. We observe that the inequality is higher for STRONG than for WEAK. Naturally, this is due to the increase in initial wealth inequality from 10 in the WEAK to 20 in STRONG. When comparing WEAK treatments we observe that

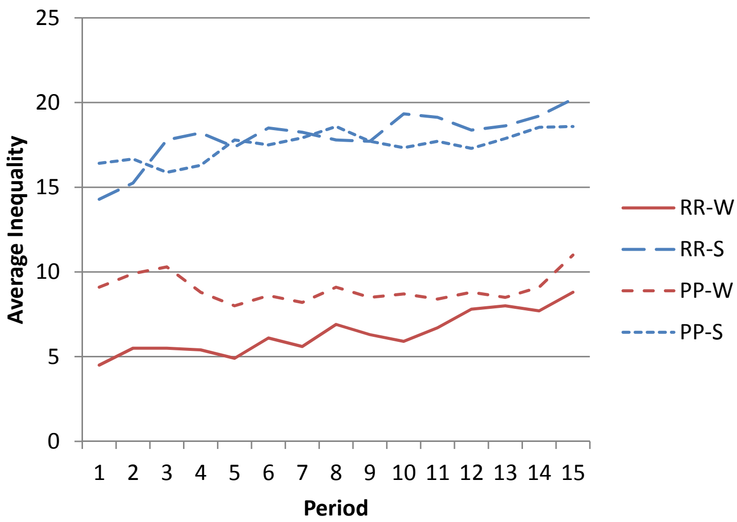


Figure 3.3: Development of average income inequality per treatment, Periods 1-15

income inequality is higher in PP-W (average difference in income inequality between rich and poor: 9.0) than in RR-W (6.4). A Mann-Whitney test on group level observations confirms that this differences is statistically significant ($z = -1.791$, $p = 0.073$). For STRONG treatments we find higher income inequality in RR-S (18.0) than in PP-S (17.5). However, this difference is not significant (Mann-Whitney test, $z = -0.520$, $p = 0.603$).

Hence, we find that under low initial wealth inequality contributions are higher and income inequality is lower if the rich players are asked to contribute first. This effect vanishes for larger initial wealth inequality. Both the rich-first and the poor-first mechanisms produce the same public good provision and result into the same final income inequality. Thus the sequence of solicited contributions is largely rendered irrelevant by the increase in wealth inequality. The observations regarding the effects of wealth inequality on the relative performance of the alternative sequential structures lead to the following result.

Result 3.4 *Increasing wealth inequality negates welfare improvements gained by the sequential contribution mechanism which solicits contributions first from wealthy individuals.*

3.6 Conclusion

The questions about what is fair and what is not have been addressed by philosophers for thousands of years. In this paper, we tackle the question on perceived and revealed fairness in the particular environment of sequential public good provision under wealth inequality. The use of experimental as well as survey data permits us to disentangle perceived fair contributions and actually revealed contribution norms. We find that people have strong inclinations to consider the equity rule to be the fairest of all. This result is robust for varying wealth distributions and for both rich and poor individuals. However, revealed contribution norms paint a different picture. While the self-serving use of fairness principles is evident under lower wealth inequality, the increase of wealth inequality dissolves any application of the commonly accepted fair contribution rule when individuals have their own money at stakes. In particular, poor second-movers switch to applying the equality norm rather than the equity norm.

Our paper also contributes to a long-standing question regarding the impact of wealth heterogeneity on public good provision. The experimental set-up enables us to examine the effect of increasing wealth heterogeneity in populations where normative contribution rules are already prevalent. By comparing the performance of sequential contribution mechanisms in treatments with low and high level of wealth heterogeneity, we provide new empirical evidence showing that increasing wealth heterogeneity decreases voluntary public good provision. At the same time, we show that increasing wealth heterogeneity hampers the relative performance of the sequential structure that solicits contributions first from the wealthier individuals.

The extrapolation of the results from our small-scale experimental investigation to the problems of growing wealth discrepancies around the world seems inappropriate. Yet, our results can offer small hints at to where the development could lead. The insight that higher asymmetry is a hindrance for cooperation between rich and poor individuals rather supports the claim for constraining the development of increasing wealth inequality.

Appendix

3.A Supplementary Tables

Table 3.5: Categorization rules when allowing for errors

	WEAK		
	Both	Rich	Poor
Equity	$c_R = \frac{5}{3}c_P \pm 1$	$[14, 16]$	$[4, 6]$
Equality	$c_R = c_P \pm 1$	$[8, 10]$	$[8, 10]$
Efficiency	$c_R + c_P \geq 38$	$[24, 25]$	$[14, 15]$
Equal Payoff	$c_R - c_P \in [9, 11]$	$[18, 20]$	$[0, 1]$
	STRONG		
	Both	Rich	Poor
Equity	$c_R = 3c_P \pm 1$	$[26, 28)$	$[2, 4]$
Equality	$c_R = c_P \pm 1$	$[8, 10]$	$[8, 9]$
Efficiency	$c_R + c_P \geq 38$	$(29, 30]$	$(9, 10]$
Equal Payoff	$c_R - c_P \in [19, 21]$	$[28, 29]$	$[0, 1]$

Chapter 4

Wealth Inequality and Fair Contribution Rules in Sequential Public Good Provision

Abstract This paper investigates the private provision of a public good in the presence of wealth inequality and sequentiality of voluntary contributions. I propose a model in which individuals value both monetary payoff and the adherence to fair contribution rules when choosing their cooperation level. Moreover, they abide by a universally effective social norm that constitutes a lower bound on cooperation. I find that wealth inequality has no impact on the contribution rule applied in the subgame perfect Nash-equilibrium if the poorer individual contributes first. Players always coordinate on equal absolute contributions, which results into inefficient provision of the public good. If the wealthy individual contributes first, coordination on a contribution rule is affected by wealth inequality and the level of the lower bound on cooperation. Players coordinate on equal relative contributions if wealth inequality and the level of the lower bound both remain below certain thresholds, resulting into efficient public good provision in equilibrium. An increase in wealth inequality hampers the ability of players to coordinate on equal relative contributions, thus resulting into inefficient public good provision.

Keywords Heterogeneous Wealth, Sequential Public Good Provision, Conditional Cooperation, Fairness Norms
JEL Classification C72, D63, H41

4.1 Introduction

A majority of people is conditionally cooperative. This is such a widely established finding in the behavioral economics literature that it has become a stylized fact.[1] If individuals collectively produce an output or can donate to a joint project benefiting everyone equally, people tend to condition their effort, time or money spent on the actions of others. Thereby, they diverge from the paradigm of free-riding usually predicted by standard preferences. In a homogenous population conditional cooperation offers simple normative guidance for individual contributions. If every individual disposes of the same amount of money to spend, the same amount of time to invest or the same abilities to put forth, and at the end everyone benefits equally from the outcome, equal time invested or equal monetary contribution is the sole fair solution. The remaining question is at which level individuals should cooperate, turning the cooperation game into a coordination game.

Everyday life is full of collective action problems with heterogeneous agents. Examples range from teamwork in sports or at the workplace to donations to charitable organizations. Coordination of cooperation becomes profoundly more difficult by asymmetry between actors. In particular wealth inequality becomes a major factor, when cooperation implies monetary contributions rather than time spent or effort put forth. Yet, if not everyone is equal, it becomes ambiguous what fair conditional cooperation is.

Distributive justice and fairness in the presence of heterogeneity has been a topic in philosophical debates for a long time and can be traced back as far as to Aristotle (1984). Two prominent approaches to fairness are the egalitarian rule and the proportional rule (Konow, 2003). While the egalitarian rule requires all individuals to be treated equally regardless of all characteristics in which they might differ, the proportional rule requires these to be taken into consideration. When being applied to wealth inequality in private public good provision, these rules translate into players cooperating on equal absolute contributions or equal relative contributions. In the economic literature, these two contribution rules have often been referred to as the equality principle and the equity principle, respectively (Reuben and Riedl, 2013; Neitzel and Sääksvuori, 2014a). Which one of these two contribution rules then becomes the social norm can differ between societies and the institutional framework.

The impact of social norms on human decision making has drawn increasing attention

[1] The arguably most influential study was conducted by Fischbacher et al. (2001). Chaudhuri (2011) provides an excellent overview of the experimental literature on conditional cooperation. Additionally, many theoretical models have been proposed to capture reciprocity and fairness considerations, e.g. Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Cox et al. (2007) and Lindroos (2008).

by the economic literature in the past decades (Elster, 1989; Ostrom, 2000; Fehr and Fischbacher, 2004). Bicchieri (2006, p.11) defines a *social norm* as a rule that fulfils two criteria. First, the rule is known to a sufficiently large subset of the population. Second, the rule is preferred to be followed by an individual if (a) she expects a sufficiently large number of others to conform to the rule and (b) she believes that a sufficiently large number of others expect her to conform to the rule. Building on this definition by Bicchieri (2006), Nikiforakis et al. (2012) define a *normative rule* as a behavioral rule that is likely to appeal to a large group of individuals within the population and constitutes a candidate of becoming a social norm. In private public goods provision with heterogeneity in wealth, the two contribution rules of equality and equity therefore translate into normative rules of fair contributions.

The self-serving application of normative fairness principles is a robust result found in multiple studies. Examples range from experimental investigations on bargaining in court trials (Babcock et al., 1995) or teacher contract negotiations (Babcock et al., 1996), to survey studies among religious groups on tithing behavior (Dahl and Ransom, 1999) and among climate negotiators on burden sharing rules in international climate policy (Lange et al., 2010).

This paper contributes to the literature on the self-serving use of fair contribution rules in public goods games under endowment heterogeneity. I propose a model in which agents value both monetary payoff and the adherence to a fair contribution rule when making their contribution choice. More specifically, agents individually choose both the contribution level and the contribution rule to follow. Deviations from the equity contribution rule and the equality contribution rule are equally costly. Additionally, I investigate the interplay of normative rules of fair contributions with a universally effective social norm which limits the individuals' contributions from below. All individuals are aware of a lowest tolerable contribution level that everyone expects all others to follow and therefore no one is willing to violate.

My work is also related to multiple theoretical studies that have examined the effect of income asymmetry on cooperation. In an early contribution, Meltzer and Richard (1981) show that an increase in the mean income relative to the median voter may result into larger public good provision if decided by the majority voting rule. Warr (1983) provides a neutrality theorem stating that the provision of a single public good is unaffected by the redistribution of income, if all individuals contribute positive levels under the original distribution. Building on these findings, Bergstrom et al. (1986) extend the result to the provision of multiple public goods.

The results of experimental studies regarding the effect of wealth inequality are rather inconclusive. While for example Chan et al. (1996, 1999) report that heterogeneous endowments increase contributions to the public good, Cherry et al. (2005) and Anderson et al. (2008) find the exact opposite. In a recent study, Neitzel and Sääksvuori (2014b) experimentally investigate the effect of increasing levels of wealth inequality on sequential public good provision. They find that increasing wealth inequality may erode existing contribution norms and thus hamper cooperation. In particular, they report the puzzling result that poor second-movers switch to applying equal absolute contributions as wealth inequality increases, departing from the self-serving interpretation of fairness.

This paper provides a model that can explain Neitzel and Sääksvuori (2014b)'s puzzling experimental findings from sequential public goods games. I show that wealth inequality affects the application of normative principles of fair contributions if the rich player has to contribute first. The analysis yields that players coordinate on equal relative contributions if the social norm bounding contributions from below and wealth inequality remain below a threshold. Then, efficient public good provision is procurable in the subgame perfect Nash-equilibrium. An increase of wealth inequality above the threshold circumvents players from coordinating on equal relative contributions and results into inefficient public good provision. If the poor individual contributes first, wealth inequality has no impact on the contribution rule applied in equilibrium. The players coordinate on equal absolute contributions, which always yields inefficient public good provision. I test the robustness of the model with an application onto the experimental set-up by Neitzel and Sääksvuori (2014b). I discover that the model qualitatively can predict all results from the study.

4.2 Economic Setting

I consider an economy in which a group of individuals interacts in order to implement a joint project. For the sake of simplicity, I consider the case where the group of individuals consists of just two players. Both players benefit equally from the provision of the project which is individually costly but globally profitable. I use the standard linear public good game to model this cooperative situation in the simplest fashion. Formally, monetary payoff of player i is given by

$$\pi_i(c_i, c_j) = w_i - c_i + \alpha(c_i + c_j),$$

where w_i is player i's endowment, $c_i \leq w_i$ is her contribution, c_j is the contribution by player $j \neq i$, and $\frac{1}{2} < \alpha < 1$ is the marginal per capita return on contributions. The two individuals differ from each other only in their initial endowment. I henceforth call the player with the larger endowment w_R the rich player R and the player with the lower endowment $w_P < w_R$ the poor player P.

In economic models contributions are usually bound by the income on the upper end and by zero contributions on the lower end. In this paper, I investigate the impact of a social norm which serves as an alternative, possibly non-negative lower bound on contributions. There is vast experimental evidence that people give away positive amounts of money even in games where no or only very limited reciprocal action by another player is possible.[2] Many alternative explanations for this behavior have been offered in the literature, from warm glow (Andreoni, 1990) to inequity aversion (Fehr and Schmidt, 1999). This model does not incorporate any of these particular theories, but I assume that there is some minimum non-negative level of cooperation which all players agree on and abide by, for whichever reason. Specifically, this is modeled by an absolute contribution level $\underline{c} \in [0, w_P)$ which is equal for both players. Thus, the rich player can choose her contribution to the public good from the interval $[\underline{c}, w_R]$ and the poor player from $[\underline{c}, w_P]$.

Besides the utility derived from monetary payoff, I assume that individuals value the adherence to a normative principle of fair contributions. The presence of wealth inequality suggests two normative principles: equal absolute contributions and equal relative contributions.[3] In accordance with the existing literature they are in the following referred to as *equality* and *equity*, respectively.

Players freely choose the normative principle they want to apply. A deviation of own contribution from the contribution demanded by the chosen normative principle leads to a deduction of utility. The farther away one's actual contribution is from the postulated normative contribution, the larger is the loss in utility. However, deviations from equality and equity induce the same amount of utility loss if violated by the same margin. The players therefore have to evaluate the utility of a combination of contribution and applied normative principle when making their contribution decision.[4] This is modeled by an

[2] Examples include Güth et al. (1982), Güth and Tietz (1990) and Forsythe et al. (1994) among many others. Engel (2011) provides a quantitative literature review of the experimental results regarding dictator games.

[3] One could also consider normative principles based on payoffs, i.e., equal payoffs or equal ratio of final payoffs as a ratio of initial endowments. However, individuals have to resolve a more complex problem to determine the contributions equalizing payoffs or the payoff ratio, while contributions can directly be transformed into own absolute and relative contributions. That is, equal absolute contributions and equal relative contributions may be determined through a direct mapping (other contribution $\rightarrow$ own contribution), while equal payoff or equal wealth ratio require indirect mapping using the payoffs (other contributions $\rightarrow$ payoffs $\rightarrow$ own contribution). Therefore, I focus on contributions.

[4] Shafir et al. (1993) develop the concept of *reason-based choice*, proposing that decision makers seek or construct reasons in order to justify their choices, may it be to themselves or to others. Spiegler

additively separable utility function with a quadratic penalty term. Thus, the utility function of individual i is given by

$$u_i(c_i, c_j, k_i) = \pi_i(c_i, c_j) - \frac{\beta}{2}\left(c_i - \left(\frac{w_i}{w_j}\right)^{k_i} c_j\right)^2,$$

where $k \in \{0, 1\}$ represents the normative principle that the player chooses to apply and $\beta > 0$ is the weight that is attached to the adherence to the normative principle. Note that only the contribution c_j by player j directly affects player i, while the choice of normative principle k_j does not. $k_i = 0$ implies that player i is using equality to justify her contributions, while $k_i = 1$ implies the use of equity. Thus, in the following player i's strategy is represented by the pair (c_i, k_i).

I study the sequential provision of the public good. Therefore, there is one first-mover and one second-mover. The second-mover observes the contribution made by the first-mover before choosing her own contribution. I assume that the first-mover anticipates the best-response of the second-mover and accounts for the second-mover's contribution when making her choice. A strategy profile of the sequential game is thus represented by $((c_i, k_i), (c_j, k_j))$, where player i is the first-mover and player j is the second-mover.

4.3 Analysis

4.3.1 Second-mover behavior

I solve the sequential game by backward induction and therefore start with the analysis of second-mover behavior. The second-mover i is faced with the maximization problem

$$\begin{aligned}
&\max_{c_i, k_i} u_i(c_i, c_j, k_i) = \pi_i(c_i, c_j) - \frac{\beta}{2}\left(c_i - \left(\frac{w_i}{w_j}\right)^{k_i} c_j\right)^2 \\
&s.t.\ c_i \in [\underline{c}, w_i] \\
&\qquad k_i \in \{0, 1\}.
\end{aligned}$$

(2002) extends the idea into economic modeling by assuming that a player needs to justify his or her strategy ex-post to a third party by a *consistent and maximally plausible belief* of the other players' strategy. In this paper, the behavior of the players may be interpreted similarly. Individuals per se have no preference for a normative principle of fair contribution, but use equity and equality as justification for their contribution choice. The choice of equity or equality itself does not have to be justified ex-post, as both represent universally acknowledged approaches to distributive justice. However, I do not explicitly model this ex-post justification of action.

It is easy to see that for the rich player contributions demanded by equality are lower than demanded by equity. The opposite applies for the poor player. Since the individuals have the monetary incentive to undercut the prescribed contribution level by a small fraction, in the interior solution it is then always the best response to contribute $c_R(c_P) = c_P - \frac{1-\alpha}{\beta}$ and choose $k_R = 0$ for the rich player, and to contribute $c_P(c_R) = \frac{w_P}{w_R}c_R - \frac{1-\alpha}{\beta}$ and choose $k_P = 1$ for the poor player. I define $x^- := x - \frac{1-\alpha}{\beta}$ and $x^+ := x + \frac{1-\alpha}{\beta}$, where x is any variable, in order to allow for a compact notation in the following.

The monetary incentive to undercut the contribution of the first-mover and the introduction of the lower bound $\underline{c} \geq 0$ give rise to corner solutions in the best-responses. Both players have to account for the fact that, for some contributions of the first-mover, they may not be able to contribute their interior utility-maximizing amount demanded by the respective normative principle. Specifically for R, P's contributions satisfying $c_P < \underline{c}^+$ do not allow R to contribute according to the interior solution for equality and contributions satisfying $c_P < \frac{w_P}{w_R}(\underline{c}^+)$ do the same for equity. Similarly for P, the rich player's contributions which satisfy $c_R < \underline{c}^+$ or $c_R < \frac{w_R}{w_P}(\underline{c}^+)$ do not allow P to choose the interior contribution level according to equality or equity, respectively. Furthermore, it is important to note that even the not undercut equity contribution $\frac{w_P}{w_R}c_R$ may be lower than $\underline{c}$. Also, contributions which do not allow P to contribute according to equality are possible if $c_R > w_P^+$. I characterize the best responses for R and P separately, beginning with the rich second-mover.

Rich second-mover. Suppose that P is the first-mover. R is then faced with a contribution level $c_P \in [\underline{c}, w_P]$. R's best responses conditional upon P's contributions are provided by the following lemma. The proofs for all lemmas and propositions are provided in Appendix 4.A.

Lemma 4.1 *If R is second-mover, then R's best response function is given by*

$$B^R(c_P) = \begin{cases} (\underline{c}, 0) & \textit{if } \underline{c} \leq c_P \leq \underline{c}^+ \\ (c_P^-, 0) & \textit{if } \underline{c}^+ < c_P \leq w_P. \end{cases}$$

That is, independent of P's contribution level R always responds with the use of $k_R = 0$. On the other hand, R's contribution is contingent on c_P. Whenever possible, R undercuts the contribution made by P by the fraction $\frac{1-\alpha}{\beta}$, and contributes at the lower bound otherwise. Note also that wealth inequality does not have any effect on R's best response.

Poor second-mover. Suppose now that R is the first-mover. P is then faced with a considerably more complex situation. As discussed before, due to wealth inequality it is possible that contributions demanded by the normative principle of equity fall below the lower bound. It follows that, in contrast to when R is second-mover, there are contribution levels for which it pays for P to apply the less preferred normative principle. In those cases contributions can even be higher than the lower bound, because the deviation from the normative principle generates larger utility deduction than the gain in monetary payoff. Also, since c_R can be higher than P's endowment, for certain levels of $\underline{c}$ there are contributions for which equity contribution, $\frac{w_P}{w_R}c_R$, is below $\underline{c}$, while equality contribution, c_R, is larger than w_P^+.

One may therefore distinguish between three different cases for the best-responses of the poor player, depending on the value of $\underline{c}$ in relation to wealth inequality. (1) A case where $\underline{c}$ is low and there is no contribution level for which it pays for P to apply equality and contribute more than $\underline{c}$. (2) A case where $\underline{c}$ is high and there are contribution levels for which equity contribution is below $\underline{c}$ and equality contribution is above w_P^+ and it pays for P to apply equality and contribute her whole endowment rather than to apply equity and contribute $\underline{c}$. (3) An intermediate case where there are contribution levels for which equity contribution is lower than $\underline{c}$ and it pays for P to apply equality and contribute more than $\underline{c}$ rather than to apply equity and contribute $\underline{c}$.

I provide the best-responses separately for the three distinct cases.

Lemma 4.2 *If P is second-mover and $\underline{c} \leq \underline{c}_{low}$, then P's best response function is given by*

$$B^P(c_R) = \begin{cases} (\underline{c}, 0) & \text{if } \underline{c} \leq c_R \leq \frac{2w_R}{w_R+w_P}\underline{c} \\ (\underline{c}, 1) & \text{if } \frac{2w_R}{w_R+w_P}\underline{c} < c_R \leq \frac{w_R}{w_P}(\underline{c}^+) \\ ((\frac{w_P}{w_R}c_R)^-, 1) & \text{if } \frac{w_R}{w_P}(\underline{c}^+) < c_R \leq w_R, \end{cases}$$

where

$$\underline{c}_{low} \equiv \frac{w_R + w_P}{w_R - w_P}\left(\frac{1-\alpha}{\beta}\right).$$

For contribution level $\underline{c}$ and levels very close to $\underline{c}$ it is obvious that equality is the best response because equity yields higher utility deductions while providing the same monetary payoff. However, an increase in R's contributions leads to the convergence of $\frac{w_P}{w_R}c_R$ to $\underline{c}$, meaning that the penalty decreases for equity and increases for equality. If $\underline{c} \leq \underline{c}_{low}$, then there already is a contribution level below $\frac{w_R}{w_P}(\underline{c}^+)$ for which it pays for P to switch to equity, because $\frac{w_P}{w_R}c_R$ is closer to $\underline{c}$ than c_R, and thus yields lower penalty while monetary payoff is equal for both. For any contribution higher than that, P is able to choose her preferred normative principle and contribute accordingly.

I continue the investigation of best-responses with the other extreme, i.e., when $\underline{c}$ is large.

Lemma 4.3 *If P is second-mover and $\underline{c} > \underline{c}_{high}$, then P's best response function is given by*

$$B^P(c_R) = \begin{cases} (\underline{c}, 0) & \text{if } \underline{c} \leq c_R \leq \underline{c}^+ \\ (c_R^-, 0) & \text{if } \underline{c}^+ < c_R \leq w_P^+ \\ (w_P, 0) & \text{if } w_P^+ < c_R \leq \tilde{c}_R \\ (\underline{c}, 1) & \text{if } \tilde{c}_R < c_R \leq \frac{w_R}{w_P}(\underline{c}^+) \\ ((\frac{w_P}{w_R} c_R)^-, 1) & \text{if } \frac{w_R}{w_P}(\underline{c}^+) < c_R \leq w_R, \end{cases}$$

where

$$\underline{c}_{high} \equiv \frac{w_P}{w_R}(w_P^+) - \frac{1-\alpha}{\beta} + \sqrt{2\frac{w_R - w_P}{w_R}\left(\frac{1-\alpha}{\beta}\right)(w_P^+)}$$

and

$$\tilde{c}_R \equiv w_R w_P \frac{w_R - \underline{c}}{w_R^2 - w_P^2} + w_R \sqrt{w_P^2 \frac{w_R - \underline{c}}{w_R^2 - w_P^2}} - (w_P^+ + \underline{c}^+).$$

In contrast to the previous case, the switch to playing the preferred normative principle does not occur before very high contributions by R, because otherwise the penalty generated by the deviation from equity is too high. That is, the gain in monetary payoff from switching to lower contributions and applying equity is more than offset by the resulting penalty. Even when R's contribution is larger than w_P^+ and thus the use of equality also generates additional utility loss through the deviation, the monetary gain is not large enough to offset the losses generated by equity.

Last, I consider intermediate lower bounds.

Lemma 4.4 *If P is second-mover and $\underline{c}_{low} < \underline{c} \leq \underline{c}_{high}$, then P's best response function is given by*

$$B^P(c_R) = \begin{cases} (\underline{c}, 0) & \text{if } \underline{c} \leq c_R \leq \underline{c}^+ \\ (c_R^-, 0) & \text{if } \underline{c}^+ < c_R \leq \hat{c}_R \\ (\underline{c}, 1) & \text{if } \hat{c}_R < c_R \leq \frac{w_R}{w_P}(\underline{c}^+) \\ ((\frac{w_P}{w_R} c_R)^-, 1) & \text{if } \frac{w_R}{w_P}(\underline{c}^+) < c_R \leq w_R, \end{cases}$$

where

$$\hat{c}_R \equiv \frac{w_R}{w_P}\left(\underline{c} + \frac{w_R}{w_P}\left(\frac{1-\alpha}{\beta}\right) - \sqrt{\frac{w_R - w_P}{w_P}\left(\frac{1-\alpha}{\beta}\right)\left(\frac{w_R + w_P}{w_P}\left(\frac{1-\alpha}{\beta}\right) + 2\underline{c}\right)}\right).$$

For an intermediate lower bound the switch from the use of equality to equity occurs within the interval of contributions where $c_R > \underline{c}^+$ and $c_R \leq \frac{w_R}{w_P}(\underline{c}^+)$, i.e. where equality contribution is fully feasible while equity contribution is not.

4.3.2 First-mover behavior and equilibria

Applying backward induction, I use the previously obtained best responses by R and P in order to determine the optimal strategy of the respective first-mover. The contribution level and applied normative principle of the first-mover then fully defines the equilibrium. As before, I begin the analysis with the poor first-mover.

Poor first-mover. Let P be the first-mover, R the second-mover and R's best responses as stated in Lemma 4.1. The analysis yields the following sub-game perfect Nash equilibria (SPNE) for the game.

Proposition 4.1 *If P is the first-mover, then the unique SPNE is $((\underline{c}, 0), (\underline{c}, 0))$ if $\beta < \beta_1$ and $\left((w_P, 0), \left(w_P^-, 0\right)\right)$ if $\beta > \beta_1$, where $\beta_1 \equiv \frac{1-\alpha^2}{2(2\alpha-1)(w_P-\underline{c})}$. If $\beta = \beta_1$, then both $((\underline{c}, 0), (\underline{c}, 0))$ and $\left((w_P, 0), \left(w_P^-, 0\right)\right)$ are SPNE.*

The intuition behind this proposition is straightforward. The rich second-mover can always match the poor first-mover's contribution, because both players are bound by the same minimum absolute cooperation level and R has a larger endowment than P. In particular, this means that contributions demanded by the equality principle are always possible and, because R prefers equality over equity, it will be applied for all contributions made by P. It follows that total contributions are maximally twice the endowment of the poor player and that fully efficient provision of the public good will never be reached under this sequential structure.

Also, P knows that she will suffer from deviation penalties because R always contributes below her contribution level. The amount that R undercuts is equal everywhere except for contributions close to the lower bound. Thus, if it is profitable for the poor player to contribute anything higher than the lower bound, then it must be profitable to fully contribute. P only has an incentive to choose low contributions if the amount of undercutting $(\frac{1-\alpha}{\beta})$ is comparatively large. That is the case if β is small and the players thus value monetary payoff a lot more than deviations from the normative principle. Observe also that the value of β_1, which marks the switch to the higher contribution equilibrium, decreases for decreasing $\underline{c}$. This property results from the increase in the difference in the monetary payoffs between contribution levels $\underline{c}$ and w_P. Particularly, the proposition also comprises of a lower bound of $\underline{c} = 0$. In that case β_1 is minimized, but still larger than zero.

Rich first-mover. Suppose now that R is the first-mover and P second-mover. I separately characterize the equilibria for the sequential game subject to the value of the lower bound, starting with $\underline{c} < \underline{c}_{low}$.

Proposition 4.2 *If R is the first-mover and $\underline{c} \leq \underline{c}_{low}$, then the unique SPNE of the sequential game is $((\underline{c}, 0), (\underline{c}, 0))$ if $\beta < \beta_2$ or $\alpha \leq \frac{w_R}{w_R + w_P}$, and $\left((w_R, 1), \left(w_P^-, 1\right)\right)$ if $\beta > \beta_2$ and $\alpha > \frac{w_R}{w_R+w_P}$, where $\beta_2 \equiv \frac{(1-\alpha)\left(\alpha + \left(\frac{w_R}{w_P}\right)^2 \frac{1-\alpha}{2}\right)}{(w_R + w_P - 2\underline{c})\alpha - (w_R - \underline{c})}$. If $\beta = \beta_2$ and $\alpha > \frac{w_R}{w_R+w_P}$, then both $((\underline{c}, 0), (\underline{c}, 0))$ and $\left((w_R, 1), \left(w_P^-, 1\right)\right)$ are SPNE.*

In equilibrium R and P either both contribute their whole endowment or close to their endowment, or both contribute exactly at the lower bound of contributions $\underline{c}$. Correspondingly, there is the possibility that the normative principle of equity is played in equilibrium and total monetary payoff is nearly maximized. If the rich player's share of total wealth is greater than or equal to α, then equity can never be played in equilibrium, because R would suffer monetary losses from cooperation. If on the other hand the rich player's share of total wealth is lower than α, then β determines whether equity or equality is played in equilibrium. For aversion parameters below the threshold β_2, then P's undercutting of the normative principle is too large and prevents R from contributing her whole endowment. If $\beta > \beta_2$, the strategy involving full contributions prevails. Observe also that, just as β_1 above, β_2 decreases with a decrease in $\underline{c}$, because the difference between $\underline{c}$ and w_R increases, leading to higher monetary gain for R when changing the contribution level from $\underline{c}$ to w_R.

Next, I characterize the equilibria for a high-valued lower bound on contributions.

Proposition 4.3 *If R is the first-mover and $\underline{c} > \underline{c}_{high}$, then the unique SPNE of the sequential game is $((\underline{c}, 0), (\underline{c}, 0))$ if $\beta < \beta_3$ and $\left(\left(w_P^+, 0\right), (w_P, 0)\right)$ if $\beta > \beta_3$, where $\beta_3 \equiv \frac{3(1-\alpha)^2}{2(w_P - \underline{c})(2\alpha - 1)}$. If $\beta = \beta_3$, then both $((\underline{c}, 0), (\underline{c}, 0))$ and $\left(\left(w_P^+, 0\right), (w_P, 0)\right)$ are SPNE.*

Again, the intuition is simple. The equilibrium in which both players apply equity has vanished, because the rich player may now extract full contributions by the poor player without having to contribute fully herself. In other words, R can force P into playing the less preferred normative principle. On the other hand, if R has to contribute a lot more than P because β is small and therefore the amount of undercutting is large, then she is more inclined to contribute at the lower bound $\underline{c}$. Again observe that β_3 increases in $\underline{c}$ by the same arguments as above.

Last, consider the intermediate case when $\underline{c} > \underline{c}_{low}$ but $\underline{c} < \underline{c}_{high}$. Here, due to the emergence of a third possible equilibrium, the results are not as clear cut as before.

Proposition 4.4 *(i) If R is the first-mover, $\underline{c}_{low} < \underline{c} \leq \underline{c}_{high}$ and $\alpha \leq \frac{w_R}{w_R+w_P}$, then the unique SPNE of the sequential game is $((\underline{c},0),(\underline{c},0))$, if $\beta < \beta_4$, and $\left((\hat{c}_R,0),\left(\hat{c}_R^-,0\right)\right)$, if $\beta > \beta_4$, where $\beta_4 \equiv \frac{2w_R(1-\alpha)\sqrt{(2-\alpha)(2\alpha-1)}+w_P\left(1-\alpha^2\right)}{2\underline{c}(w_R-w_P)(2\alpha-1)}$. If $\beta = \beta_4$, then both $((\underline{c},0),(\underline{c},0))$ and $\left((\hat{c}_R,0),\left(\hat{c}_R^-,0\right)\right)$ are SPNE.*

(ii) If R is the first-mover, $\underline{c}_{low} < \underline{c} \leq \underline{c}_{high}$ and $\alpha > \frac{w_R}{w_R+w_P}$, then the unique SPNE of the sequential game is $((\underline{c},0),(\underline{c},0))$, if $\beta < \min\{\beta_2,\beta_4\}$. If $\beta = \min\{\beta_2,\beta_4\} = \beta_2$, then both $((\underline{c},0),(\underline{c},0))$ and $\left((w_R,1),\left(w_P^-,1\right)\right)$ are SPNE, and if $\beta = \min\{\beta_2,\beta_4\} = \beta_4$, then both $((\underline{c},0),(\underline{c},0))$ and $\left((\hat{c}_R,0),\left(\hat{c}_R^-,0\right)\right)$ are SPNE. If $\beta > \min\{\beta_2,\beta_4\}$, then either $\left((\hat{c}_R,0),\left(\hat{c}_R^-,0\right)\right)$ is unique SPNE or $\left((w_R,1),\left(w_P^-,1\right)\right)$ is unique SPNE or both are SPNE.

For intermediate lower bounds there are three equilibrium candidates. In addition to the two equilibria already found in Proposition 4.2, there is an equilibrium in which both players apply the normative principle of equality, but contribute more than $\underline{c}$ and less than w_P^+. It originates from the rich player's option to choose a contribution which renders the use of equity too costly for P. Under these conditions, R can thus again force P into playing her less preferred contribution principle.

If cooperation on grounds of equity is not profitable for the rich player, hence, if $\alpha \leq \frac{w_R}{w_R+w_P}$, then R will only choose between the two options which follow the equality principle. For small β, this leads to equilibrium contributions of $\underline{c}$ by both players. However, the costlier the deviation from the normative principle is for the players, the smaller will P's deviation be and thus the higher R's utility for contributing $\hat{c}_R$. Eventually, for all $\beta > \beta_4$, $\hat{c}_R$ is the equilibrium contribution. It is straightforward to see that β_4 decreases both in $\underline{c}$ and wealth inequality, thus resulting into either increasing or constant public good provision for a given β.

For $\alpha > \frac{w_R}{w_R+w_P}$ the analysis is not as straightforward. I distinguish between two cases. If β is lower than the minimum of β_2 and β_4, then obviously $\underline{c}$ is the sole equilibrium contribution. If β is larger than either β_2 or β_4, or even larger than both, the equilibrium strategy additionally depends on wealth inequality and the level of $\underline{c}$. In this case, I refrain from calculating the exact values of β for which the respective equilibria are played, because, due to the complexity of the model, they offer no additional insights. Instead, I perform a comparative statics analysis on the effects of wealth inequality and the level of the lower bound on the resulting equilibrium. However, since β is larger than β_2 or β_4 or both, only $(\hat{c}_R,0)$ and $(w_R,1)$ can be equilibrium strategies. The indirect utility functions needed for the analysis are provided in the appendix at the end of the proof of Proposition 4.4.

I make two observations. First, increases in wealth inequality and the lower bound have similar effects on the equilibrium, as both increase the appeal of strategy $(\hat{c}_R, 0)$ compared to strategy $(w_R, 1)$. Second, both $u_R(\hat{c}_R, \hat{c}_R^-, 0)$ and $u_R(w_R, w_P^-, 1)$ increase in β, approaching a limit.

The first observation follows from the fact that $u_R(w_R, w_P^-, 1)$ is independent of $\underline{c}$, while $u_R(\hat{c}_R, \hat{c}_R^-, 0)$ increases. Furthermore, an increase in wealth inequality, while keeping total wealth constant, increases $\hat{c}_R$ and therefore also $u_R(\hat{c}_R, \hat{c}_R^-, 0)$.[5] However, $u_R(w_R, w_P^-, 1)$ decreases when wealth inequality is increased. As a result, β_2 increases in both wealth inequality and the lower bound, while β_4 decreases in both.

Regarding the second observation, I compare the utilities at the limit. When $\beta \to \infty$, then $u_R(\hat{c}_R, \hat{c}_R^-, 0) > u_R(w_R, w_P^-, 1)$ if and only if

$$\underline{c} > w_P \left(\frac{(w_R + w_P)\alpha - w_R}{2w_R\alpha - w_R} \right).$$

The right-hand side is always less than w_P. Therefore, there always exists a $\underline{c}$ for which the condition is satisfied. Furthermore, the higher wealth inequality, the lower is the threshold. This result further supports the interdependency of wealth inequality and lower bound.

Taken together, these observations indicate that $(w_R, 1)$ is the equilibrium strategy for all $\beta > \beta_2$, if wealth inequality and lower bound are sufficiently low. On the other hand, $(\hat{c}_R, 0)$ is the equilibrium strategy for all $\beta > \beta_4$, if wealth inequality and lower bound are sufficiently high.

Figure 4.1 exemplarily illustrates the effects of wealth inequality, the lower bound and β on the equilibrium. β is plotted on the abscissa while utility and $\underline{c}$ are plotted on the ordinate. The three lines at the top of each graph represent the three utility values $u_R(\underline{c}, \underline{c}, 0)$ (continuous), $u_R(\hat{c}_R, \hat{c}_R^-, 0)$ (dashed) and $u_R(w_R, w_P^-, 1)$ (dotted). The upper ridge of the three lines thus constitutes R's equilibrium utility for the respective value of β. The two dashed-dotted lines at the bottom of each graph represent $\underline{c}_{low}$ (lower line) and $\underline{c}_{high}$ (upper line), while the continuous line represents $\underline{c}$. That is, for every β where the continuous line is between the two dashed-dotted lines, the condition $\underline{c} \in (\underline{c}_{low}, \underline{c}_{high}]$ is satisfied and thus $\underline{c}$ constitutes an intermediate lower bound. The level $\underline{c}$ is increased from left to right and wealth inequality is increased from top to bottom.

[5] I only consider constant wealth in order to focus on the distribution of wealth instead of possible effects of growth.

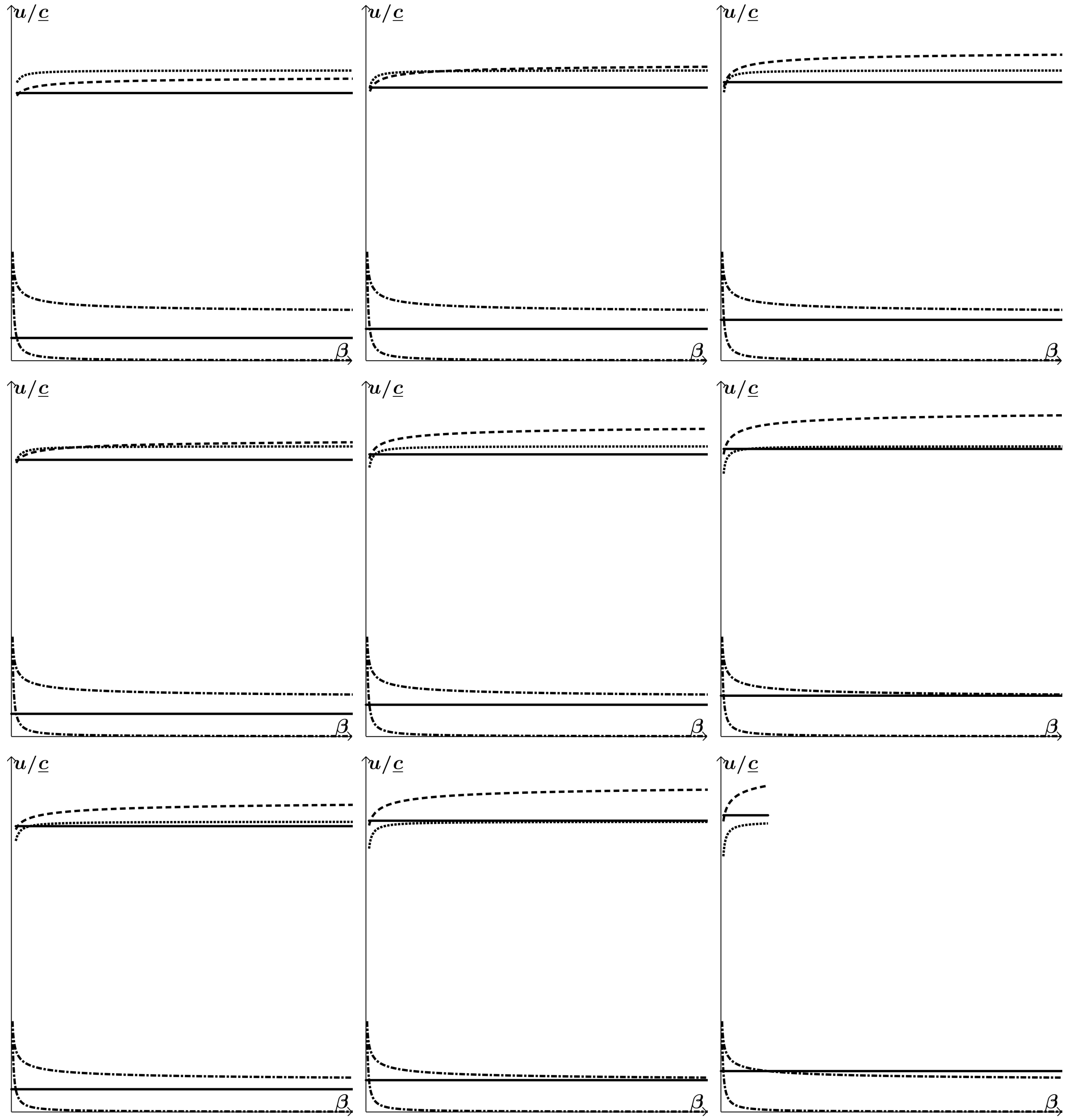


Figure 4.1: The effects of the increase in wealth inequality, $\underline{c}$ and β on the equilibrium for intermediate lower bounds. The three lines at the top of each graph represent the three utility values $u_R(\underline{c}, \underline{c}, 0)$ (continuous), $u_R(\hat{c}_R, \hat{c}_R^-, 0)$ (dashed) and $u_R(w_R, w_P^-, 1)$ (dotted). The two dashed-dotted lines at the bottom of each graph represent $\underline{c}_{low}$ and $\underline{c}_{high}$, while the continuous line represents $\underline{c}$. The level of the lower bound is increased from left to right and wealth inequality from top to bottom.

As described above, we observe that an increase in wealth inequality shifts the curves $u_R(\underline{c}, \underline{c}, 0)$ and $u_R(\hat{c}_R, \hat{c}_R^-, 0)$ upwards while $u_R(w_R, w_P^-, 1)$ is slightly shifted down. Similarly, an increase in $\underline{c}$ shifts $u_R(\underline{c}, \underline{c}, 0)$ and $u_R(\hat{c}_R, \hat{c}_R^-, 0)$ upwards while $u_R(w_R, w_P^-, 1)$ remains unchanged. Resulting thereof, low wealth inequality and a low lower bound yield the equilibrium $((w_R, 1), (w_P^-, 1))$ for all $\beta > \beta_2$, and high wealth inequality and a high lower bound yield the equilibrium $((\hat{c}_R, 0), (\hat{c}_R^-, 0))$ for all $\beta > \beta_4$. For intermediate levels of wealth inequality and lower bound, there is a β_5 for which the equilibrium strategy of the rich player switches from $(w_R, 1)$ to $(\hat{c}_R, 0)$.

I summarize the main findings of this section as follows. If P is the first-mover, then the normative principle of equality is always applied by both P and R and thus efficient provision of the public good is impossible. If R is the first-mover, then (almost) efficient provision is possible if the rich player's share of total wealth is less than the individual return from the public good and the aversion against breaking the normative principle is sufficiently large. Otherwise, equity is never played in equilibrium and efficient provision is again impossible. But even if the rich player's share of the total wealth is less than the individual return on investment, there exists a level of wealth inequality and a level of the lower bound such that for every higher level, equity is never played in equilibrium and efficient public good provision is never procurable.

4.4 Application to Data

Neitzel and Sääksvuori (2014b) implement a laboratory experiment to investigate the impact of wealth inequality in a four-player sequential voluntary contribution mechanism with two first-movers and two second-movers. In each group there are two rich players and two poor players and the distribution of endowments is determined through a real-effort tournament task in advance of the contribution game. The authors then compare the effects of the increase of the level wealth inequality on the individual behavior and the resulting public good provision in sequential structures where either both rich players or both poor players contribute first.

They find that, if the poor players have to choose their contributions first, the use of equality is observed under both wealth distributions. On the other hand, if the rich players have to choose their contributions first, the use of equity is observed for the lower wealth inequality treatment, while the use of equality is observed for the higher wealth inequality treatment. As a result, contributions decrease for both sequential structures if wealth inequality is increased. They also find that under low wealth inequality total contributions are higher when solicited first from the rich players. However, this effect vanishes as wealth inequality is increased.

In order to test the accuracy of the model, I apply it to the wealth distributions used in the experiment by Neitzel and Sääksvuori (2014b). Since both β and $\underline{c}$ are unobservable, I compute the equilibria and total contributions for various combinations of β and $\underline{c}$. Tables 4.1(a) and 4.1(b) provide excerpts of the calculations for the poor first-mover and rich first-mover mechanisms, respectively.[6]

I observe that the model can correctly predict the experimental results by Neitzel and Sääksvuori (2014b). In particular, for a large set of parameter combinations it reproduces the switch in applied contribution norm from equity to equality when the rich player contributes first. Furthermore, while the model cannot correctly predict the overall contribution level from the experiments, it can however predict the correct ordering across sequential structure and level of wealth inequality. For example, for $\underline{c} = 5$ and $\beta = 0.5$, contributions for both sequential structures are lower under high wealth inequality than under low wealth inequality. While under low wealth inequality contributions are significantly higher if the rich player contributes first, the difference is minimal under high wealth inequality.

Altogether, taking the model to the data adds further support to the underlying assumptions. However, the results have to be weighted carefully. Since the experiments were designed for four players and the model only accounts for two players, the incentive structure needed to be adjusted for the calculations. Specifically, the calculations used a marginal per capita return of $\alpha = 0.8$, instead of $\alpha = 0.4$ used in the experiment. I am not aware of any well-defined rules for adjusting social dilemma games so as to keep incentives equal across different group sizes. As a robustness check, I therefore ran the same calculations for $\alpha = 0.7$. I observe similar results for adjusted combinations of $\underline{c}$ and β. I therefore conclude that the qualitative results are robust against small variations in α. The tables for $\alpha = 0.7$ are shown in Appendix 4.B.

[6] The values in the tables were chosen as to provide the most conclusive overview of the observed patterns. In particular, $\underline{c} = 4$ was chosen because the structure of equilibria is the same as for $\underline{c} = 2$ and $\underline{c} = 3$. Similarly for $\underline{c} = 5$ and the equilibria for $\underline{c} = 6$ and $\underline{c} = 7$.

Table 4.1: Calculations of SPNE and total contributions, $\alpha = 0.8$

(a) Poor first-mover

	$\underline{c}$	1		4		5		8	
β		EQ	C	EQ	C	EQ	C	EQ	C
0.1	low	w_P	28	w_P	28	w_P	28	w_P	28
	high	w_P	18	w_P	18	w_P	18	$\underline{c}$	16
0.2	low	w_P	29	w_P	29	w_P	29	w_P	29
	high	w_P	19	w_P	19	w_P	19	w_P	19
0.5	low	w_P	29.6	w_P	29.6	w_P	29.6	w_P	29.6
	high	w_P	19.6	w_P	19.6	w_P	19.6	w_P	19.6
1	low	w_P	29.8	w_P	29.8	w_P	29.8	w_P	29.8
	high	w_P	19.8	w_P	19.8	w_P	19.8	w_P	19.8
1.5	low	w_P	29.9	w_P	29.9	w_P	29.9	w_P	29.9
	high	w_P	19.9	w_P	19.9	w_P	19.9	w_P	19.9
2	low	w_P	29.9	w_P	29.9	w_P	29.9	w_P	29.9
	high	w_P	19.9	w_P	19.9	w_P	19.9	w_P	19.9

(b) Rich first-mover

	$\underline{c}$	1		4		5		8	
β		EQ	C	EQ	C	EQ	C	EQ	C
0.1	low	$\underline{c}$	2	$\underline{c}$	8	$\underline{c}$	10	$\underline{c}$	16
	high	$\underline{c}$	2	$\underline{c}$	8	$\underline{c}$	10	w_P^+	22
0.2	low	w_R	39	w_R	39	w_R	39	$\hat{c}_R$	19.5
	high	$\underline{c}$	2	$\hat{c}_R$	11.6	$\hat{c}_R$	15.3	w_P^+	21
0.5	low	w_R	39.6	w_R	39.6	w_R	39.6	w_R	39.6
	high	$\underline{c}$	2	$\hat{c}_R$	14.2	$\hat{c}_R$	18.5	w_P^+	20.4
1	low	w_R	39.8	w_R	39.8	w_R	39.8	w_R	39.8
	high	w_R	39.8	$\hat{c}_R$	16.1	w_P^+	20.2	w_P^+	20.2
1.5	low	w_R	39.9	w_R	39.9	w_R	39.9	$\hat{c}_R$	23.3
	high	w_R	39.9	$\hat{c}_R$	17.2	w_P^+	20.1	w_P^+	20.1
2	low	w_R	39.9	w_R	39.9	w_R	39.9	$\hat{c}_R$	23.7
	high	w_R	39.9	$\hat{c}_R$	17.9	w_P^+	20.1	w_P^+	20.1

Notes: The columns labeled EQ comprise of the predicted equilibria, where $\underline{c}$, w_P, w_R, $\hat{c}_R$ and w_P^+ represent $((\underline{c}, 0), (\underline{c}, 0))$, $((w_P, 0), (w_P^-, 0))$, $((w_R, 1), (w_R^-, 1))$, $((\hat{c}_R, 0), (\hat{c}_R^-, 0))$ and $((w_P^+, 0), (w_P, 0))$, respectively. The columns labeled C comprise of the total contributions in the predicted equilibrium. Rows labeled *low* present the predictions for wealth distribution $w_R = 25$ and $w_P = 15$, while rows labeled *high* exhibit the predictions for wealth distribution $w_R = 30$ and $w_P = 10$.

4.5 Concluding Remarks

Human cooperation is put to test every time when individual and collective interests conflict. It becomes even more difficult when heterogeneity between actors comes into play. There is ample evidence that people apply normative principles for cooperation in a self-serving way, causing an additional conflict of interest. In the case of public good provision under wealth inequality, poor and rich players disagree about whether cooperation should be based on equal absolute or equal relative contributions, with poor players preferring relative contributions and rich players absolute contributions.

In this paper I provide a model of sequential public good provision that investigates the interplay of the self-serving use of normative principles of fair contributions and a universally effective social norm that bounds absolute contributions from below. I show that full cooperation is supported if and only if the rich player moves first and wealth inequality is sufficiently low. Higher wealth inequality has a negative impact on the provision of the public good, regardless of the sequence of play. If the poor player moves first, then the contributions of either individual never exceed the endowment of the poor player. If the rich player moves first, then low inequality and a small lower bound on contributions facilitate cooperation on equal relative contributions and enable efficient provision of the public good, but higher levels of wealth inequality or a higher value of the lower bound on contribution hamper cooperation and decrease public good provision.

I furthermore find that a higher lower bound on contribution may have a negative effect on public good provision, if the rich player is first-mover. At first sight this seems counterintuitive, as an increase in the lower bound usually should be associated with increasing contributions because the minimal cooperation level is increased. In could be interpreted similarly to the crowding out of prosocial behavior often observed in other settings (see Frey (1997) and Frey and Jegen (2001) for literature reviews on the topic). That is, a high lower bound on cooperation has the same effect as monetary rewards for blood donations. The extrinsic motivation of the social norm then leads to the erosion of the intrinsic motivation to apply the normative principle that maximizes total payoff. On the other hand, whenever the poor player contributes first, the increase of the lower bound always has a positive effect on public good provision and crowding out is not observed.

The results derived in this paper are substantially influenced by the introduction of the social norm that serves as a lower bound on absolute contributions. There are of course many more ways of defining a lower bound on contributions and any alternative formulation would very likely also change the resulting equilibria. For example, the obvious alternative formulation of a relative contribution lower bound, i.e., a minimal percentage contribution of initial endowments, presumably causes an opposite effect on the individuals' strategic options. In particular, the poor first-mover could choose a contribution which renders cooperation based on equal absolute contributions impossible for the rich second-mover, and thus force her to apply equal relative contributions. The exact equilibria and the question whether this alternative formulation of the model better represents human interaction remains open for future research.

Appendix

4.A Proofs

Proof of Lemma 4.1:

Let $\frac{w_P}{w_R}\left(\underline{c}+\frac{1-\alpha}{\beta}\right) \geq \underline{c}$ and $c_P \in \left[\underline{c}, \frac{w_P}{w_R}\left(\underline{c}+\frac{1-\alpha}{\beta}\right)\right]$. Then for both $k_R = 0$ and $k_R = 1$ it is optimal for R to contribute $c_R = \underline{c}$ and the monetary payoff is equal. However, since $c_P < \frac{w_R}{w_P}c_P$, the penalty from undercutting the prescribed normative principle is lower for $k_R = 0$ than for $k_R = 1$. Let now $c_P \in \left(\frac{w_P}{w_R}(\underline{c}+\frac{1-\alpha}{\beta}), \underline{c}+\frac{1-\alpha}{\beta}\right]$ or, if $\frac{w_P}{w_R}\left(\underline{c}+\frac{1-\alpha}{\beta}\right) < \underline{c}$, $c_P \in \left[\underline{c}, \underline{c}+\frac{1-\alpha}{\beta}\right]$. In both cases $k_R = 0$ again requires contributions of $c_R = \underline{c}$, while the contributions are $c_R = \frac{w_R}{w_P}c_P - \frac{1-\alpha}{\beta}$ for $k_R = 1$. Hence, the monetary payoff is larger for $k_R = 0$ (since $\underline{c} < \frac{w_R}{w_P}c_P - \frac{1-\alpha}{\beta}$) and the penalty from undercutting the prescribed normative principle is less or equal, because $c_P - \underline{c} \leq \frac{1-\alpha}{\beta}$. Therefore, if follows that $(\underline{c}, 0)$ is the best response if $c_P \leq \underline{c}+\frac{1-\alpha}{\beta}$.

If $c_P > \underline{c}+\frac{1-\alpha}{\beta}$, then both normative principles produce undercutting of P's contribution and relative contribution, respectively, by $\frac{1-\alpha}{\beta}$. Hence the penalty is equal for both, but the monetary payoff is larger for $k_R = 0$. It follows that $\left(c_P - \frac{1-\alpha}{\beta}, 0\right)$ is the best response. $\square$

Proof of Lemma 4.2:

Let $c_R < \underline{c}+\frac{1-\alpha}{\beta}$ and thus also $c_R < \frac{w_R}{w_P}\left(\underline{c}+\frac{1-\alpha}{\beta}\right)$. Then P chooses $k_P = 1$ if and only if $u(\underline{c}, c_R, 1) > u(\underline{c}, c_R, 0)$. Solving for c_R yields $c_R > \frac{2w_R}{w_R+w_P}\underline{c}$. Vice versa, if $c_R \leq \frac{2w_R}{w_R+w_P}$, $k_P = 0$ yields higher utility. Testing if $\frac{2w_R}{w_R+w_P}\underline{c} \leq \underline{c}+\frac{1-\alpha}{\beta}$ leads to the condition that $\underline{c} \leq \frac{w_R+w_P}{w_R-w_P}\left(\frac{1-\alpha}{\beta}\right)$. For $c_R > \frac{w_R}{w_P}\left(\underline{c}+\frac{1-\alpha}{\beta}\right)$, it is then obvious that P's best response is to choose $k_P = 1$ and $c_P = \frac{w_P}{w_R}c_R - \frac{1-\alpha}{\beta}$ because monetary payoff is higher and penalty is equal if $k_P = 0$. $\square$

Proof of Lemma 4.3:

Let $\underline{c}$ and c_R be such that $\frac{w_P}{w_R}c_R < \underline{c} \wedge c_R - \frac{1-\alpha}{\beta} > w_P$. P chooses to play $(w_P, 0)$ instead of $(\underline{c}, 1)$ if and only if

$$
\begin{aligned}
& & u(\underline{c}, c_R, 1) &\leq u(w_P, c_R, 0) \\
&\Leftrightarrow & \pi_P(\underline{c}, c_R) - \frac{\beta}{2}\left(\underline{c} - \frac{w_R}{w_P}c_R\right)^2 &\leq \pi_P(w_P, c_R) - \frac{\beta}{2}\left(w_P - c_R\right)^2 \\
&\Leftrightarrow & \pi_P(\underline{c}, c_R) - \pi_P(w_P, c_R) &\leq \frac{\beta}{2}\left(\left(\underline{c} - \frac{w_R}{w_P}c_R\right)^2 - \left(w_P - c_R\right)^2\right) \\
&\Leftrightarrow & (1-\alpha)\left(w_p - \underline{c}\right) &\leq \frac{\beta}{2}\left(\left(\underline{c} - \frac{w_R}{w_P}c_R\right)^2 - \left(w_P - c_R\right)^2\right).
\end{aligned}
$$

In particular, by using $c_R = w_P + \frac{1-\alpha}{\beta}$ and solving for $\underline{c}$ we determine the lower limit for $\underline{c}$ such that $u(\underline{c}, c_R, 1) \leq u(w_P, c_R, 0)$ can even hold:

$$
\underline{c} > \frac{w_P}{w_R}\left(w_P + \frac{1-\alpha}{\beta}\right) - \frac{1-\alpha}{\beta} + \sqrt{2\frac{w_R - w_P}{w_R}\left(\frac{1-\alpha}{\beta}\right)\left(w_P + \frac{1-\alpha}{\beta}\right)}.
$$

Furthermore, solving the inequality $u(\underline{c}, c_R, 1) \leq u(w_P, c_R, 0)$ for $c_R > w_P + \frac{1-\alpha}{\beta}$ yields

$$
c_R \leq w_R w_P \frac{w_R - \underline{c}}{w_R^2 - w_P^2} + w_R\sqrt{w_P^2\frac{w_R - \underline{c}}{w_R^2 - w_P^2} - \left(w_P + \underline{c} + 2\frac{1-\alpha}{\beta}\right)}.
$$

Denote this threshold by $\tilde{c}_R$

It is then trivial that P applies $k_P = 0$ for all contributions $c_R \leq \tilde{c}_R$ and accordingly contributes either $\underline{c}$ if $c_R \leq \underline{c} + \frac{1-\alpha}{\beta}$, or $c_R - \frac{1-\alpha}{\beta}$ if c_R is within the interval $\left(\underline{c} + \frac{1-\alpha}{\beta}, w_P + \frac{1-\alpha}{\beta}\right]$, or w_P if $c_R > w_P + \frac{1-\alpha}{\beta}$. For all contributions $c_R > \tilde{c}_R$ P responds with $k_P = 1$ and contributes either $\underline{c}$ if $c_R \leq \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$ or $\frac{w_P}{w_R}c_R - \frac{1-\alpha}{\beta}$ if $c_R > \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$. □

Proof of Lemma 4.4:

Let $\underline{c}_{low} < \underline{c} \leq \underline{c}_{high}$ and $c_R < \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$. Then P chooses to play $(\underline{c}, 1)$ instead of $\left(c_R - \frac{1-\alpha}{\beta}, 1\right)$ if

$$\begin{aligned}
& u(\underline{c}, c_R, 1) > u(c_R - \frac{1-\alpha}{\beta}, c_R, 0) \\
\Leftrightarrow \quad & \pi_P(\underline{c}, c_R) - \frac{\beta}{2}\left(\underline{c} - \frac{w_P}{w_R}c_R\right)^2 > \pi_P(c_R - \frac{1-\alpha}{\beta}, c_R) - \frac{\beta}{2}\left(\frac{1-\alpha}{\beta}\right)^2 \\
\Leftrightarrow \quad & \pi_P(\underline{c}, c_R) - \pi_P(c_R - \frac{1-\alpha}{\beta}, c_R) > \frac{\beta}{2}\left(\left(\underline{c} - \frac{w_P}{w_R}c_R\right)^2 - \left(\frac{1-\alpha}{\beta}\right)^2\right) \\
\Leftrightarrow \quad & (1-\alpha)(c_R - \frac{1-\alpha}{\beta} - \underline{c}) > \frac{\beta}{2}\left(\left(\underline{c} - \frac{w_P}{w_R}c_R\right)^2 - \left(\frac{1-\alpha}{\beta}\right)^2\right)
\end{aligned}$$

Solving the inequality for $c_R < \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$ returns

$$c_R \leq \frac{w_R}{w_P}\left(\underline{c} + \frac{w_R}{w_P}\left(\frac{1-\alpha}{\beta}\right) - \sqrt{\frac{w_R - w_P}{w_P}\left(\frac{1-\alpha}{\beta}\right)\left(\frac{w_R + w_P}{w_P}\left(\frac{1-\alpha}{\beta}\right) + 2\underline{c}\right)}\right).$$

Denote this threshold by $\hat{c}_R$.

For all contributions $c_R \leq \hat{c}_R$ P then trivially applies $k_P = 0$ and contributes either $\underline{c}$ if $c_R \leq \underline{c} + \frac{1-\alpha}{\beta}$ or $c_R - \frac{1-\alpha}{\beta}$ if $c_R > \underline{c} + \frac{1-\alpha}{\beta}$. For all contributions $c_R > \hat{c}_R$ P responds with $k_P = 1$ and contributes either $\underline{c}$ if $c_R \leq \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$ or $\frac{w_P}{w_R}c_R - uc$ if $c_R > \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$.
□

Proof of Proposition 4.1:

I prove the proposition in three steps. First, I show that P always applies the same normative principle as R. Second, I determine the optimal contributions for P in the sub-intervals. Last, I deduce the condition for β where P switches from playing the contribution in one sub-interval to the other.

1. Let $c_P \in \left[\underline{c}, \underline{c} + \frac{1-\alpha}{\beta}\right]$ and thus R contribute $\underline{c}$. Then

$$\begin{aligned} u_P(c_P, \underline{c}, 0) &= \pi_P(c_P, \underline{c}) - \frac{\beta}{2}\left(c_P - \underline{c}\right)^2 \\ &> \pi_P(c_P, \underline{c}) - \frac{\beta}{2}\left(c_P - \frac{w_P}{w_R}\underline{c}\right)^2 \\ &= u_P(c_P, \underline{c}, 1), \end{aligned}$$

since $c_P \geq \underline{c}$ and $\frac{w_P}{w_R} < 1$. Similarly, let $c_P \in \left(\underline{c} + \frac{1-\alpha}{\beta}, w_P\right]$ and thus R contribute $c_P - \frac{1-\alpha}{\beta}$. Then, for all c_P within the range,

$$\begin{aligned} u_P(c_P, c_P - \frac{1-\alpha}{\beta}, 0)) &= \pi_P(c_P, c_P - \frac{1-\alpha}{\beta}) - \frac{\beta}{2}\left(c_P - \left(c_P - \frac{1-\alpha}{\beta}\right)\right)^2 \\ &> \pi_P(c_P, c_P - \frac{1-\alpha}{\beta}) - \frac{\beta}{2}\left(c_P - \frac{w_P}{w_R}\left(c_P - \frac{1-\alpha}{\beta}\right)\right)^2 \\ &= u_P(c_P, c_P - \frac{1-\alpha}{\beta}, 1), \end{aligned}$$

since $c_P > c_P - \frac{1-\alpha}{\beta}$ and $\frac{w_P}{w_R} < 1$. Hence, P always chooses $k_P = 0$.

2. By partial derivation of u_P with respect to c_P, we get $\frac{\partial u_P}{\partial c_P}(c_P, \underline{c}, 0) = \alpha - 1 - \beta\left(c_P - \underline{c}\right)$. Because by definition $\alpha < 1$ and $c_P \geq \underline{c}$, utility is decreasing on the entire interval. Hence, u_P is maximized at $c_P = \underline{c}$.

   Analogous, for $u_P(c_P, c_P - \frac{1-\alpha}{\beta}, 0)$ we obtain $\frac{\partial u_P}{\partial c_P}(c_P, c_P - \frac{1-\alpha}{\beta}, 0) = 2\alpha - 1$, which by definition is greater than zero and thus u_P within this interval is maximized at $c_P = w_P$.

3. The comparison of the utilities from $(\underline{c}, 0)$ and $(w_P, 0)$ yields

$$\begin{aligned} & u_P(\underline{c}, \underline{c}, 0) \leq u_P(w_P, w_P - \frac{1-\alpha}{\beta}, 0) \\ \Leftrightarrow \quad & w_P + (2\alpha - 1)\,\underline{c} \leq 2\alpha w_P - \frac{1-\alpha^2}{2\beta} \\ \Leftrightarrow \quad & \frac{1-\alpha^2}{2\,(2\alpha-1)\,(w_P - \underline{c})} \leq \beta. \end{aligned}$$

□

Proof of Proposition 4.2:

As in Proposition 4.1, I prove the proposition in three steps. First, I show that P always applies the same normative principle as R. Second, I determine the optimal contributions for P in the sub-intervals. Last, I deduce the condition for β for which R switches from playing $\underline{c}$ to w_R.

1. R is faced with contributions of either $\underline{c}$ if $c_R \in \left[\underline{c}, \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)\right]$ or else $\frac{w_P}{w_R}c_R - \frac{1-\alpha}{\beta}$.

   In the former case the utility maximizing contribution for $k_R = 0$ obviously is $\underline{c}$, because $\frac{\partial u_R}{\partial c_R} = -1 + \alpha - \beta(c_R - \underline{c}) < 0 \quad \forall c_R \geq \underline{c}$. The utility maximizing contribution for $k_R = 1$ is either (a) $c_R = \frac{w_R}{w_P}\underline{c} - \frac{1-\alpha}{\beta}$ if $\frac{w_R}{w_P}\underline{c} - \frac{1-\alpha}{\beta} > \underline{c}$ or (b) $c_R = \underline{c}$ otherwise. In both cases $k_R = 0$ induces higher utility because for (a) both monetary payoff is lower and deviation is higher than under $k_R = 1$ and for (b) monetary payoff is equal and deviation is higher.

   In the latter case, R always chooses $k_R = 1$. For R to use $k_R = 0$, the penalty from deviating needs to be smaller for $k_R = 0$ than for $k_R = 1$ (because monetary payoff is equal for any given c_R). Thus, the inequality

$$\frac{\beta}{2}\left(c_R - \frac{w_P}{w_R}c_R + \frac{1-\alpha}{\beta}\right)^2 < \frac{\beta}{2}\left(\frac{w_R}{w_P}\frac{1-\alpha}{\beta}\right)^2$$

   has to be satisfied for some c_R. Because $1 - \frac{w_P}{w_R} > 0$, the left side of the inequality increases in c_R. Therefore, at least $c_R = \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$ must satisfy the condition. Inserting into the left side of the inequality yields

$$\left(\frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)\left(1 - \frac{w_P}{w_R}\right) + \frac{1-\alpha}{\beta}\right)^2 = \left(\underline{c}\left(\frac{w_R}{w_P} - 1\right) + \frac{w_R}{w_P}\frac{1-\alpha}{\beta}\right)^2$$

   and because $\underline{c}\left(\frac{w_R}{w_P} - 1\right) > 0$, it is always larger than $\left(\frac{w_R}{w_P}\frac{1-\alpha}{\beta}\right)^2$. Hence, there exists no $c_R > \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$ for which $k_R = 0$ achieves higher utility than $k_R = 1$.

2. From R's best response we know that she chooses to play the strategy $(\underline{c}, 0)$ if P's contribution is such that $c_P \in \left[\underline{c}, \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)\right]$.

   In the interval $\left(\frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right), w_R\right]$, I determine the optimal contribution through

$$\frac{\partial u_R}{\partial c_R}(c_R, \frac{w_P}{w_R}c_R - \frac{1-\alpha}{\beta}, 1) = -1 + \alpha + \alpha\frac{w_P}{w_R}.$$

Thus, I have to distinguish between three cases. (i) If $\alpha > \frac{w_R}{w_R+w_P}$, R's optimal contribution is w_R, because utility is increasing in c_R. (ii) If $\alpha < \frac{w_R}{w_R+w_P}$, utility is decreasing in c_R and (iii) if $\alpha = \frac{w_R}{w_R+w_P}$, all contributions return the same utility. However, I show that in cases (ii) and (iii) R will never choose to contribute $c_R > \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$ because $(\underline{c}, 0)$ yields higher utility.

Let $\alpha < \frac{w_R}{w_R+w_P}$. Since utility is decreasing in c_R, $c_R = \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$ constitutes an upper bound for the utility within this interval. It follows that $c_P = \underline{c}$ and therefore

$$u_R(\frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right), \underline{c}, 1) = w_R + \underline{c}\left(\frac{w_R}{w_P}(\alpha-1) + \alpha\right) - \beta\left(\frac{w_R}{w_P} + \frac{1}{2}\right)\left(\frac{1-\alpha}{\beta}\right)^2.$$

This utility can never be larger than the utility from playing $(\underline{c}, 0)$, because rearranging

$$w_R + \underline{c}\left(\frac{w_R}{w_P}(\alpha-1) + \alpha\right) - \beta\left(\frac{w_R}{w_P} + \frac{1}{2}\right)\left(\frac{1-\alpha}{\beta}\right)^2 < w_R + \underline{c}\,(2\alpha - 1)$$

yields

$$\underline{c}\left(\frac{w_R}{w_P} - 1\right)(\alpha - 1) - \beta\left(\frac{w_R}{w_P} + \frac{1}{2}\right)\left(\frac{1-\alpha}{\beta}\right)^2 < 0,$$

which is always satisfied by definition of $\frac{w_R}{w_P} > 1$ and $\alpha < 1$.

Let $\alpha = \frac{w_R}{w_R+w_P}$. Then any contribution within the interval yields the same utility. Assume therefore that $c_R = w_R$ and thus $c_P = w_P - \frac{1-\alpha}{\beta}$. The utility level is

$$u_R(w_R, w_P - \frac{1-\alpha}{\beta}, 1) = \alpha\left(w_R + w_P - \frac{1-\alpha}{\beta}\right) - \frac{\beta}{2}\left(\frac{w_R}{w_P}\right)^2\left(\frac{1-\alpha}{\beta}\right)^2.$$

Substituting α into the utility yields

$$\begin{aligned}
&\frac{w_R}{w_R+w_P}\left(w_R + w_P - \frac{w_P}{\beta(w_R+w_P)}\right) - \frac{\beta}{2}\left(\frac{w_R}{w_P}\right)^2\left(\frac{w_P}{\beta(w_R+w_P)}\right)^2 \\
= \quad & w_R - \frac{w_R w_P}{\beta(w_R+w_P)^2} - \frac{w_R^2}{2\beta(w_R+w_P)^2} \\
= \quad & w_R - \frac{w_R^2 + 2w_R w_P}{2\beta(w_R+w_P)^2},
\end{aligned}$$

which is less than $u_R(\underline{c}, \underline{c}, 0) = w_R + \underline{c}\,(2\alpha - 1) = w_R + \underline{c}\left(\frac{w_R - w_P}{w_R + w_P}\right)$.

Hence, $\alpha > \frac{w_R}{w_R+w_P}$ is a necessary condition for R to choose $(w_R, 1)$.

3. Let $\alpha > \frac{w_R}{w_R+w_P}$. The comparison of the utilities from $(w_R, 1)$ and $(\underline{c}, 0)$ yields

$$\begin{aligned}
& u_R(\underline{c}, \underline{c}, 0) \leq u_R(w_R, w_P - \frac{1-\alpha}{\beta}, 1) \\
\Leftrightarrow \quad & w_R + \underline{c}(2\alpha - 1) \leq (w_R + w_P)\alpha - \frac{1-\alpha}{\beta}\left(\alpha + \left(\frac{w_R}{w_P}\right)^2 \frac{1-\alpha}{2}\right) \\
\Leftrightarrow \quad & \frac{(1-\alpha)\left(\alpha + \left(\frac{w_R}{w_P}\right)^2 \frac{1-\alpha}{2}\right)}{(w_R + w_P - 2\underline{c})\alpha - (w_R - \underline{c})} \leq \beta
\end{aligned}$$

□

Proof of Proposition 4.3:

The steps are similar to preceding proofs.

1. We may again use the proofs of the previous propositions to rule out ranges of contributions.

   From the previous proofs we already know that within the lowest interval $\left[\underline{c}, \underline{c} + \frac{1-\alpha}{\beta}\right]$, $(\underline{c}, 0)$ is the best response. Therefore, no contribution in the fourth interval $\left(\tilde{c}_R, \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)\right]$ can ever be optimal.

   Similarly, in the second interval $\left(\underline{c} + \frac{1-\alpha}{\beta}, w_P + \frac{1-\alpha}{\beta}\right]$ utility is maximized by $c_R = w_P + \frac{1-\alpha}{\beta}$ and $k_R = 0$. Therefore, no contribution in the third interval $\left(w_P + \frac{1-\alpha}{\beta}, \tilde{c}_R\right]$ can be optimal. Last, if $c_R \in \left(\frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right), w_R\right]$, R's best strategy is to play $(w_R, 1)$ if α is large enough. However, notice that the utility of the strategy profile $\left(\left(w_P + \frac{1-\alpha}{\beta}, 0\right), (w_P, 0)\right)$ is

$$u_R\left(w_P + \frac{1-\alpha}{\beta}, w_P, 0\right) = w_R + w_P(2\alpha - 1) - \frac{3}{2}\frac{(1-\alpha)^2}{\beta},$$

   and that of the strategy profile $\left((w_R, 1), \left(w_P - \frac{1-\alpha}{\beta}, 1\right)\right)$ is

$$u_R\left(w_R, w_P - \frac{1-\alpha}{\beta}, 1\right) = (w_R + w_P)\alpha - \frac{1-\alpha}{\beta}\left(\alpha + \left(\frac{w_R}{w_P}\right)^2 \frac{1-\alpha}{2}\right).$$

Comparing these two utilities yields that $\left(\left(w_P + \frac{1-\alpha}{\beta}, 0\right), (w_P, 0)\right)$ gives higher utility if

$$\beta > \frac{3 - 5\alpha - (1-\alpha)\left(\frac{w_R}{w_P}\right)^2}{2(w_R - w_P)}.$$

But the numerator of the fraction is only positive if

$$\left(\frac{w_R}{w_P}\right)^2 < \frac{3-5\alpha}{1-\alpha}.$$

However, that would imply that $\frac{3-5\alpha}{1-\alpha} > 1 \Leftrightarrow \alpha < \frac{1}{2}$, which is impossible by definition. Therefore, R also never contributes $c_R = w_R$.

2. Having excluded all other equilibrium candidates but two, the comparison of the utilities from $(\underline{c}, 0)$ and $\left(w_P + \frac{1-\alpha}{\beta}, 0\right)$ yields

$$\begin{aligned} & u_R\left(w_P + \frac{1-\alpha}{\beta}, w_P, 0\right) \leq u_R(\underline{c}, \underline{c}, 0) \\ \Leftrightarrow \quad & w_R + w_P(2\alpha - 1) - \frac{3}{2}\frac{(1-\alpha)^2}{\beta} \leq w_R + (2\alpha - 1)\,\underline{c} \\ \Leftrightarrow \quad & \beta \leq \frac{3(1-\alpha)^2}{2(w_P - \underline{c})(2\alpha - 1)}. \end{aligned}$$

□

Proof of Proposition 4.4:

1. R is faced with a quadripartite utility function, but only with three different contribution levels.

   As before know that $(\underline{c}, 0)$ is the optimal choice within $\left[\underline{c}, \underline{c} + \frac{1-\alpha}{\beta}\right]$. Then it is quite obvious that R will never use $c_R \in \left(\hat{c}_R, \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)\right]$ irrespective of the applied normative principle. This is due to the fact that $(\underline{c}, 0)$ yields the highest utility whenever P contributes $\underline{c}$.

   The arguments regarding $c_R > \frac{w_R}{w_P}\left(\underline{c} + \frac{1-\alpha}{\beta}\right)$ carry over, too. Thus for $\alpha > \frac{w_R}{w_R + w_P}$, $(w_R, 1)$ is the best-response in this particular sub-interval.

   The sub-interval $\left(\underline{c} + \frac{1-\alpha}{\beta}, \hat{c}_R\right]$ remains to be investigated. It is optimal for R to use $k_R = 0$, because the monetary payoff is equal for any c_R and the deviation from the normative principle is lower than if $k_R = 1$. The latter is shown as follows.

$$\frac{\beta}{2}\left(c_R - c_R + \frac{1-\alpha}{\beta}\right)^2 < \frac{\beta}{2}\left(c_R - \frac{w_R}{w_P}\left(c_R + \frac{1-\alpha}{\beta}\right)\right)^2$$
$$\frac{1-\alpha}{\beta} < |c_R - \frac{w_R}{w_P}\left(c_R + \frac{1-\alpha}{\beta}\right)|$$

For $c_R > \frac{w_R}{w_P}\left(c_R + \frac{1-\alpha}{\beta}\right)$ the inequality translates to $\frac{1-\alpha}{\beta} < c_R$. The condition is always satisfied because $c_R > \underline{c} + \frac{1-\alpha}{\beta}$ and $\underline{c} > 0$. If on the other hand $c_R < \frac{w_R}{w_P}\left(c_R + \frac{1-\alpha}{\beta}\right)$, then the inequality turns into $0 < \left(c_R + \frac{1-\alpha}{\beta}\right)\left(\frac{w_R}{w_P} - 1\right)$. Again, because $\frac{w_R}{w_P} > 1$, the inequality is always satisfied and $k_R = 1$ yields higher deviation from the normative principle than $k_R = 0$.

The optimal contribution level within the interval is determined by partial derivation of the utility function with respect to c_R:

$$\frac{\partial u_R}{\partial c_R}(c_R, c_R - \frac{1-\alpha}{\beta}, 0) = 2\alpha - 1,$$

which is always greater than zero and thus utility is maximized within that interval by the contribution $c_R = \hat{c}_R$.

2. There are three candidates for the equilibrium with utilities

$$\begin{aligned} u_R(\underline{c}, \underline{c}, 0) &= w_R + (2\alpha - 1)\underline{c}, \\ u_R(\hat{c}_R, \hat{c}_R - \frac{1-\alpha}{\beta}, 0) &= w_R + (2\alpha - 1)\frac{w_R}{w_P}(\underline{c} - \\ &\sqrt{\frac{w_R - w_P}{w_P}\left(\frac{1-\alpha}{\beta}\right)\left(\frac{w_R + w_P}{w_P}\left(\frac{1-\alpha}{\beta}\right) + 2\underline{c}\right)} \\ &-\left(\frac{w_R}{w_P}\right)^2 \frac{2\alpha^2 - 3\alpha + 1}{\beta} - \frac{1-\alpha^2}{2\beta} \end{aligned}$$

and

$$u_R(w_R, w_P - \frac{1-\alpha}{\beta}, 1) = (w_R + w_P)\alpha - \frac{1-\alpha}{\beta}\left(\alpha + \left(\frac{w_R}{w_P}\right)^2 \frac{1-\alpha}{2}\right).$$

Suppose that $\alpha \leq \frac{w_R}{w_R + w_P}$. Then only $((\underline{c}, 0), (\underline{c}, 0))$ and $((\hat{c}_R, 0), (\hat{c}_R - \frac{1-\alpha}{\beta}, 0))$ are possible equilibria. The comparison of the utilities and solving for $\beta \geq 0$ yields

$$
\begin{aligned}
& u_R(\hat{c}_R, \hat{c}_R - \frac{1-\alpha}{\beta}, 0) \leq u_R(\underline{c}, \underline{c}, 0) \\
\Leftrightarrow \quad & \beta \leq \frac{2w_R(1-\alpha)\sqrt{(2-\alpha)(2\alpha-1)} + w_P\,(1-\alpha^2)}{2\underline{c}(w_R - w_P)(2\alpha-1)}.
\end{aligned}
$$

$\square$

4.B Supplementary Tables

Table 4.2: Alternative calculations of SPNE and total contributions, $\alpha = 0.7$

(a) Poor first-mover

	$\underline{c}$	1		4		5		8	
β		EQ	C	EQ	C	EQ	C	EQ	C
0.1	low	w_P	27	w_P	27	w_P	27	w_P	27
	high	w_P	17	$\underline{c}$	8	$\underline{c}$	10	$\underline{c}$	16
0.2	low	w_P	28.5	w_P	28.5	w_P	28.5	w_P	28.5
	high	w_P	19	w_P	19	w_P	19	$\underline{c}$	16
0.5	low	w_P	29.4	w_P	29.4	w_P	29.4	w_P	29.4
	high	w_P	19.4	w_P	19.4	w_P	19.4	w_P	19.4
1	low	w_P	29.7	w_P	29.7	w_P	29.7	w_P	29.7
	high	w_P	19.7	w_P	19.7	w_P	19.7	w_P	19.7
1.5	low	w_P	29.8	w_P	29.8	w_P	29.8	w_P	29.8
	high	w_P	19.8	w_P	19.8	w_P	19.8	w_P	19.8
2	low	w_P	29.9	w_P	29.9	w_P	29.9	w_P	29.9
	high	w_P	19.9	w_P	19.9	w_P	19.9	w_P	19.9

(b) Rich first-mover

	$\underline{c}$	1		4		5		8	
β		EQ	C	EQ	C	EQ	C	EQ	C
0.1	low	$\underline{c}$	2	$\underline{c}$	8	$\underline{c}$	10	$\underline{c}$	16
	high	$\underline{c}$	2	$\underline{c}$	8	$\underline{c}$	10	$\underline{c}$	16
0.2	low	$\underline{c}$	2	$\underline{c}$	8	$\underline{c}$	10	$\underline{c}$	16
	high	$\underline{c}$	2	$\underline{c}$	8	$\underline{c}$	10	w_P^+	21.5
0.5	low	w_R	39.4	$\underline{c}$	8	$\hat{c}_R$	12.2	$\hat{c}_R$	20.6
	high	$\underline{c}$	2	$\hat{c}_R$	13	$\hat{c}_R$	17.1	w_P^+	20.6
1	low	w_R	39.7	w_R	39.7	w_R	39.7	$\hat{c}_R$	21.9
	high	$\underline{c}$	2	$\hat{c}_R$	15	$\hat{c}_R$	19.5	w_P^+	20.3
1.5	low	w_R	39.8	w_R	39.8	w_R	39.8	$\hat{c}_R$	22.6
	high	$\hat{c}_R$	3.1	$\hat{c}_R$	16.1	w_P^+	20.2	w_P^+	20.2
2	low	w_R	39.9	w_R	39.9	w_R	39.9	$\hat{c}_R$	23.1
	high	$\hat{c}_R$	3.3	$\hat{c}_R$	16.9	w_P^+	20.2	w_P^+	20.2

Chapter 5

Minimum Participation Rules for the Provision of Public Goods[1]

Abstract This paper considers the endogenous formation of an institution to provide a public good. If the institution governs only its members, players have an incentive to free ride on the institution formation of others and the social dilemma is simply shifted to a higher level. Addressing this second-order social dilemma, we study the effectiveness of three different minimum participation requirements: 1. full participation / unanimity rule; 2. partial participation; 3. unanimity first and in case of failure partial participation. While unanimity is most effective once established, one might suspect that a weaker minimum participation rule is preferable in practice as it might facilitate the formation of the institution. The data of our laboratory experiment do not support this latter view, though. In fact, weakening the participation requirement does not increase the number of implemented institutions. Thus, we conclude that the most effective participation requirement is the unanimity rule which leaves no room for free riding on either level of the social dilemma.

Keywords Public Goods, Coalition Formation, Endogenous Institutions
JEL Classification C72, C92, H41, D02

[1] A shorter version of this chapter has appeared in the European Economic Review, 2013, 64. It is co-authored by Anke Gerber and Philipp C. Wichardt.

5.1 Introduction

The provision of public goods is a social dilemma that has attracted a lot of attention ever since the seminal article of Samuelson (1954). Many solutions have been proposed so far, all relying on some institution that sets the rule of the game such as to provide individuals the incentive to contribute the efficient amount to the public good. These institutions can be characterized either by centralized or decentralized sanctioning (punishment and reward) and the sanctioning can either be formal (monetary transfers) or informal (e.g. social ostracism).[2]

While institutions have been shown to be effective not only theoretically but also empirically, the main hindrance to their adoption is the fact that the implementation itself is a public goods problem. Obviously, any selfish individual prefers others to install the institution and provide the public good. Hence, it is questionable whether the institution will be implemented at all if membership is voluntary. In the present paper, we focus on this second order public goods problem and study the endogenous formation of an institution that, once established, enforces efficient contributions to a public good by all its members.

There is a large theoretical literature on endogenous institution (coalition) formation in the context of global environmental problems (see e.g. Carraro and Siniscalco 1993; Barrett 1994; Carraro and Marchiori 2003; Finus and Rundshagen 2003 yielding the pessimistic result that stable coalitions usually comprise only few members and hence efficiency is low.[3] In these models there are no restrictions on the size of coalitions. A straightforward question therefore is whether efficiency could be increased under appropriate restrictions on coalition size. A natural restriction would be a minimum participation requirement and in fact minimum participation rules are very common in international environmental agreements.

The Kyoto protocol, for example, had to be ratified by at least 55 parties accounting for at least 55 percent of greenhouse gas emissions in 1990 before entering into force. Other treaties like the Convention for the Protection of the Marine Environment of the North-East Atlantic required ratification by all abutting nations. Rutz (2001) has analyzed data provided by the International Center for Earth Science Information Network (CIESIN) showing that the number of international environmental treaties that do not contain any

[2] Examples for centralized formal sanctioning institutions are the mechanisms proposed by Groves and Ledyard (1977), Moore and Repullo (1988), Abreu and Sen (1990), Palfrey and Srivastava (1991), Jackson (1992), Falkinger (1996) and more recently Gerber and Wichardt (2009, 2013). Decentralized sanctioning with punishment or rewards being executed by players themselves have been studied among others by Fehr and Gächter (2000, 2002), Masclet et al. (2003) and Sefton et al. (2007).

[3] Stability here refers to the notion of cartel stability introduced by D'Aspremont et al. (1983).

minimum participation requirement is negligible.

Focusing on the consequences of minimum participation requirements, this paper addresses the empirical question which minimum participation rule is most effective in terms of maximizing overall efficiency in the context of a simple linear public goods game. While a stricter requirement leads to the implementation of larger institutions, and, hence to an increase in efficiency whenever an institution is actually implemented, a weaker requirement may be more effective overall since it renders the implementation of an institution less vulnerable to a coordination failure.

We conducted a laboratory experiment, where groups of four players interacted in an institution formation game under three different minimum participation rules: In treatment 1 an institution was only implemented when all four group member joined the institution (unanimity rule). In treatment 2 at least three group members had to join the institution and in treatment 3 the minimum participation requirement was relaxed to three in a second institution formation stage whenever institution formation failed under the unanimity rule in the first stage.[4]

According to the data, overall efficiency turns out to be largest in those treatments that restrict to or start with the unanimity rule (treatments 1 and 3). What is more, the total number of institutions is not larger under a weaker minimum participation rule than under the unanimity rule.[5] From an applied point of view, the results of our experiment, thus, show that weakening participation requirements does not improve outcomes. And, although the conditions in our experiment are very special (e.g. homogenous players, small groups, perfect enforcement of contributions by an institution), we believe that the basic findings are likely to hold in more complex settings as well.

Furthermore, in view of the debate about the importance of social preferences, it is interesting to note that a substantial proportion of groups implemented institutions of size three in treatments 2 and 3. This implies that these groups apparently tolerate that one player free rides on the public goods contributions of those who joined the institution, which in turn can be interpreted as evidence against a large proportion of inequality averse subjects.

Regarding the existing literature, our paper is related to a number of experimental

[4] Note that the institution formation game itself bears some resemblance to so called threshold public good games. In these games, the creation of the public good – here the institution – is achieved once a predefined level of contributions – here the conditional commitment of 3 (4) players – has been reached (for a further discussion of threshold or provision point public goods see, for example, Bagnoli and McKee 1991; Cadsby and Maynes 1999; Croson and Marks 2000; see Bagnoli and Lipman 1989, for a theoretical argument).

[5] The total number of institutions under the weak participation requirement is even smaller than under the unanimity rule. However, the difference is not significant.

papers on endogenous institution formation, most of which, however, consider the case where the institution governs all players and hence there is no second order public goods problem (Walker et al., 2000; Gürerk et al., 2006; Tyran and Feld, 2006; Kroll et al., 2007; Ertan et al., 2009; Sutter et al., 2010). Only few papers study coalition formation when the coalition only governs its members: Kosfeld et al. (2009) provide experimental evidence on the endogenous formation of a punishing institution showing that in most cases all players become members of the institution if an institution is formed at all. Dannenberg et al. (2010) compare the effectiveness of institutions that differ in the level of public good provision required from its members. The experimental data shows that weakening the rules such as to lower the free riding incentives in general does not increase overall efficiency which is akin to our result that efficiency is not increased under a weaker minimum participation rule. In a follow-up paper, Dannenberg (2012) studies coalition formation when members can vote on a binding minimum provision level of the public good and finds that social welfare is independent of the voting rule. In all of these papers players could form institutions of arbitrary size, while we set out to study institution formation under different minimum participation rules. A further interesting exception is Hamman et al. (2011), who study the effects of endogenous delegation of contribution decisions to an allocator whose decision only affects those who decided to vote. They find that delegation in general improves contributions to the public good and that allowing players to communicate increases the frequency of (endogenous) delegation.

Moreover, our paper is related to Carraro et al. (2009) who analyze a model of coalition formation under a minimum participation rule which is endogenously determined by unanimity voting before the coalition formation stage.[6] The authors derive conditions on the players' payoff functions under which a particular minimum participation requirement (including the unanimity rule) is chosen in equilibrium.

5.2 Model and Theoretical Predictions

5.2.1 The Basic Model

Consider the following symmetric linear public goods game PG with $n \geq 2$ players. Each player i has a private endowment $w > 0$ and can choose to contribute $g_i \in [0, w]$ to the

[6] This modeling raises the obvious question why the unanimity voting rule is not chosen endogenously as well. It is self-evident that there is no limit to the number of stages that could be added in order to endogenize the rules under which the rules of the game are determined.

public good. For given contributions $g = (g_1, \ldots, g_n)$ player i's payoff is

$$\pi_i(g) = w - g_i + a \sum_{j=1}^{n} g_j \tag{5.1}$$

where a with $\frac{1}{n} < a < 1$ is the marginal benefit from contributions to the public good. Since $a < 1$, $g_i^0 = 0$ is a dominant strategy for all i, which implies that the game has a unique Nash equilibrium where no one contributes to the public good. However, since $na > 1$, the welfare maximizing strategy profile is $g^* = (w, \ldots, w)$.

Suppose now that, before playing PG, players can decide whether to join an institution that enforces full contributions to the public good.[7] For the sake of argument, we assume that the institution is costless (the subsequent results do not change as long as the welfare gain of full public good contributions outweighs the costs, though). The institution is implemented if and only if an exogenously given minimum participation requirement is met. If the institution is implemented, all members are forced to contribute their full endowment to the public good, while all nonmembers can freely choose their contribution level. More precisely, for any given minimum participation requirement m with $1 \leq m \leq n$ we consider the following two-stage institution formation game IFm:

Stage 1: All players simultaneously decide whether to join the institution or not. If at least m players join the institution, the institution is implemented. Otherwise, it is not implemented. In both cases the game continues with stage 2.

Stage 2: PG is played with the following rules: If an institution was implemented in stage 1, then all members i of the institution are restricted to contribute their full endowment w to the public good, i.e. $g_i = w$, while all nonmembers j simultaneously choose their contribution $g_j \in [0, w]$. If no institution was implemented, all players i simultaneously choose their contribution $g_i \in [0, w]$. Players' payoffs are given by (5.1).

To capture the possibility of renegotiation of minimum participation requirements we also consider another institution formation game, where the minimum participation requirement is weakened if no institution is implemented under a stricter minimum participation requirement. In particular, we consider the following three-stage institution formation game, IFmk, where m is the minimum participation requirement in a first participation phase and $k < m$ is the minimum participation requirement in a second

[7] We treat the institution as a black box because the specific mechanism which implements the desired contributions is of no relevance for our analysis.

participation phase:

Stage 1a: All players simultaneously decide whether to join the institution or not. If at least m players join the institution, the institution is implemented and the game continues with stage 2. Otherwise, the institution is not implemented and the game continues with stage 1b.

Stage 1b: All players decide simultaneously whether to join the institution or not. If at least k players join the institution, the institution is implemented. Otherwise, the institution is not implemented. In both cases the game continues with stage 2.

Stage 2: PG is played with the following rules: If an institution was implemented in stage 1a or in stage 1b, then all members i of the institution are restricted to contribute their full endowment w to the public good, i.e. $g_i = w$, while all nonmembers j simultaneously choose their contribution $g_j \in [0, w]$. If neither in stage 1a nor in stage 1b an institution was implemented, all players i simultaneously choose their contribution $g_i \in [0, w]$. Players' payoffs are given by (5.1).

In the following, we characterize the set of subgame perfect Nash equilibria for standard preferences as well as for social preferences as proposed by Fehr and Schmidt (1999). All proofs are in Appendix 5.A.

5.2.2 Institution Formation with Standard Preferences

Under standard preferences a player's utility equals her payoff, i.e. $u_i(g) = \pi_i(g)$ for all contribution profiles $g = (g_1, \ldots, g_n)$. The following proposition provides a characterization of the pure strategy subgame perfect Nash equilibria of the institution formation games.[8]

Proposition 5.1 *Let $u_i = \pi_i$ for all players i.*

(i) Let $ma > 1$. Then in any pure strategy subgame perfect Nash equilibrium of IFm *either an institution with exactly m members is implemented or no institution is implemented. In any subgame perfect Nash equilibrium nonmembers of an institution always contribute zero to the public good.*

(ii) Let $m > k$ and $ka > 1$. Then in any pure strategy subgame perfect Nash equilibrium of IFmk *either an institution with exactly m members is implemented in stage*

[8] There also exist subgame perfect Nash equilibria in mixed strategies.

1a or an institution with exactly k members is implemented in stage 1b or no institution is implemented. In any subgame perfect Nash equilibrium nonmembers of an institution always contribute zero to the public good.

Note that, by Proposition 5.1, all institution formation games have subgame perfect equilibria in which no institution is implemented. From a social planner's point of view, this might seem unsatisfactory as it implies that in theory none of the institution formation games is strictly preferable. However, the different socially undesirable equilibria are all such that all players are indifferent between joining and not joining the institution; this is due to the fact that the commitment entailed in the decision to join is only binding if sufficiently many others join as well. One may therefore argue that such equilibria are inherently unstable. Facing a similar problem, Kosfeld et al. 2009 focus on strict equilibria:[9]

Definition 5.1 *A subgame perfect Nash equilibrium of an extensive form game is called* ***stagewise strict****, if in every stage game every player's strategy is a unique best response to the equilibrium strategies of the other players.*

Using this strictness refinement the following result is straightforward.

Proposition 5.2 *Let $u_i = \pi_i$ for all players i and let $ma > 1$. Then, in any stagewise strict subgame perfect Nash equilibrium of* IFm*, an institution with exactly m members is implemented.*

Unfortunately, the strictness requirement is too strong for IFmk. In all subgame perfect Nash equilibria of IFmk, no institution is formed in at least one stage of the game and hence, for at least one player the corresponding equilibrium strategy is not a unique best response. Thus, there exists no stagewise strict subgame perfect Nash equilibrium in IFmk.

[9] In the sequel, we refer to these equilibria as "stagewise strict" in order to avoid any confusion with the notion of a strict Nash equilibrium of a normal form game. Note that the notion of a stagewise strict Nash equilibrium is equivalent to the notion of a strict Nash equilibrium in the agent normal form of the extensive game.

5.2.3 Institution Formation with Social Preferences

Experimental research over the last years has provided ample evidence for behavior being governed by social preferences. In the context of institution formation, players with social preferences may prefer no institution over an institution that is implemented by a subgroup of players, because the latter yields unequal payoffs to members and nonmembers if nonmembers do not contribute to the public good. For tractability reasons and in order to compare our results to those of Kosfeld et al. (2009), we apply the social preference model by Fehr and Schmidt (1999), where players are assumed to be inequality averse.[10] The payoffs of the players are given by $\pi = (\pi_1, \dots, \pi_n)$ as above. The utility of a player i is then defined as

$$u_i(g) = \pi_i(g) - \frac{\alpha_i}{n-1} \sum_{j \neq i} \max\{\pi_j(g) - \pi_i(g), 0\} - \frac{\beta_i}{n-1} \sum_{j \neq i} \max\{\pi_i(g) - \pi_j(g), 0\}. \quad (5.2)$$

The two parameters α_i and β_i measure the reduction of player i's utility due to disadvantageous inequality and advantageous inequality, respectively. Typically, it is assumed that $\alpha_i \geq \beta_i$ for all i and that $0 \leq \beta_i < 1$, which we also assume here.

In the sequel, we analyze the subgame perfect Nash equilibria of the institution formation games IF4, IF3 and IF43 conditional on the inequality aversion parameters of the players. We set $a := 0.4$ which will be the parameter value in our experiment.

To begin with, we note that, for the linear public goods game with payoffs given in (5.1), Fehr and Schmidt (1999, Prop. 4) show that it is a dominant strategy for each player with $a + \beta_i < 1$ to choose $g_i = 0$. Furthermore, if h denotes the number of players with $a + \beta_i < 1$ and the condition $\frac{h}{n-1} > \frac{a}{2}$ is met, the unique Nash equilibrium is $g_i = 0$ for all i. In order to characterize the set of subgame perfect Nash equilibria of IF4, IF3 and IF43, we define the following player types:

Type 1: $\beta_i \leq \alpha_i \leq 0.6$ and $\beta_i < 0.6$. This includes players with standard preferences who have parameters $\alpha_i = \beta_i = 0$. A player of type 1 is weakly averse against disadvantageous and advantageous inequality. It is a dominant strategy for this type to choose $g_i = 0$ whenever the player is not a member of an institution.

Type 2: $\alpha_i > 0.6 > \beta_i$. A player of type 2 is strongly averse against disadvantageous inequality but only weakly averse against advantageous inequality. As for type 1, it is a dominant strategy for type 2 to choose $g_i = 0$ whenever the player is not a member of

[10] We note that there are other models of social preferences and more specifically inequity aversion, among them Bolton and Ockenfels (2000), Charness and Rabin (2002), and Cox et al. (2007).

an institution.

Type 3: $\alpha_i \geq \beta_i \geq 0.6$. A player of type 3 is strongly averse against disadvantageous and advantageous inequality. If $\beta_i > 0.6$, it is a dominant strategy for this type to choose $g_i = w$ whenever an institution forms and the player is not a member of the institution. If $\beta_i = 0.6$, this type is indifferent between all contributions in $[0, w]$ whenever an institution forms and the player is not a member.

The subgame perfect Nash equilibria of IF4 then can be characterized as follows:

Proposition 5.3 *Let u_i be given as in (5.2) for all i. Then, for all preference parameters of the players, in any subgame perfect Nash equilibrium of* IF4*, either an institution with four members is implemented or no institution is implemented. Concerning the equilibrium contributions to the public good if no institution is implemented, there are two cases:*

(i) If all players are of type 3, then all players i contribute $g_i = g \in [0, w]$.

(ii) If there is at least one player of type 1 or 2, then all players contribute zero to the public good.

Next, we consider IF3 and IF43 which both allow for the formation of a 3-player institution. Types 1 and 2 always prefer to be nonmember of a 3-player institution over being a member of a 4-player institution as these types are only weakly averse against advantageous inequality: A player of type 1 or 2 contributes $g_i = 0$ as a nonmember and hence receives utility $u_i = 2.2w - \beta_i w$ if the other players implement a 3-player institution, while he receives $u_i = 1.6w < 2.2w - \beta_i w$ as a member of a 4-player institution. Hence, for weaker minimum participation requirements only type 3 players will form 4-player institutions. Moreover, no player of type 2 or 3 would want to be part of a 3-player institution if the nonmember contributes zero to the public good as both types are strongly averse against disadvantageous inequality. Type 1 players on the other hand have no reservations against being members of 3-player institutions even if the nonmember contributes zero.

Thus, using the above type-classification, we obtain the following straightforward result for IF3.

Proposition 5.4 *Let u_i be given as in (5.2) for all i and consider the institution formation game* IF3.

(i) The implementation of a 4-player institution is supported as a subgame perfect Nash equilibrium if and only if all players are of type 3. *The following contributions*

support this equilibrium outcome: Whenever no institution is implemented in stage 1, all players i contribute $g_i = g \in [0, w]$. Whenever a 3-player institution is implemented in stage 1, the nonmember i contributes $g_i = w$, if $\beta_i > 0.6$, and he contributes any amount $g_i \in [0, w]$, if $\beta_i = 0.6$.

(ii) *The implementation of a 3-player institution is supported as a subgame perfect Nash equilibrium if and only if either at least one player is of type 3 or at least three players are of type 1. The following contributions support this equilibrium outcome: Whenever no institution is implemented in stage 1, equilibrium contributions are $g_i = g \in [0, w]$ for all i if all players are of type 3 and $g_i = 0$ for all i, otherwise. Whenever a 3-player institution is implemented in stage 1, and the nonmember is of type 3, he contributes $g_i = w$. If the nonmember is of type 1 or 2, he contributes $g_i = 0$.*

(iii) *No institution is always supported as a subgame perfect Nash equilibrium and it is the only subgame perfect Nash equilibrium outcome if at least two players are of type 2 and no player is of type 3. The following contributions support this equilibrium outcome: Whenever no institution is implemented in stage 1, equilibrium contributions are $g_i = g \in [0, w]$ for all i if all players are of type 3 and $g_i = 0$ for all i, otherwise. Whenever a 3-player institution is implemented in stage 1, and the nonmember is of type 3, he contributes $g_i = w$, if $\beta_i > 0.6$, and he contributes any amount $g_i \in [0, w]$, if $\beta_i = 0.6$. If the nonmember is of type 1 or 2, he contributes $g_i = 0$.*

From Proposition 5.4 it follows that social preferences can be detrimental for efficiency under a weak minimum participation requirement: If no player has a strong aversion against advantageous inequality ($\beta_i < 0.6$) and at least two players have a strong aversion against disadvantageous inequality ($\alpha_i > 0.6$), then no institution is formed and welfare is minimized in equilibrium.

Finally, we derive the equilibrium outcomes of IF43. Since the equilibrium contributions are the same as in the corresponding cases of Proposition 5.3 and 5.4 we confine ourselves to stating the equilibrium outcome in terms of the implemented institutions:

Proposition 5.5 *Let u_i be given as in (5.2) for all i and consider the institution formation game* IF43.

(i) *The implementation of a 4-player institution in stage 1a is always supported as a subgame perfect Nash equilibrium.*

(ii) The implementation of a 4*-player institution in stage 1b is supported as a subgame perfect Nash equilibrium if and only if all players are of type* 3.

(iii) The implementation of a 3*-player institution in stage 1b is supported as a subgame perfect Nash equilibrium if and only if either at least one player is of type* 3 *or at least three players are of type* 1*.*

(iv) No institution is always supported as a subgame perfect Nash equilibrium.

5.2.4 Predictions

From the previous analysis of standard preferences, we can derive a number of testable predictions. The first prediction follows from the fact that in equilibrium nonmembers of an institution contribute zero to the public good.

Prediction 1 *The formation of an institution increases the level of public good provision.*

Concerning the average size of an institution, the theoretical analysis yields a clear prediction for IF4 and IF3: Under standard preferences no institution of size four is implemented in IF3. For IF43 the theoretical results are ambiguous, since both 4- and 3-player institutions can be implemented in equilibrium. Nevertheless, as long as both types of institutions are implemented with positive probability we can conclude that the average size of implemented institutions in IF43 is larger than in IF3 and smaller than in IF4 which leads to the following prediction:

Prediction 2 *The average size of the institutions that are implemented is larger the stricter the minimum participation requirement, i.e. the average size of an institution in* IF3 *is smaller than the average size of an institution in* IF43*, which in turn is smaller than the average size of an institution in* IF4*.*

Concerning the number of implemented institutions and, hence, overall efficiency achieved by the different minimum participation rules, our theoretical analysis yields ambiguous predictions due to the multiplicity of equilibria. On the one hand one may predict that more institutions are implemented and hence the level of public good provision is higher under a stricter minimum participation requirement, since the formation of a 3-player institution involves a coordination problem, which may very likely result in miscoordination and a complete failure to implement an institution. Following this line of argument we would predict that more institutions are implemented in IF4 than in IF3 and that IF43 is somewhere in between. On the other hand, under all minimum participation rules the implementation of an institution may simply fail because some players expect that everyone is playing the no institution equilibrium. This kind of miscoordination is more likely to lead to a failure to implement an institution under a strict than under a weak minimum participation requirement. If we follow this argument we would predict that more institutions are implemented in IF3 than in IF4 and that again IF43 is somewhere in between.

Given that we do not know whether one of these opposing effects dominates the other or whether they just cancel out, the theoretical analysis does not yield a clear prediction concerning the number of implemented institutions for different minimum participation rules. As a consequence, we also cannot derive a prediction about the level of public good provision for different minimum participation rules. This theoretical nondeterminacy further motivates our laboratory experiment.

5.3 Experimental Design and Procedure

Our experimental design consists of the three treatments IF4, IF3 and IF43 with endowment $w = 20$ and marginal benefit parameter $a = 0.4$. The experiment was conducted at the experimental laboratory of the School of Business, Economics and Social Sciences at the University of Hamburg between January and May 2011. In total, we ran nine sessions (three per treatment) with 196 mainly undergraduate students, predominantly from the social sciences. No participant had previously participated in a public goods experiment. In each session $20 - 24$ students participated, resulting in 17 groups for IF4 and 16 groups each for treatments IF3 and IF43. We used z-tree (Fischbacher, 2007) for programming and ORSEE (Greiner, 2004) for recruiting.

Each session started with a short introduction after which the instructions were read aloud, so that all participants knew that everyone received the same instructions. The actual experiment did not start until each subject had correctly answered a set of control questions.[11] Subjects were then randomly matched into groups of four players who stayed together throughout the entire experiment (partner matching), however the identities of group members remained unknown. At the end of each round, a summary screen was shown with the subject's individual contribution to the public good, the total group contribution and the subject's individual payoff in this round.

At the end of the experiment, one of the ten rounds was randomly selected for pay-

[11] An English translation of the instructions and control questions can be found in Appendix 5.C.

ment.[12] The exchange rate between Euro and experimental currency units (ECU) was 1 : 3. Additionally, subjects received a show-up fee of 4 Euro. The payment was carried out in private after a final questionnaire which included socio-demographic data and the Big-Five-Inventory-Shortversion (BFI-S).

The sessions lasted 60-90 minutes including instructions, control questions, questionnaire and payment. On average, a subject earned 13 Euro.

5.4 Results

In the sequel, we present the different results from the experiment.

5.4.1 Contributions to the Public Good

Figure 5.1 shows the average contributions to the public good in all periods. In total, average contributions are 13.6 in IF4, 12.4 in IF43 and 9.3 in IF3 respectively. In fact, IF4 and IF43 generate higher contributions to the public good than IF3 (Mann-Whitney test, 5% significance, where the average contribution per group is one independent observation), while there is no significant difference in the public good provision between IF4 and IF43 ($p = 0.43$).[13]

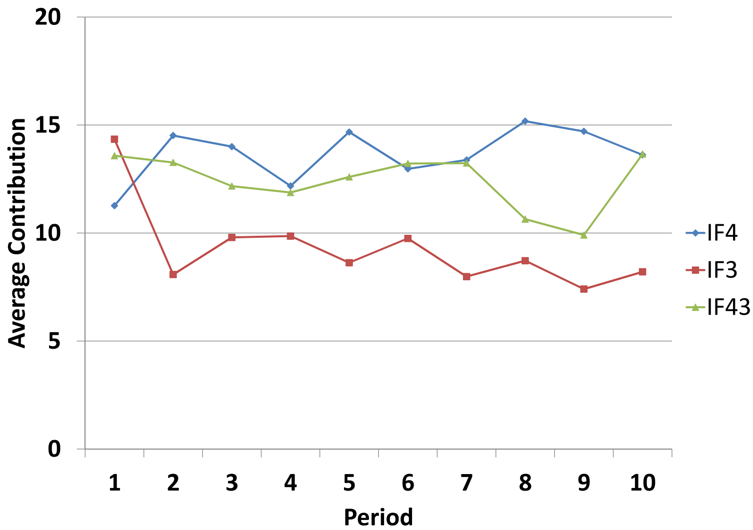


Figure 5.1: Average contributions per subject across treatments in periods 1-10.

[12] A ten-sided dice was thrown in public by the supervisor of the experiment in order to select the payment round.

[13] Here and in the following we report the results of two-sided Mann-Whitney tests.

Table 5.1: GLS regression results for public good contributions

	Contribution	
IF3	-4.382^{***}	(1.450)
IF43	-1.132	(1.643)
Const.	9.416^{**}	(4.157)
Observations	$N = 1960$	

Notes: Contributions in IF4 are baseline. IF3 (IF43) is a dummy variable that takes the value 1 if the treatment is IF3 (IF43) and 0 otherwise. Std. errors in parentheses are adjusted for group clusters. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

These results are confirmed by the regression depicted in Table 5.1, where IF4 is used as the baseline treatment. We also included age, sex and the big five personality traits from the questionnaire in the regression. However, the effects on the coefficients for IF3 and IF43 are negligible and in all our regressions the socio-demographic variables and BIG5 personality traits show no consistent and mostly insignificant effects on the behavior of the subjects. We therefore do not report the coefficients of these variables in our tables. Table 5.1 shows that, the contributions to the public good are significantly lower in IF3 than in IF4 (1% significance), while there is no significant difference between IF4 and IF43. We summarize these findings as follows:

Result 5.1 *Contributions to the public good are higher in* IF4 *and* IF43 *than in* IF3*, while there is no significant difference between* IF4 *and* IF43*.*

Next we have a look at the public good contributions conditional on whether an institution was implemented or not. To begin with, we consider the general effect of institutions being present. Figure 5.2 depicts average contribution levels when an institution is formed and when no institution is formed. As hypothesized, for all three treatments there are more contributions to the public good when an institution is implemented. For IF4, the average contribution obviously is 20 if an institution is formed, whereas it is only approximately 5.2 when no institution is formed. IF3 records mean contributions of 16.5 if an institution is formed and 3.7 if not. Subjects in IF43 on average contributed 18.6 to the public good when an institution was formed and 5.5 when no institution was formed. These differences between contributions with and without institutions are confirmed by several Mann-Whitney tests using the average contribution per group over all periods as one independent observation. For all treatments, the differences between the contributions when an institution is implemented and when no institution is implemented are significant on a 1%-level, which confirms Prediction 1:

Result 5.2 *In all treatments, average contributions to the public good are higher with than without an institution.*

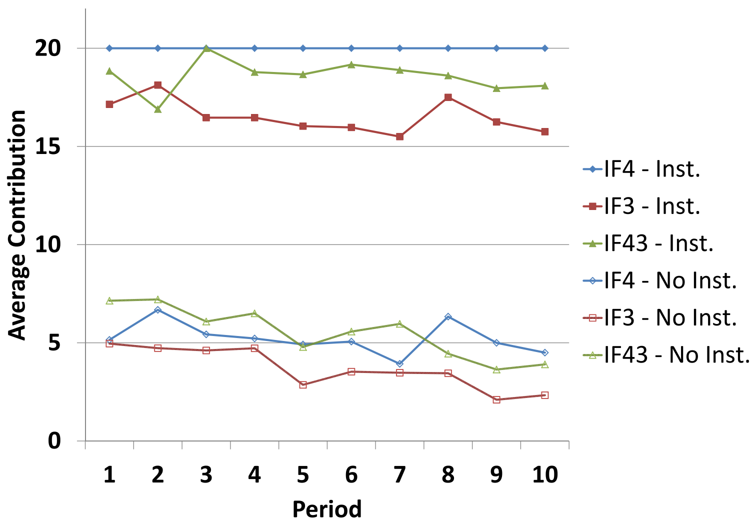


Figure 5.2: Average contributions per subject in periods 1-10 conditional on the existence/nonexistence of an institution.

Figure 5.2 also shows that average contributions to the public good when no institution is formed are only slightly different in the three treatments (5.2 in IF4, 3.7 in IF3 and 5.5 in IF43). Taking the average contribution per group as one independent observation, this result is confirmed by Mann-Whitney tests, which report no significant differences. Moreover, whenever an institution of size three was implemented, the nonmember in IF3 and IF43 on average contributes 2.5 to the public good with no significant difference between the two treatments. We also note that in IF3 the contribution of a nonmember when an institution of size three was implemented (average contribution 2.5) is significantly lower (5% significance) than the contribution of a subject if no institution has been formed (average contribution 3.7). In IF43, however, the difference is not significant.[14] We summarize these findings in the following result:

[14] The insignificance of the difference in IF43 is surprising at first sight, but has a simple explanation. Our data shows that the contributions of nonmembers of an institution are lower in IF3 than in IF43, both in the case where an institution is implemented and in the case where no institution is implemented. While in both cases this difference between IF3 and IF43 is not significant, the change destroys the significant difference in the contribution of nonmembers when an institution is formed and when no institution is formed that we observe in IF3.

Result 5.3 *In* IF3 *the contribution of the nonmember in the presence of a* 3*-player institution is lower than the contribution of a player if no institution was implemented. In* IF43 *the difference is not significant. Overall the contribution of a player is the same across all treatments whenever no institution was implemented and the contribution of the nonmember of a* 3*-player institution is the same in* IF3 *and* IF43*.*

5.4.2 Institution Formation

In the following we have a closer look at institution formation in the different treatments. A straightforward conjecture is that Result 5.1 can be explained by the number and size of the institutions implemented in the different treatments. Table 5.2 reports the proportion of implemented institution sizes across treatments.

Table 5.2: Proportion of institutions of size three and four

Inst. Size	**IF**4	**IF**3	**IF**43
4	57%	9%	35%
3	0%	35%	18%
Total	57%	44%	53%

We find that the observed differences in contributions to the public good cannot be explained with the total number of formed institutions. Although there are more institutions implemented in IF4 (institutions implemented in 57% of the cases) than in IF43 (53%) and IF3 (44%), these differences are not significant.[15] This confirms the ambiguous theoretical prediction concerning the number of implemented institutions in the different treatments:

Result 5.4 *There is no significant difference in the number of institutions implemented in all treatments.*

We continue by comparing the different sizes of institutions across treatments. The percentages of institutions with four members are 35% in IF43 and 9% in IF3. Again applying Mann-Whitney tests, we can establish that there are more institutions with four members in IF4 than in IF43 (10% significance) and IF3 (1% significance). Also, groups form 4-player institutions more frequently in IF43 than in IF3 (5% significance). At the same time there are more institutions with three members in IF3, 35%, than in IF43, 18% (1% significance). We observe that in IF3 only about 20% of the implemented institutions have size four, while in IF43 about 66% of the implemented institutions have

[15] For these tests, the number of institutions per group was used as one independent observation.

size four. Thus, different from the findings by Kosfeld et al. (2009), where most subjects are not willing to implement institutions that do not involve all group members, our subjects do not shy away from implementing 3-player institutions. Hence, we do not find much evidence for inequity aversion in our subject pool.

Overall, Mann-Whitney tests show that the average size of an institution formed in IF4 is larger than the average size of an institution in IF3 or IF43 (1% significance each). Furthermore, the average size of an institution in IF43 is larger than in IF3 (5% significance). This confirms Prediction 2:

Result 5.5 *On average, the largest institutions are formed in* IF4*, and institutions are larger in* IF43 *than in* IF3*.*

Finally, we explore whether institutions in IF43 are predominantly formed in stage 1a or in stage 1b. Figure 5.3 depicts the proportion of cases in which an institution of size three or four is formed in stage 1a and 1b, respectively. 62% of the institutions are formed already in stage 1a and only 38% in stage 1b. A Mann-Whitney test confirms that this difference is significant (5%-level). Furthermore, three institutions of size four were even formed in stage 1b, which implies that overall 66% of all implemented institutions are of size four.

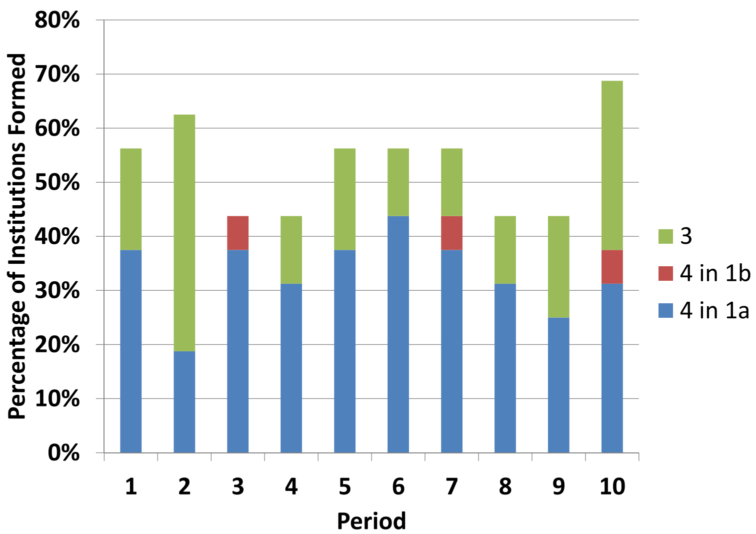


Figure 5.3: Institutions formed in IF43 in Periods 1-10 (proportion of cases with an institution of size three or four in stage 1a and 1b).

Summarizing, the observations concerning the number and size of institutions support the findings regarding the contributions across treatments: While there is no difference in the total number of institutions in IF3 and IF43, the latter has significantly more

4-player institutions leading to higher average contributions in IF43 than in IF3. There are also more 4-player institutions in IF4 than in IF43 but the difference is only weakly significant. This, together with the fact that there is also a sizable number of 3-player institutions in IF43 may explain why we do not observe a significant difference in the contributions between IF4 and IF43.

Having analyzed the differences in aggregate outcomes across treatments, we now proceed to explore individual behavior under the different minimum participation requirements.

5.4.3 Individual Participation Decisions

Table 5.3: Probit regression results for the probability to join the institution

	Join Institution			
	(1)		(2)	
IF3∨IF43	-0.833^{***}	(0.279)	-1.163^{***}	(0.263)
IF3	-0.642^{**}	(0.265)	-0.265	(0.246)
Const.	1.852^{***}	(0.214)	1.796^{***}	(0.202)
Observations	$N = 1960$		$N = 1960$	

Notes: Regressions (1) and (2) differ in the definition of the participation decision in IF43. In regression (1) we use the choice in stage 1a, whereas in (2) we apply the choice in the ultimate stage, i.e., the decision of stage 1a if an institution was implemented in stage 1a and the decision of stage 1b otherwise. The decision to join the institution in IF4 is baseline. IF3∨IF43 (IF3) is a dummy variable that takes the value 1 if the treatment is IF3 or IF43 (IF3) and 0 otherwise. Std. errors in parentheses adjusted for group clusters. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

In order to further understand the participation decisions of the subjects across treatments, we performed two probit regressions. The results are presented in Table 5.3. The regressions analyze the impact of the different treatments on the probability to join the institution in the institution formation stage (stage 1 in IF4 and IF3 and stage 1a, resp. stage 1b in IF43) independent of whether the institution was finally implemented or not. The two regressions differ in the way that the "join" decision is defined for IF43; in (1) only stage 1a is considered, in (2) the ultimate decision in each period is used. That is, if no institution has been formed in stage 1a, then the decision in stage 1b is used in the regression. We find that in treatments IF3 and IF43 the probability that a subject joins the institution is lower compared to IF4 (1% significance), and that the probability is even less for IF3 than for IF43 (5% significance). Regression (2) shows that, when considering the decision in stage 1*b* of IF43, the willingness to join the institution is even less. It is then not significantly higher in IF43 than in IF3.

Result 5.6 *The probability that a subject joins an institution is higher in* IF4 *than in* IF3 *and* IF43*.*

Observe that Result 5.6 is consistent with Results 5.4 and 5.5: The higher the probability to join an institution the higher the probability that a 4-player institution is formed. On the other hand, for given probabilities to join an institution, the probability that an institution is formed decreases with the minimum participation requirement. In our case the two effects of a higher probability to join an institution and a lower probability that an institution is implemented at all cancel out, so that there is no difference in the number of institutions formed in the different treatments.

We again used a Mann-Whitney test to compare IF3 with stage 1*b* of IF43. For this test, the number of institutions per group is not suitable to count as an observation, as the groups in IF43 do not always reach the second stage. Therefore, we used the number of institutions per group per stage 1*b* played as one independent observation.[16] We find that there are more institutions formed in IF3 than in stage 1*b* of IF43 (Mann-Whitney test 5%-level). Hence, the willingness to join the institution is higher in IF3 than in stage 1*b* of IF43. There are two potential explanations for this observation. First, the result may be driven by a selection effect in IF43, since only those groups, who fail to form an institution in stage 1*a*, reach stage 1*b*. If the groups that reach stage 1*b* have more members with a low inclination to join an institution than the average group, there will naturally be less institutions in stage 1*b* of IF43 than in IF3. As we can see from Table 5.4, except for three groups (3, 11, 15) who almost never reach stage 1*b*, all other groups reach stage 1*b* in at least five out of ten periods. While this speaks against a strong selection effect, we also note from Table 5.4 that six out of the seven groups, who almost always reach stage 1*b* (4, 7, 8, 9, 10, 13, 16), have one member who either never joins an institution or only joins in very few cases. The remaining groups (1, 2, 3, 5, 6, 11, 12, 14, 15) do not have such a member with a low likelihood to join an institution. Thus, the majority of observations in stage 1*b* come from groups who are more prone to fail in the formation of an institution than the average group, which provides some evidence in favor of a selection effect in IF43.

A second explanation for the fact that more institutions are formed in IF3 than in stage 1*b* of IF43 could be that subjects in IF43 interpret the failure to form an institution

[16] The played 1*b* stages vary from one up to ten across the different groups.

Table 5.4: Institution formation and join decisions in IF43

IF43		4a	4b	3	Join-a	Join-b
Group 1	S1	5	-	-	1	0.2
	S2				0.6	0
	S3				0.9	1
	S4				1	0.6
Group 2	S1	4	-	2	1	0.83
	S2				0.9	0.33
	S3				0.4	0
	S4				0.9	0.67
Group 3	S1	9	-	-	0.9	0
	S2				1	1
	S3				1	0
	S4				1	1
Group 4	S1	-	-	4	0.7	0.8
	S2				0.6	0.5
	S3				1	1
	S4				0.1	0
Group 5	S1	4	-	1	0.4	0.17
	S2				1	0
	S3				1	0.83
	S4				1	0.83
Group 6	S1	3	-	3	0.5	0.29
	S2				1	1
	S3				0.6	0.14
	S4				1	0.86
Group 7	S1	-	-	1	1	0.1
	S2				0	0
	S3				0.4	0.7
	S4				1	1
Group 8	S1	1	3	1	0.6	0.67
	S2				0.4	0.33
	S3				0.5	0.67
	S4				0.8	1
Group 9	S1	1	-	5	0.2	0
	S2				0.5	0.56
	S3				1	0.78
	S4				0.8	0.78
Group 10	S1	1	-	3	0.9	0.78
	S2				1	1
	S3				0.6	0.33
	S4				0.1	0
Group 11	S1	8	-	-	1	1
	S2				1	0.5
	S3				0.9	0
	S4				0.9	0.5
Group 12	S1	5	-	3	0.9	0.8
	S2				0.9	0.2
	S3				0.6	0.4
	S4				0.9	0.8
Group 13	S1	-	-	2	0	0
	S2				0.8	0.9
	S3				0.4	0.3
	S4				0.7	0.6
Group 14	S1	3	-	1	0.8	0.71
	S2				0.9	0.14
	S3				0.4	0.43
	S4				0.7	0.14
Group 15	S1	9	-	-	1	1
	S2				1	1
	S3				0.9	0
	S4				0.9	0
Group 16	S1	-	-	2	0.9	0.6
	S2				1	1
	S3				0	0
	S4				0.3	0.3

Notes: Column 4a (4b) shows the number of periods in which a group implemented a 4-player institution in stage 1a (1b). Column 3 shows the number of periods in which a group implemented a 3-player institution in stage 1b. Column Join-a (Join-b) shows the frequency with which the respective group member (subjects $S1$, $S2$, $S3$, and $S4$) joined the institution in stage 1a (1b).

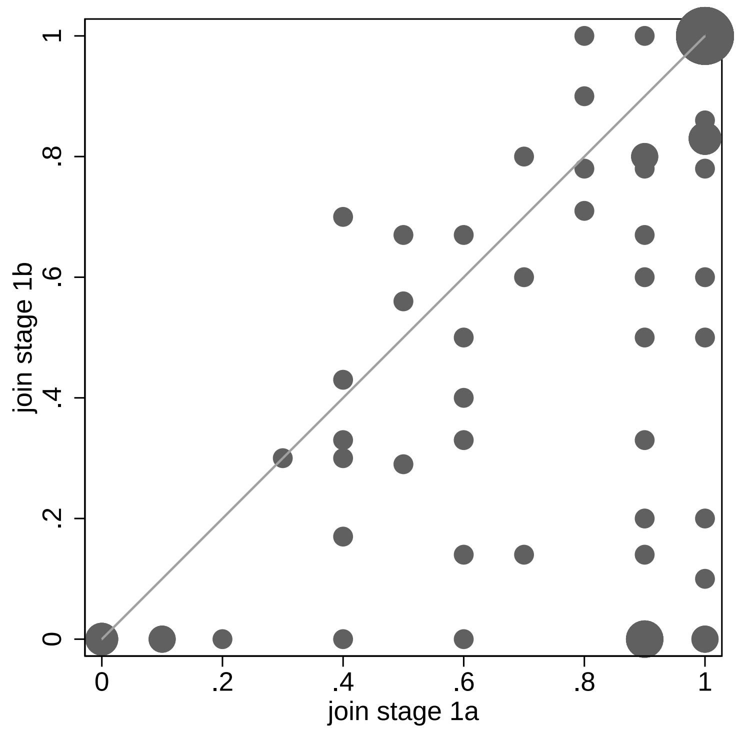


Figure 5.4: Proportion of cases in which a subject joined the institution in stage 1a and stage 1b of IF43. Small bubbles corresponds to one subject and the larger the size of a bubble the larger the number of subjects with the specific join decisions.

in stage 1a as a signal that at least one player intends to free ride on the public good provision of the others. This may induce subjects with social preferences not to join the institution in stage 1b. If this conjecture is true, we should observe that subjects are less likely to join the institution in stage 1b than in stage 1a. The scatter plot in Figure 5.4 depicts for every subject the proportion of cases in which the subject joined the institution in stage 1a and stage 1b. It shows that the vast majority of subjects is indeed less inclined to join the institution in stage 1b than in stage 1a. Eleven subjects even reduce the likelihood to join the institution to zero in stage 1b whenever the institution formation fails in stage 1a. Hence, behavior of most subjects in stage 1b of IF43 is potentially influenced by a weak form of inequity aversion. However, only those eleven out of 64 subjects, who do not join the institution in stage 1b any more, could be classified as truly inequity averse. Interestingly, we do not observe a similar reaction to the failure of institution formation across periods in IF3, where subjects are more likely to join the institution in period t if the institution formation failed in period $t-1$ than if it was successful (see the discussion below).

Table 5.5: Probit regression results for the probability to join the institution

	Join Institution			
	IF4	IF3	IF43-1a	IF43-1b
Join1a_t				1.772***
				(0.225)
Join1a_{t-1}	0.853***	1.210***	0.594***	0.231
	(0.282)	(0.220)	(0.228)	(0.253)
Join1b_{t-1}			0.013	1.140***
			(0.226)	(0.301)
Inst$_{t-1}$	0.406*	-0.669^{**}	-0.005	-0.025
	(0.233)	(0.284)	(0.230)	(0.255)
Inst3$_{t-1}$		-0.266	-0.284	-1.308^{***}
		(0.253)	(0.291)	(0.357)
Const.	-0.607	-4.587^{***}	-0.375	-2.305^{*}
	(1.306)	(1.514)	(1.391)	(1.319)
Observations	$N = 612$	$N = 576$	$N = 576$	$N = 388$

Notes: The decision not to join the institution is baseline. Join1a_t is a dummy variable for IF43 that takes the value 1 if the subject joined the institution in stage 1a of the current period and is 0 otherwise. Join1a_{t-1} is a dummy variable that takes the value 1 if the subject joined the institution in the previous period (in stage 1a of the previous period for IF43) and is 0 otherwise. Join1b_{t-1} is a dummy variable for IF43 that takes the value 1 if the subject joined the institution in stage 1b of the previous period and is 0 otherwise. Inst$_{t-1}$ is a dummy variable that takes the value 1 if an institution was implemented in the previous period and is 0 otherwise. Inst3$_{t-1}$ is a dummy variable that takes the value 1 if a 3-player institution was implemented in the previous period and is 0 otherwise. Std. errors in parentheses adjusted for group clusters. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

To complement our analysis of the subjects' individual participation decisions, we performed separate probit regressions for all treatments. The regression results are summarized in Table 5.5. As treatment IF43 has two participation stages, we performed regressions for each of the possible stages $1a$ and $1b$. In all treatments we find a strong persistence of behavior, i.e. the decision to join an institution is positively correlated over time. For IF43 this property needs to be specified for the different stages, though. Here, only the decisions in stage $1a$ ($1b$) are positively correlated over time, while there is no correlation between the decision in stage $1a$ in period $t-1$ and the decision in stage $1b$ in period t. Similarly, there is no correlation between the decision in stage $1b$ in period $t-1$ and the decision in stage $1a$ in period t. Finally, in IF43 the decision to join the institution in stage $1a$ and in stage $1b$ are positively correlated.

Furthermore, we observe that the formation of an institution in period $t-1$ influences the subjects' decision to join an institution in period t. In IF4 the formation of an institution in $t-1$ has a weakly significant and positive impact on the probability to join an institution in period t, while in IF3 the impact is strongly significant and negative. In IF43 only the formation of a 3-player institution in $t-1$ has a significant and negative impact. We included a dummy variable for the membership in a 3-player institution in the previous period in the regressions for IF3 and IF43. However, this variable and the dummy variable for a 3-player institution in the previous period are highly correlated ($|\rho| \geq 0.79$ in each treatment) and we therefore only report the dummy variable for a 3-player institution in the regressions.

5.4.4 Learning

In order to determine if there is any learning over the ten periods, we used the Spearman-Rank-Order test. We find that there is no correlation between contributions to the public good and the periods in IF4 (Spearman-$\rho = 0.4$) and IF43 ($\rho = -0.2$). Yet, we find that there is a small negative correlation (10% significance, $\rho = -0.6$) for IF3. However, this significance vanishes if we exclude period 1. For the number of institutions formed per period, there is a slightly significant positive correlation for IF4 ($\rho = 0.55$, 10% significance). The other two treatments show no correlations ($\rho = -0.27$ for IF3 and $\rho = -0.01$ for IF43). Furthermore, there is no significant correlation between the number of choices to join an institution and the period. All results are quite robust against start-game and end-game effects, i.e. we obtain the same results if we only consider periods 3 through 8 or 4 through 7.

Result 5.7 *In no treatment there is a significant change in behavior over time.*

5.4.5 Incentives to Opt for Weaker Minimum Participation Rules

Finally, we were interested in the empirical incentives for subjects to free ride. More specifically, while Result 5.1 shows that on average players' payoffs are highest in IF4 and IF43, i.e. whenever the institution formation game starts with a strict minimum participation requirement, one may ask whether nevertheless certain player behavior generates higher payoffs under a weaker minimum participation requirement. In particular, it could be that a player who intends to free ride on the institution formation of others earns a higher payoff in IF3 and in IF43 than in IF4. However, it turns out that this is not the case. The average payoff of a player in IF4 is 28.19. Not surprisingly, the payoff conditional on joining the institution in IF4 is even higher, namely 29.00. By comparison, the average payoff of a player who does not join the institution (a free rider) is 27.09 in IF3 and 27.23 in IF43. Although the differences in payoffs are not significant, this shows that free-riders do not profit from the weakening of the minimum participation requirement in IF3 or IF43.

Result 5.8 *There is no significant difference between the average payoff of a free rider in* IF3 *and* IF43 *and a player in* IF4.

5.5 Discussion

In the following, we discuss the data from our experiment in light of the theoretical results for players with social preferences derived in Section 5.2.3. More precisely, we will use four rules to estimate the distribution of player types from our data. As the three types are best distinguishable in their attitude towards the formation of 3-player institutions and the contributions of nonmembers when a 3-player institution has formed, these rules are based on the stages when the minimum participation constraint is 3, i.e. IF3 and stage 1*b* of IF43. We apply the following four rules:

1. Subjects who choose to join the institution in a stage with minimum participation constraint 3 even though there was a nonmember in a previous period who contributed nothing are classified as type 1; these subjects reveal weak aversion against disadvantageous inequality.

2. Subjects who choose not to join the institution in a stage with minimum participation constraint 3 in all periods after observing a nonmember contributing less than her full endowment in a previous period are classified as type 2 or 3; these subjects reveal strong aversion against disadvantageous inequality.

3. Subjects who contribute a positive amount to the public good when being nonmembers of 3-player institutions are classified as type 3; these subjects reveal a noticeable aversion against advantageous inequality as their contributions are a costly act to reduce the inequality.

4. Subjects in groups which repeatedly form 4-player institutions in a stage with minimum participation constraint 3 are classified as type 3; these subjects reveal a noticeable aversion against advantageous inequality as they repeatedly pass over the opportunity to free ride on the contributions of others.

As we have seen, with social preferences the equilibrium predictions are very sensitive to the type distribution within a group. In order to draw conclusions about the type distribution from observed behavior, we therefore have to assume that players form correct beliefs about the types of other group members from their actions. Furthermore, in order to allow classifications by the first two rules, we also have to assume that a subject expects the same subject to be nonmember in future 3-player institutions. We are aware of the fact that these are rather strong assumptions. However, it is the only way to match the theoretical discussion with the data.

To begin with, we consider treatment IF3. Here, we are able to identify 37 subjects as type 1. Moreover, there are 3 subjects who may be classified as either type 2 or 3, and 13 subjects can be classified as type 3 as they contribute a positive amount to the public good when being nonmember. However, 5 of the latter subjects can also be classified as type 1 based on their behavior. As these subjects show behavior which is inconsistent / inexplicable within our model of social preferences, we label them as "no type". Last, we classify the members of one group which repeatedly (in 6 out of 10 periods) forms 4-player institutions as type 3.[17] The remaining 13 subjects cannot be classified with the rules presented above. See Appendix 5.B for a detailed breakdown of the classification.

In a second step, we consider treatment IF43. The first thing to note is that the analysis reveals a much higher percentage of unclassifiable subjects. This is due to the fact that there is less data from stages with participation constraint 3 in IF43 than in IF3. Yet, we can still classify 15 subjects as type 1 and 7 subjects of type 2 or 3. Furthermore, 8 subjects contribute positive amounts as nonmembers and are, thus, classified as type 3. Again, there is one subject who is identified as both type 1 and 3 and whose behavior cannot be explained with the model. This subject is classified as "no type". In addition, we classify 3 more subjects as type 3 because their group repeatedly

[17] There is another group that formes a 4-player institution more than once. However, we consider this group to display a coordination problem rather than all players being of type 3 because several players can also be classified as type 1.

Table 5.6: Estimation of the type distribution in IF3 and IF43

	IF3	**IF**43
type 1	32 (50%)	14 (22%)
type 2 **or** 3	3 (5%)	7 (11%)
type 3	11 (17%)	10 (16%)
no type	5 (8%)	1 (2%)
unknown	13 (20%)	32 (50%)

Notes: The total number of subjects is 64 in both treatments, percentages in parentheses.

forms 4-player institutions in stage 1*b*. The remaining 32 subjects cannot be classified according to our criteria. Again, we provide a detailed overview over the classification of subjects in Appendix 5.B. Table 5.6 summarizes the analysis for both IF3 and IF43.

As we see from Table 5.6, behavior in our experiment is largely consistent with subjects having standard or social preferences of the Fehr and Schmidt (1999) type. Only few subjects (8% in IF3 and 2% in IF43) show behavior that is not consistent with either behavioral model. Interestingly, about 70% (45%) of the subjects in IF3 (IF43) who can be classified according to our rules are of type 1, i.e. either have standard preferences or are only weakly inequality averse.[18]

Hence, the results of our experiment are to a large extent not driven by a strong inequality aversion of the subjects. In particular, the observed failure to form institutions under a weaker minimum participation requirement is in most cases not due to inequality aversion.

5.6 Conclusion

In this paper, we have studied the effectiveness of different minimum participation rules for the formation of an institution to provide a public good. Due to a multiplicity of equilibria it is not clear which rule is optimal from a theoretical point of view: All minimum participation rules allow for a subgame perfect Nash equilibrium, where the formation of an institution fails and hence welfare is minimized.

In our experiment overall efficiency is highest in those treatments that start with or are restricted to the strict unanimity rule. In particular, the intuitive conjecture, that in practice a weaker minimum participation requirement increases the number of

[18] The lower percentage for IF43 may simply be due to the fact that there are many more subjects in IF43 than in IF3 who cannot be classified according to our rules, because the game never reaches stage 1b.

institutions and hence increases efficiency is not confirmed by our laboratory experiment. In fact, the total number of institutions is independent of the minimum participation rule. While more large institutions are formed under the strict rule that requires the participation of all players, small (3-player) institutions are frequently formed if the minimum participation requirement is weakened either from the very beginning or else in a second stage. This result stands in contrast to the findings by Kosfeld et al. (2009) who report that most subjects are not willing to implement institutions that do not involve all group members. While their results can only be explained with a large proportion of inequity averse subjects, our results are mostly consistent with standard preferences.

While we are hesitant to draw any bold policy conclusions from our small scale laboratory experiment, we nevertheless believe that our results support the idea that using a strict minimum participation rule is best whenever it comes to the implementation of a policy which is a Pareto improvement over the unregulated noncooperative outcome with no or too low contributions to the public good. In particular, the fear that a strict rule exacerbates the implementation of an institution in this case appears to be unwarranted.

Appendix

5.A Proofs

Proof of Proposition 5.1

Let $u_i = \pi_i$ for all players i. We use backwards induction in order to characterize the set of pure strategy subgame perfect Nash equilibria.

(i) Let $ma > 1$. From $a < 1$ it follows that $m \geq 2$. Consider stage 2 of game IFm. For nonmembers of an institution it is always a dominant strategy to contribute zero to the public good. Hence, there is a unique Nash equilibrium in stage 2, where player i contributes $g_i^* = w$ if an institution was implemented and i joined the institution in stage 1, and she contributes $g_i^* = 0$ otherwise. Next, consider stage 1 of the game. Suppose that no more than $m-2$ players join the institution in stage 1. Then no institution is implemented and all players i get $u_i = w$. An individual deviation of one player in stage 1 does not change the outcome since at least m players have to join the institution for it to be implemented. Hence, any strategy profile, where at most $m-2$ players join the institution in stage 1 and players' contributions in stage 2 are as derived before, is a subgame perfect Nash equilibrium.

Suppose exactly m players join the institution in stage 1. Then an institution is implemented and player i gets $u_i = maw$ if i joined the institution and $u_i = (1+ma)w$ if i did not join the institution and $m < n$. Suppose i joined the institution. Then, if i deviates and does not join the institution, no institution is implemented and i gets $w < maw$ which is worse. Suppose i did not join the institution. Then, if i deviates and joins the institution, it is implemented and i gets $(m+1)aw < (1+ma)w$ since $a < 1$. Hence, i is worse off after the deviation. Therefore, any strategy profile, where exactly m players join the institution in stage 1 and players' contributions in stage 2 are as derived before, is a subgame perfect Nash equilibrium.

Finally, suppose $M > m$ players join the institution in stage 1. Then an institution is implemented and player i gets $u_i = Maw$ if i joined the institution and $u_i = (1+Ma)w$ if i did not join the institution and $M < n$. Suppose i joined the institution. Since $M > m$, if i deviates and does not join the institution, it is still implemented and i gets $(1+(M-1)a)w > Maw$ since $a < 1$. Hence, there exists no subgame perfect Nash equilibrium, where more than m players join the institution in stage 1. This proves the claim that in any subgame perfect Nash equilibrium of IFm either an institution with exactly m members is implemented or no institution is implemented.

(ii) Let $m > k$ and $ka > 1$. From $a < 1$ it follows that $k \geq 2$. Consider stage 2 of game IFmk. Similar to the proof of (i) it is straightforward to show that there is a unique Nash equilibrium in stage 2, where player i contributes $g_i^* = w$ if an institution was implemented and i is a member of the institution, and she contributes $g_i^* = 0$ otherwise. Next, consider stage 1b of the game. Similar to the proof of (i) one can show that in any subgame perfect Nash equilibrium of the subgame starting in stage 1b, either exactly k players join the institution and an institution is implemented or at most $k-2$ players join the institution and no institution is implemented.

Next, consider stage 1a of the game and consider the continuation equilibrium, where in stage 1b an institution with exactly k members is implemented. Suppose that no more than $m-2$ players join the institution in stage 1a. Then the game moves to stage 1b, where an institution with exactly k members is implemented. All members i of the institution get $u_i = kaw$ and all nonmembers j get $u_j = (1 + ka)w$. An individual deviation of one player in stage 1a does not change the outcome. Hence, any strategy profile, where at most $m-2$ players join the institution in stage 1a, exactly k players join the institution in stage k and players' contributions in stage 2 are as derived before, is a subgame perfect Nash equilibrium.

Consider now the continuation equilibrium, where in stage 1b no institution is implemented. Suppose exactly m players join the institution in stage 1a. Then an institution with exactly m members is implemented and members i get $u_i = maw$ and nonmembers get $u_i = (1 + ma)w$. Obviously, a player who did not join the institution in stage 1a cannot improve by joining it because this would change her payoff to $(m+1)aw < (1+ma)w$. Also, a player who joined the institution in stage 1a cannot improve by leaving it because this would change her payoff to $w < maw$ since no institution will be implemented in that case. Hence, any strategy profile, where exactly m players join the institution in stage 1a and at most $k-2$ players join the institution in stage 1b, is a subgame perfect Nash equilibrium.

It is straightforward to verify that any strategy profile, where at most $m-2$ players join the institution in stage 1a and at most $k-2$ players join the institution in stage 1b, and contributions in stage 2 are as derived before, is a subgame perfect Nash equilibrium, where no institution is implemented.

Finally, similar to the proof of (i) one can show that there exists no subgame perfect Nash equilibrium, where more than m players join the institution in stage 1a. This proves the claim that in any subgame perfect Nash equilibrium of IFmk either an institution with exactly m members is implemented in stage 1a or an institution with exactly k members is implemented in stage 1b or no institution is implemented.

□

Proof of Proposition 5.2

$u_i = \pi_i$ for all players i and let $ma > 1$. By Proposition 5.1 in any subgame perfect Nash equilibrium of IFm either an institution with exactly m members is implemented or no institution is implemented. Consider a subgame perfect Nash equilibrium, where exactly m players join the institution in stage 1, and in stage 2 nonmembers of the institution contribute zero to the public good. Consider stage 2. Obviously, contributing $g_i^* = 0$ is a strictly dominant strategy for all nonmembers of the institution. Consider stage 1. If a member of the institution deviates and does not join the institution, then her payoff strictly decreases from maw to w. If a nonmember deviates and joins the institution, then her payoff strictly decreases from $(1 + ma)w$ to $(m + 1)aw$. Hence, in every stage game, every player's strategy is a unique best response to the equilibrium strategies of the other players and hence, the subgame perfect Nash equilibrium, where an institution with exactly m members is implemented, is stagewise strict.

By contrast, the subgame perfect Nash equilibrium, where at most $m - 2$ players join the institution in stage 1, is not stagewise strict, since in stage 1 every player is indifferent between joining and not joining the institution given the equilibrium strategies of the other players.

□

Proof of Proposition 5.2

$u_i = \pi_i$ for all players i and let $ma > 1$. By Proposition 5.1 in any subgame perfect Nash equilibrium of IFm either an institution with exactly m members is implemented or no institution is implemented. Consider a subgame perfect Nash equilibrium, where exactly m players join the institution in stage 1, and in stage 2 nonmembers of the institution

contribute zero to the public good. Consider stage 2. Obviously, contributing $g_i^* = 0$ is a strictly dominant strategy for all nonmembers of the institution. Consider stage 1. If a member of the institution deviates and does not join the institution, then her payoff strictly decreases from maw to w. If a nonmember deviates and joins the institution, then her payoff strictly decreases from $(1 + ma)w$ to $(m + 1)aw$. Hence, in every stage game, every player's strategy is a unique best response to the equilibrium strategies of the other players and hence, the subgame perfect Nash equilibrium, where an institution with exactly m members is implemented, is stagewise strict.

By contrast, the subgame perfect Nash equilibrium, where at most $m - 2$ players join the institution in stage 1, is not stagewise strict, since in stage 1 every player is indifferent between joining and not joining the institution given the equilibrium strategies of the other players.

□

Proof of Proposition 5.3

Consider stage 2 first. If no institution was formed in stage 1, and if all players are of type 3 ($\beta_i \geq 0.6$), then from Fehr and Schmidt (1999, Prop. 4) it follows that any contribution profile with $g_i = g \in [0, w]$ for all i is a Nash equilibrium in stage 2. If instead, there exists at least one player i of type 1 or 2 ($\beta_i < 0.6$), then again by Fehr and Schmidt (1999, Prop. 4) it follows that there exists a unique Nash equilibrium in stage 2, where all players contribute zero to the public good.

Next, consider stage 1. By the same argument as in the proof of Proposition 5.1 it follows that there always exists a subgame perfect Nash equilibrium, where no institution is formed. Suppose now that all players join the institution in stage 1. Then, the institution is implemented and very player's utility is $u_i = 1.6w$. If one player deviates and does not join the institution, then no institution is implemented and by the equilibrium contribution profile(s) derived above, the deviating player does not earn more utility than $1.6w$ since all Nash equilibria in stage 2 are symmetric. Hence, there exists a subgame perfect Nash equilibrium, where a four member institution is implemented.

□

Proof of Proposition 5.4

Consider stage 2 first. Suppose no institution was implemented in stage 1. Then, by the same argument as in the proof of Proposition 5.3, if all players are of type 3, any contribution profile with $g_i = g \in [0, w]$ for all i is a Nash equilibrium in stage 2. If instead, there exists at least one player of type 1 or 2 then there exists a unique Nash equilibrium in stage 2, where all players contribute zero to the public good. Suppose next that a 3-player institution was implemented in stage 1. Then in equilibrium, the nonmember i contributes $g_i = 0$, if $\beta_i < 0.6$, $g_i = w$, if $\beta_i > 0.6$, and any amount $g_i \in [0, w]$, if $\beta_i = 0.6$.

Next, consider stage 1. Suppose all players join the institution in stage 1. Then, the institution is implemented and every player i's utility is $u_i = 1.6w$. If player i deviates and does not join the institution, a 3-player institution is implemented. In this case, if player i is of type 1 or 2 ($\beta_i < 0.6$), he contributes zero to the public good and receives utility $u_i = 2.2w - \beta_i w > 1.6w$. Hence, if at least one player is of type 1 or 2, no 4-player institution is formed in a subgame perfect Nash equilibrium. If the deviating player i is of type 3 instead, he either contributes $g_i = w$ (if $\beta_i > 0.6$) or any amount $g_i \in [0, w]$ (if $\beta_i = 0.6$) as a nonmember, which in both cases yields the same utility $u_i = 1.6w$ as in a 4-player institution. Hence, if all players are of type 3, the implementation of a 4-player institution is supported as a subgame perfect Nash equilibrium. This proves (i).

Suppose now that exactly three players join the institution in stage 1. Then, the nonmember i receives utility $u_i = 2.2w - \beta_i w > 1.6w$ if i is of type 1 or 2, and he receives $u_i = 1.6w$ if he is of type 3. Hence, in both cases the nonmember cannot improve by joining the institution. Suppose now that no player is of type 3 and at least two players are of type 2. In this case, at least one member of a 3-player institution is of type 2. Since no player is of type 3, the nonmember contributes zero and all members i receive utility $u_i = 1.2w - \frac{\alpha_i}{3}w$. For the type 2 member, $\alpha_i > 0.6$ and hence, $u_i < w$, where w is i's utility if no institution is formed and all players contribute zero in the unique Nash equilibrium (there are no type 3 players). Therefore, if no player is of type 3 and at least two players are of type 2, there exists no subgame perfect Nash equilibrium, where a 3-player institution is implemented.

If, instead, at least one player is of type 3, then it is easy to see that the following strategy profile is a subgame perfect Nash equilibrium: The type 3 player i does not join the institution in stage 1, while all other players join the institution. As a nonmember of a 3-player institution the type 3 player contributes $g_i = w$. Apart from this specification the contributions of all players as nonmembers of a 3-player institution and in the case where no institution was implemented in stage 1 are such that they satisfy the conditions derived in the first part of this proof.

Finally, consider the case, where at least three players are of type 1. Then it is easy to see that the following strategy profile is a subgame perfect Nash equilibrium: Three type 1 players join the institution while the remaining player does not join the institution. The contributions of all players as nonmembers of a 3-player institution and in the case where no institution was implemented in stage 1, are such that they satisfy the conditions derived in the first part of this proof. This proves (ii).

By the same argument as in the proof of Proposition 5.1 it follows that there always exists a subgame perfect Nash equilibrium, where no institution is implemented. The equilibrium contributions are as derived in the first part of this proof. From (i) and (ii) it then follows that no institution is the unique subgame perfect Nash equilibrium outcome if at least two players are of type 2 and no player is of type 3. This proves (iii).

□

Proof of Proposition 5.5

(i) The implementation of a 4-player institution in stage 1a can be supported as a subgame perfect Nash equilibrium as follows: All players join the institution in stage 1a. If no institution was implemented in stage 1a, then no player joins the institution in stage 1b. Equilibrium contributions in stage 2 are such that they satisfy the conditions derived in the first part of the proof of Proposition 5.4. To verify that this strategy profile is a subgame perfect Nash equilibrium one can use the same proof as in Proposition 5.3.

(ii) Suppose all players are of type 3. Then, similar to the proof of Proposition 5.4 (i) one can verify that the following strategy profile is a subgame perfect Nash equilibrium, where a 4-player institution is implemented in stage 1b: In stage 1a no player joins the institution and in stage 1b all players join the institution. Whenever no institution is implemented in stage 1a or 1b, all players i contribute $g_i = g \in [0, w]$. Whenever a 3-player institution is implemented in stage 1b, the nonmember contributes $g_i = w$, if $\beta_i > 0.6$ and he contributes any amount $g_i \in [0, w]$, if $\beta_i = 0.6$. Also, the formation of a 4-player institution in stage 1b is not supported as a subgame perfect Nash equilibrium if at least one player i is of type 1 or 2, since i can always improve by deviating in stage 1b an not joining the institution.

(iii) Similar to the proof of Proposition 5.4 (ii) one can verify that the implementation of a 3-player institution in stage 1b is supported as a subgame perfect Nash equilibrium if and only if either at least one player is of type 3 or at least three players are of type 1. The following strategy profile supports this equilibrium outcome: No player joins the institution in stage 1a. If at least one player i is of type 3, then i does not join the institution in stage 1b, while all other players join the institution in stage 1b. As a non-member of a 3-player institution the type 3 player i contributes $g_i = w$ in stage 2. Apart from this specification the contributions of all players as nonmembers of a 3-player institution and in the case where no institution was implemented in stage 1 are such that they satisfy the conditions derived in the first part of the proof of Proposition 5.4. If at least three players are of type 1, then three type 1 players join the institution in stage 1b, while the remaining player does not join the institution. The contributions of all players as nonmembers of a 3-player institution and in the case where no institution was implemented in stage 1, are such that they satisfy the conditions derived in the first part of the proof of Proposition 5.4.

(iv) It is straightforward to verify that the following strategy profile is a subgame perfect Nash equilibrium, where no institution is implemented: No player joins the institution in stage 1a and in stage 1b. are such that they satisfy the conditions derived in the first part of the proof of Proposition 5.4.

□

5.B Type Classifications

The following four tables depict the detailed type classifications from Section 5.5. The tables may be split in five parts: subject identification, institution formation, individual participation decisions, nonmember contributions and classification.

- Subject identification: Columns one and two comprise the group and subject number, respectively.
- Institution formation: The columns labeled by "4" and "3" in IF3 contain the number of times an institution with the respective size was formed in the group. Tables for IF43 feature three columns for institution formation. Size four institutions are split such that there is one column for size four institutions in stage 1a (labeled "4a") and another one for stage 1b ("'4b").
- Individual participation decisions: The "Join" column in IF3 tables presents the number of times that the subject chose to join the institution. In IF43, there is a "Join-a" column for decisions in stage 1a and a "Join-b" column for decisions in stage 1b.
- Nonmember contributions: Column "NM" captures the number of times a size three institution was formed where this subject was a nonmember. Column "Contributions" then presents the contributions of this subject for each single case. Contributions are sorted by order of occurrence, separated by commas.
- Classification: The "Classification" column contains the final classification of the subject. A "?" symbolizes that we were not able to classify the subject according to our classification rules. In the "Justification" column we present the rules according to which a subject was classified.

Table 5.6: Type classifications in IF3, groups 1 - 8.

IF3		4	3	Join	NM	Contributions	Classification	Justification
Group 1	S1	-	5	9	-		Type 1	Rule 1
	S2			-	5	1,0,10,1,1	Type 3	Rule 3
	S3			10	-		Type 1	Rule 1
	S4			5	-		Type 1	Rule 1
Group 2	S1	1	3	2	2	2,0	No type	Inconsistent: Rules 1 and 3
	S2			4	1	0	Type 1	Rule 1
	S3			7	-		Type 1	Rule 1
	S4			10	-		Type 1	Rule 1
Group 3	S1	1	6	10	-		Type 1	Rule 1
	S2			5	2	0,0	Type 1	Rule 1
	S3			3	4	0,0,10,15	No type	Inconsistent: Rules 1 and 3
	S4			10	-		Type 1	Rule 1
Group 4	S1	-	3	-	3	0,0,0	?	
	S2			10	-		Type 1	Rule 1
	S3			4	-		Type 1	Rule 1
	S4			7	-		Type 1	Rule 1
Group 5	S1	-	6	-	6	10,20,0,5,0,0	Type 3	Rule 3
	S2			8	-		Type 1	Rule 1
	S3			7	-		Type 1	Rule 1
	S4			10	-		Type 1	Rule 1
Group 6	S1	1	1	10	-		Type 1	Rule 1
	S2			3	-		Type 2 or 3	Rule 2
	S3			4	-		Type 2 or 3	Rule 2
	S4			4	1	0	?	
Group 7	S1	2	4	7	-		Type 1	Rule 1
	S2			4	3	0,0,0	?	
	S3			6	1	10	No type	Inconsistent: Rules 1 and 3
	S4			10	-		Type 1	Rule 1
Group 8	S1	1	5	8	-		Type 1	Rule 1
	S2			3	3	0,0,0	Type 1	Rule 1
	S3			9	-		Type 1	Rule 1
	S4			5	2	0,0	Type 1	Rule 1

Table 5.7: Type classifications in IF3, groups 9 - 16.

IF3								
		4	3	Join	NM	Contributions	Classification	Justification
Group 9	S1	-	3	9	-		Type 1	Rule 1
	S2			7	-		Type 1	Rule 1
	S3			-	3	0,0,0	?	
	S4			5	-		Type 1	Rule 1
Group 10	S1	1	5	4	2	10,1	Type 3	Rule 3
	S2			10	-		?	
	S3			10	-		?	
	S4			3	3	1,1,1	Type 3	Rule 3
Group 11	S1	-	4	2	2	5,5	No type	Inconsistent: Rules 1 and 3
	S2			3	1	0	?	
	S3			7	1	5	No type	Inconsistent: Rules 1 and 3
	S4			10	-		Type 1	Rule 1
Group 12	S1	1	1	2	-		Type 2 or 3	Rule 2
	S2			1	1	0	?	
	S3			9	-		Type 1	Rule 1
	S4			9	-		Type 1	Rule 1
Group 13	S1	6	2	9	-		Type 3	Rule 4
	S2			9	-		Type 3	Rule 4
	S3			10	-		Type 3	Rule 4
	S4			6	2	10,10	Type 3	Rules 3 and 4
Group 14	S1	-	-	-	-		?	
	S2			-	-		?	
	S3			9	-		?	
	S4			6	-		?	
Group 15	S1	-	4	2	2	2,0	Type 3	Rule 3
	S2			10	-		Type 1	Rule 1
	S3			3	1	1	Type 3	Rules 2 and 3
	S4			5	1	1	Type 3	Rules 2 and 3
Group 16	S1	-	4	-	4	0,0,0,0	?	
	S2			5	-		Type 1	Rule 1
	S3			10	-		Type 1	Rule 1
	S4			7	-		Type 1	Rule 1

Table 5.8: Type classifications in IF43, groups 1 - 8.

IF43		4a	4b	3	Join-a	Join-b	NM	Contributions	Classification	Justification
Group 1	S1	5	-	-	10	1	-		?	
	S2				6	-	-		?	
	S3				9	5	-		?	
	S4				10	3	-		?	
Group 2	S1	4	-	2	10	5	-		Type 1	Rule 1
	S2				9	2	-		Type 1	Rule 1
	S3				4	-	2	0,0	?	
	S4				9	4	-		Type 1	Rule 1
Group 3	S1	9	-	-	9	-	-		?	
	S2				10	1	-		?	
	S3				10	-	-		?	
	S4				10	1	-		?	
Group 4	S1	-	-	4	7	8	-		Type 1	Rule 1
	S2				6	5	-		Type 1	Rule 1
	S3				10	10	-		Type 1	Rule 1
	S4				1	-	4	0,0,0,0	?	
Group 5	S1	4	-	1	4	1	-		Type 2 or 3	Rule 2
	S2				10	-	1	1	Type 3	Rule 3
	S3				10	5	-		Type 2 or 3	Rule 2
	S4				10	5	-		?	
Group 6	S1	3	-	3	5	2	1	5	No type	Inconsistent: Rules 1 and 3
	S2				10	7	-		Type 1	Rule 1
	S3				6	1	2	0,0	Type 2 or 3	Rule 2
	S4				10	6	-		Type 1	Rule 1
Group 7	S1	-	-	1	10	1	-		Type 2 or 3	Rule 2
	S2				-	-	1	2	Type 3	Rule 3
	S3				4	7	-		Type 2 or 3	Rule 2
	S4				10	10	-		?	
Group 8	S1	1	3	1	6	6	-		Type 3	Rule 4
	S2				4	3	1	8	Type 3	Rules 3 and 4
	S3				5	6	-		Type 3	Rule 4
	S4				8	9	-		Type 3	Rule 4

Table 5.9: Type classifications in IF43, groups 9 - 16.

IF43										
		4a	4b	3	Join-a	Join-b	NM	Contributions	Classification	Justification
Group 9	S1	1	-	5	2	-	5	1,1,1,1,1	Type 3	Rule 3
	S2				5	5	-		?	
	S3				10	7	-		?	
	S4				8	7	-		?	
Group 10	S1	1	-	3	9	7	-		?	
	S2				10	9	-		?	
	S3				6	3	-		?	
	S4				1	-	3	10,10,10	Type 3	Rule 3
Group 11	S1	8	-	-	10	2	-		?	
	S2				10	1	-		?	
	S3				9	-	-		?	
	S4				9	1	-		?	
Group 12	S1	5	-	3	9	4	-		Type 1	Rule 1
	S2				9	1	2	7,5	Type 3	Rules 2 and 3
	S3				6	2	1	0	?	
	S4				9	4	-		Type 1	Rule 1
Group 13	S1	-	-	2	-	-	2	0,0	?	
	S2				8	9	-		Type 1	Rule 1
	S3				4	3	-		Type 1	Rule 1
	S4				7	6	-		Type 1	Rule 1
Group 14	S1	3	-	1	8	5	-		Type 1	Rule 1
	S2				9	1	-		Type 2 or 3	Rule 2
	S3				4	3	1	0	?	
	S4				7	1	-		Type 2 or 3	Rule 2
Group 15	S1	9	-	-	10	1	-		?	
	S2				10	1	-		?	
	S3				9	-	-		?	
	S4				9	-	-		?	
Group 16	S1	-	-	2	9	6	-		?	
	S2				10	10	-		?	
	S3				-	-	2	5,5	Type 3	Rule 3
	S4				3	3	-		?	

5.C Experimental Instructions

In the following we present an English translation of the instructions and control questions for the treatments IF4, IF3 and IF43. Since the experiment was conducted in Germany, all embedded screenshots are in German.

5.C.1 Experimental instructions for IF4

General information

Welcome. You will now participate in a decision making experiment and thereby earn money. How much money you earn will depend on your decisions and the decisions of the other participants in the experiment. Your total earnings in the experiment will be calculated in tokens which will at the end be converted into Euro and then paid to you in cash without other participants knowing how much money you receive. The exchange rate is

3 Tokens = 1 Euro.

Your decisions in the experiment remain anonymous, i.e., no other participant receives information about your identity, neither during nor after the experiment. In the same way, you will not receive any information about the identities of the other participants. It is important that you fully understand these instructions. Please read the following pages carefully. If you have questions, please raise your hand and the experimenter will come to you and answer them. In order to make sure that you have understood the instructions correctly, we ask you to answer a set of control questions after the instruction phase.
During the experiment it is not allowed to communicate with any other participant. Please have your mobile phones turned off for the entire experiment. You are only allowed to use those features on the computer that are needed for the course of the experiment. Communication and misuse of the computer lead to an exclusion from the experiment and the loss of all earnings.

Experiment description

All participants are faced with the same decision problem and have the same decision options. At the beginning of the experiment the computer randomly partitions all participants into groups with four members. The group members remain anonymous, i.e.,

neither during nor after the experiment you will learn which other three persons are or were in your group. You remain in the same group for the entire experiment.
The experiment consists of **10 identical rounds**. You will receive an endowment of 20 tokens at the beginning of each round. Your task is to decide how many tokens you want to invest in a common project and how many tokens you want to keep for yourself. Your **total earnings** in each round consist of two parts: your **private earnings** and the **project earnings**. Your private earnings are your endowment of 20 tokens minus your contribution to the common project. The project earnings are the sum of contributions by all group members multiplied by 0.4. In an equation, your total earnings are represented as follows:

$$\textbf{Total earnings} = \underbrace{\textbf{20} - \textbf{own contribution}}_{\text{Private earnings}} + \underbrace{\textbf{0.4} * \textbf{sum of contributions}}_{\text{Project earnings}}$$

This implies that every token that you keep for yourself increases your total earnings by one token. Every token that you contribute to the project increases your total earnings by 0.4 tokens. At the same time, every token that another group member contributes to the project increases your total earnings by 0.4. The same is true vice versa. With every token that you contribute to the project, the total earnings of all other group members are increased by 0.4 tokens.
Before you invest in the project, you and the other group members can form a team. A team is formed only if all four group members decide to form a team. If a team is formed, the entire endowment of 20 tokens is automatically contributed to the project by each team member. In the following, we explain the sequence of decisions in detail.

Experiment procedure

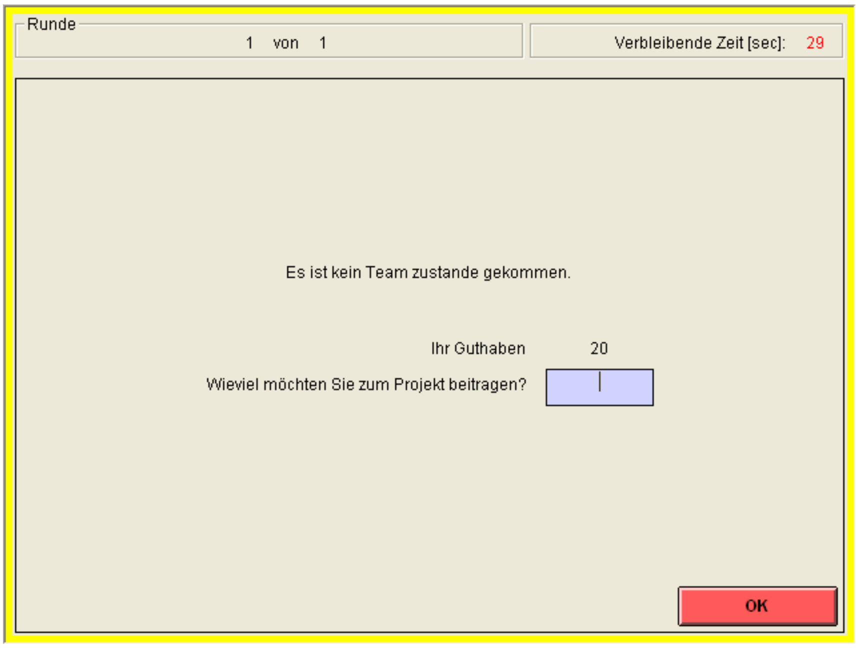
Each round of the experiment consists of two consecutive phases, the **joining phase** and the **investment phase**.

Joining phase:
In this phase you are asked if you want to join the team. On the screen you see:

By joining the team you irrevocably commit yourself to invest your entire endowment of 20 tokens in the following investment phase, if a team is formed. A team is only formed if all four group members join the team.

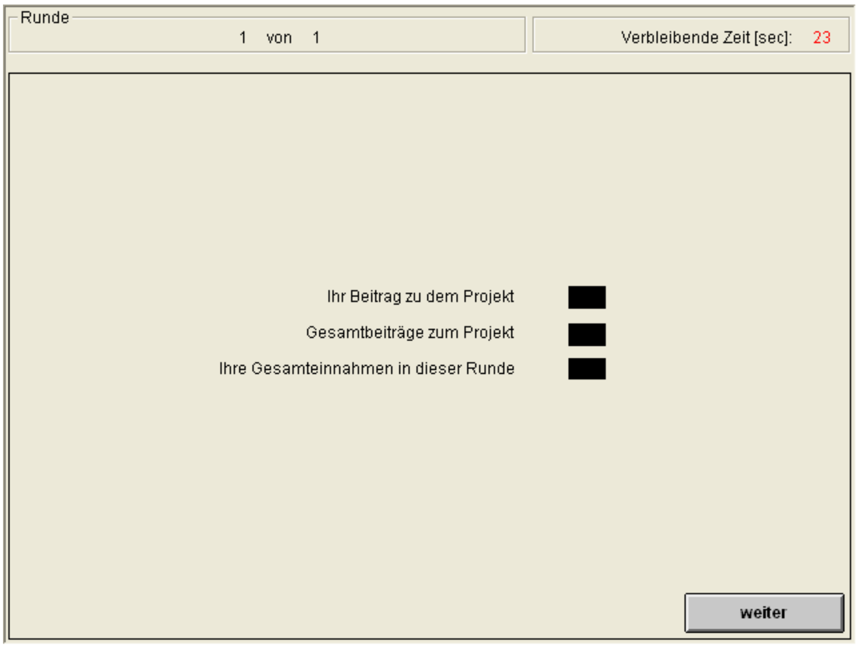
Investment phase:
If a team was formed, i.e., if all group members joined the team, then you do not take any decision in this phase. Your entire endowment of 20 tokens is automatically invested in the project and you see the screen:

Click on “OK” if you have read the information on the screen in order for the experiment to continue.
If no team was formed because not all group members joined the team, then you can now decide how many tokens you want to invest in the project. On the screen you see:



Enter your contribution to the project into the input box and click on "OK". Keep in mind that you can only enter integers 0, 1, 2, .., 20.

At the end of the investment phase the round ends with a screen informing you about your contribution to the project, the total contributions to the project by the whole group and your total earnings in this round (numbers are blackened).



Click on "continue" if you have read the information on the screen in order for the experiment to continue with the next round which is identical to the previous round. The experiment ends after the 10th round.

Payment at the end of the experiment

At the end of the ten rounds of the experiment one round is randomly drawn. Your total earnings in that round are converted from tokens into Euro and are paid to you in cash. Independent of your earnings you additionally receive a fixed payment of 4 Euro for your participation in the experiment.

5.C.2 Experimental instructions for IF3

General information

Welcome. You will now participate in a decision making experiment and thereby earn money. How much money you earn will depend on your decisions and the decisions of the other participants in the experiment. Your total earnings in the experiment will be calculated in tokens which will at the end be converted into Euro and then paid to you in cash without other participants knowing how much money you receive. The exchange rate is

3 Tokens = 1 Euro.

Your decisions in the experiment remain anonymous, i.e., no other participant receives information about your identity, neither during nor after the experiment. In the same way, you will not receive any information about the identities of the other participants. It is important that you fully understand these instructions. Please read the following pages carefully. If you have questions, please raise your hand and the experimenter will come to you and answer them. In order to make sure that you have understood the instructions correctly, we ask you to answer a set of control questions after the instruction phase.
During the experiment it is not allowed to communicate with any other participant. Please have your mobile phones turned off for the entire experiment. You are only allowed to use those features on the computer that are needed for the course of the experiment. Communication and misuse of the computer lead to an exclusion from the experiment and the loss of all earnings.

Experiment description

All participants are faced with the same decision problem and have the same decision options. At the beginning of the experiment the computer randomly partitions all participants into groups with four members. The group members remain anonymous, i.e., neither during nor after the experiment you will learn which other three persons are or were in your group. You remain in the same group for the entire experiment.
The experiment consists of **10 identical rounds**. You will receive an endowment of 20 tokens at the beginning of each round. Your task is to decide how many tokens you want to invest in a common project and how many tokens you want to keep for yourself. Your **total earnings** in each round consist of two parts: your **private earnings** and

the **project earnings**. Your private earnings are your endowment of 20 tokens minus your contribution to the common project. The project earnings are the sum of contributions by all group members multiplied by 0.4. In an equation, your total earnings are represented as follows:

$$\textbf{Total earnings} = \underbrace{\textbf{20} - \textbf{own contribution}}_{\text{Private earnings}} + \underbrace{\textbf{0.4} * \textbf{sum of contributions}}_{\text{Project earnings}}$$

This implies that every token that you keep for yourself increases your total earnings by one token. Every token that you contribute to the project increases your total earnings by 0.4 tokens. At the same time, every token that another group member contributes to the project increases your total earnings by 0.4. The same is true vice versa. With every token that you contribute to the project, the total earnings of all other group members are increased by 0.4 tokens.

Before you invest in the project, you and the other group members can form a team. A team is formed only if at least three group members decide to form a team. If a team is formed, the entire endowment of 20 tokens is automatically contributed to the project by each team member. In the following, we explain the sequence of decisions in detail.

Experiment procedure

Each round of the experiment consists of two consecutive phases, the **joining phase** and the **investment phase**.

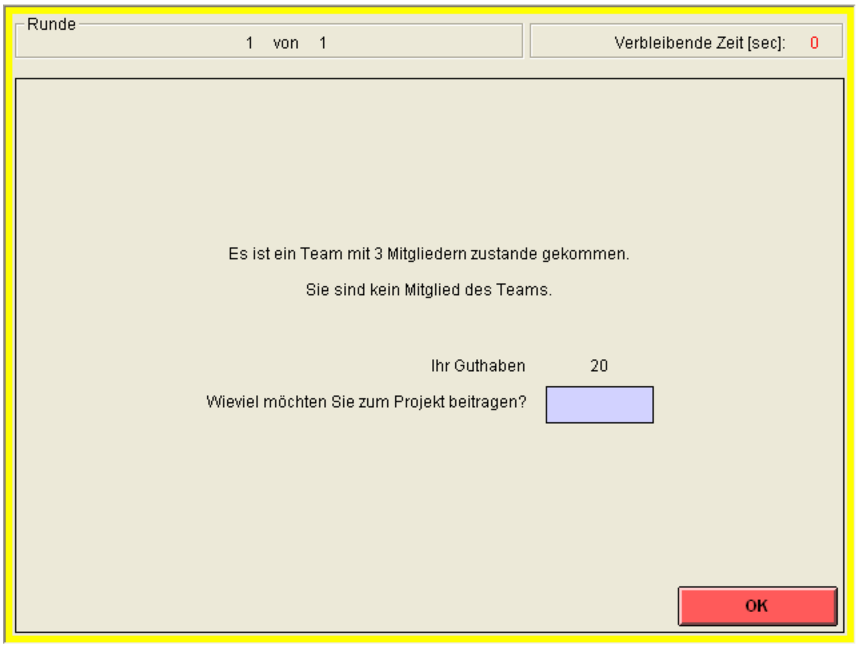
Joining phase:

In this phase you are asked if you want to join the team. On the screen you see:

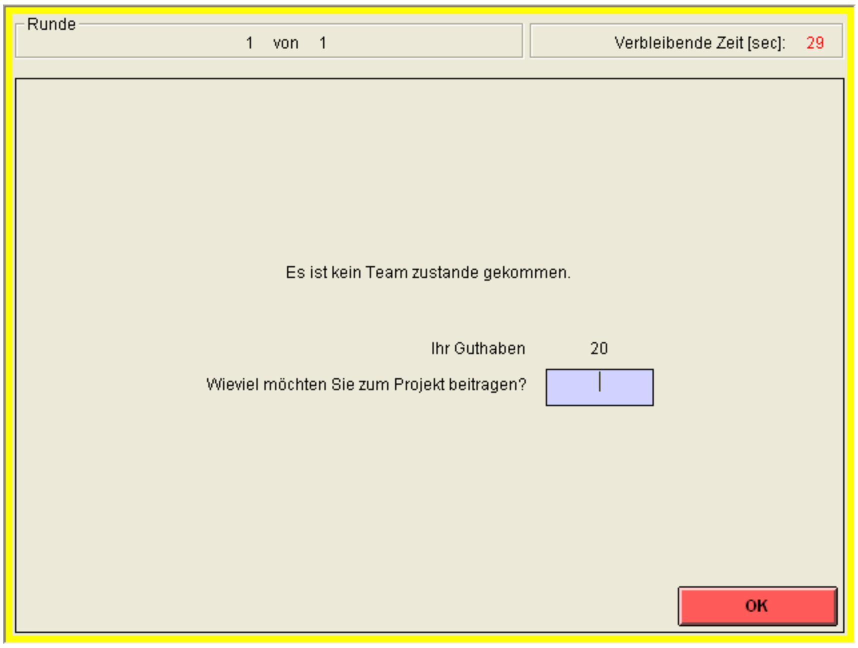
By joining the team you irrevocably commit yourself to invest your entire endowment of 20 tokens in the following investment phase, if a team is formed. A team is only formed if at least three group members join the team.

Investment phase:
If a team was formed, i.e., if at least three group members joined the team and you are in the team, then you do not take any decision in this phase. Your entire endowment of 20 tokens is automatically invested in the project and you see the screen:

Click on “OK” if you have read the information on the screen in order for the experiment to continue.
If a team was formed but you are not member of the team, then you can now decide how many tokens you want to invest in the project. On the screen you see:

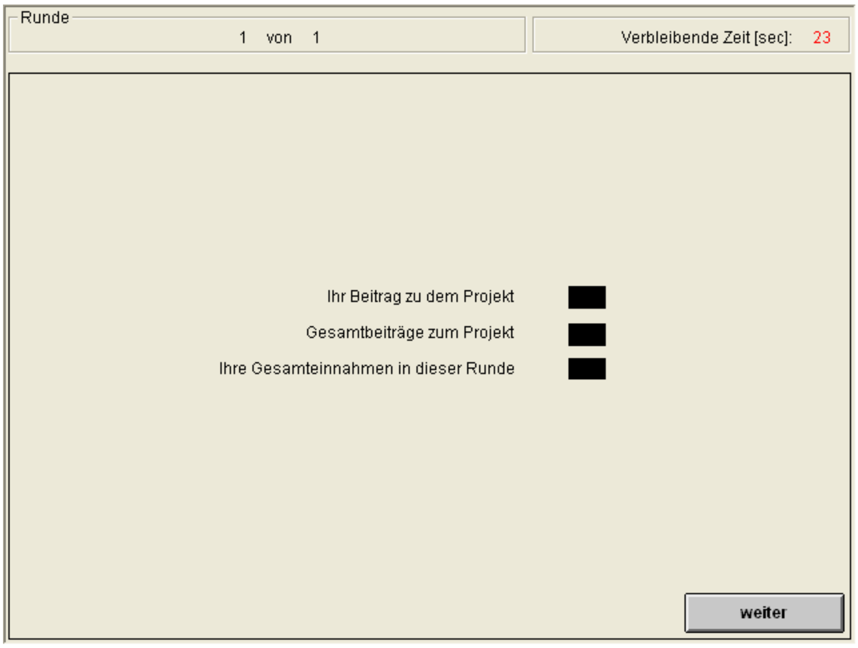


Enter your contribution to the project into the input box and click on “OK”. Keep in mind that you can only enter integers 0, 1, 2, .., 20.
If no team was formed because less than three group members joined the team, then you can now decide how many tokens you want to invest in the project. On the screen you see:



Enter your contribution to the project into the input box and click on "OK". Keep in mind that you can only enter integers 0, 1, 2, .., 20.

At the end of the investment phase the round ends with a screen informing you about your contribution to the project, the total contributions to the project by the whole group and your total earnings in this round (numbers are blackened).



Click on "continue" if you have read the information on the screen in order for the experiment to continue with the next round which is identical to the previous round. The experiment ends after the 10th round.

Payment at the end of the experiment

At the end of the ten rounds of the experiment one round is randomly drawn. Your total earnings in that round are converted from tokens into Euro and are paid to you in cash. Independent of your earnings you additionally receive a fixed payment of 4 Euro for your participation in the experiment.

5.C.3 Experimental instructions for IF43

General information

Welcome. You will now participate in a decision making experiment and thereby earn money. How much money you earn will depend on your decisions and the decisions of the other participants in the experiment. Your total earnings in the experiment will be calculated in tokens which will at the end be converted into Euro and then paid to you in cash without other participants knowing how much money you receive. The exchange rate is

3 Tokens = 1 Euro.

Your decisions in the experiment remain anonymous, i.e., no other participant receives information about your identity, neither during nor after the experiment. In the same way, you will not receive any information about the identities of the other participants. It is important that you fully understand these instructions. Please read the following pages carefully. If you have questions, please raise your hand and the experimenter will come to you and answer them. In order to make sure that you have understood the instructions correctly, we ask you to answer a set of control questions after the instruction phase.
During the experiment it is not allowed to communicate with any other participant. Please have your mobile phones turned off for the entire experiment. You are only allowed to use those features on the computer that are needed for the course of the experiment. Communication and misuse of the computer lead to an exclusion from the experiment and the loss of all earnings.

Experiment description

All participants are faced with the same decision problem and have the same decision options. At the beginning of the experiment the computer randomly partitions all participants into groups with four members. The group members remain anonymous, i.e., neither during nor after the experiment you will learn which other three persons are or were in your group. You remain in the same group for the entire experiment.
The experiment consists of **10 identical rounds**. You will receive an endowment of 20 tokens at the beginning of each round. Your task is to decide how many tokens you want to invest in a common project and how many tokens you want to keep for yourself. Your **total earnings** in each round consist of two parts: your **private earnings** and

the **project earnings**. Your private earnings are your endowment of 20 tokens minus your contribution to the common project. The project earnings are the sum of contributions by all group members multiplied by 0.4. In an equation, your total earnings are represented as follows:

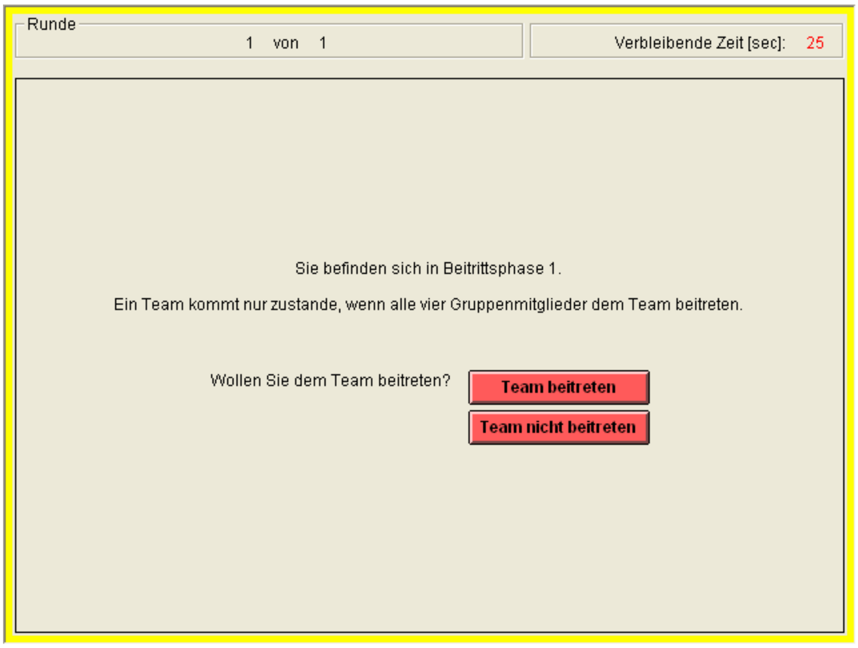
$$\textbf{Total earnings} = \underbrace{\textbf{20} - \textbf{own contribution}}_{\text{Private earnings}} + \underbrace{\textbf{0.4} * \textbf{sum of contributions}}_{\text{Project earnings}}$$

This implies that every token that you keep for yourself increases your total earnings by one token. Every token that you contribute to the project increases your total earnings by 0.4 tokens. At the same time, every token that another group member contributes to the project increases your total earnings by 0.4. The same is true vice versa. With every token that you contribute to the project, the total earnings of all other group members are increased by 0.4 tokens.
Before you invest in the project, you and the other group members can form a team. In the first step a team is formed only if all four group members decide to form a team. If no team is formed in this step, a team may be formed by three group members in a second step. If a team is formed, the entire endowment of 20 tokens is automatically contributed to the project by each team member. In the following, we explain the sequence of decisions in detail.

Experiment procedure

Each round of the experiment consists of three consecutive phases, the **joining phases 1 and 2** and the **investment phase**.

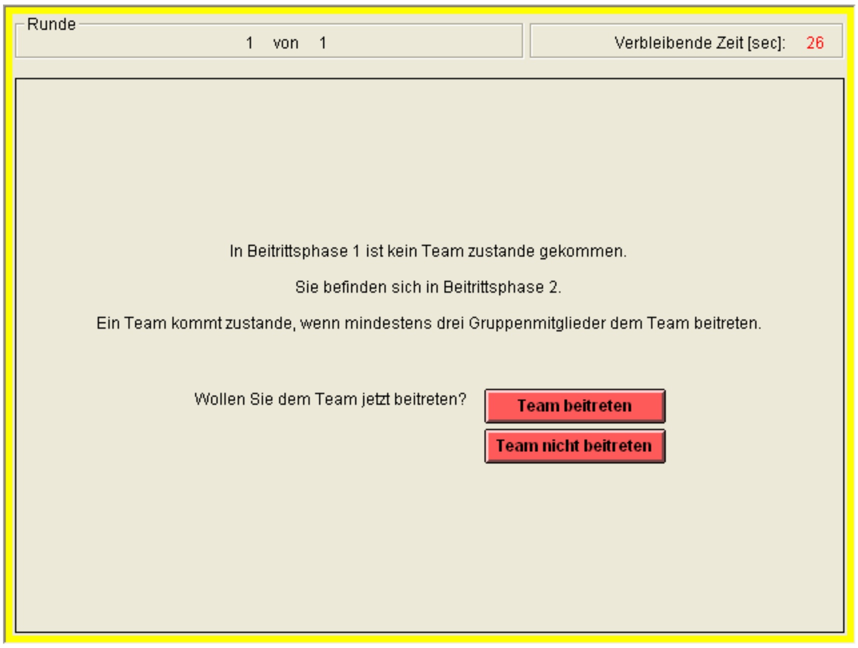
Joining phase 1:
In this phase a team is only formed if all four group members join the team. You are asked if you want to join the team. On the screen you see:



By joining the team you irrevocably commit yourself to invest your entire endowment of 20 tokens in the following investment phase, if a team is formed.

Joining phase 2:

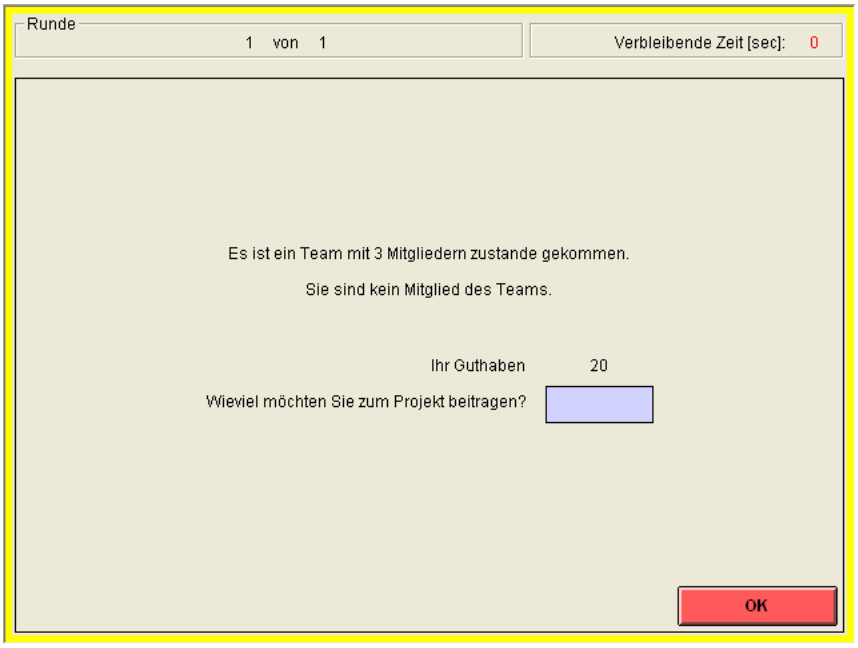
If a team was formed in joining phase 1 then this phase is skipped and the round continues directly with the investment phase. If no team was formed in joining phase 1, i.e., if not all group members decided to join the team, then in this phase you are again asked if you want to join the team. Your decision from joining phase 1 has no influence on your decision alternatives in joining phase 2. In this phase a team is formed if at least three group members join the team. On the screen you see:



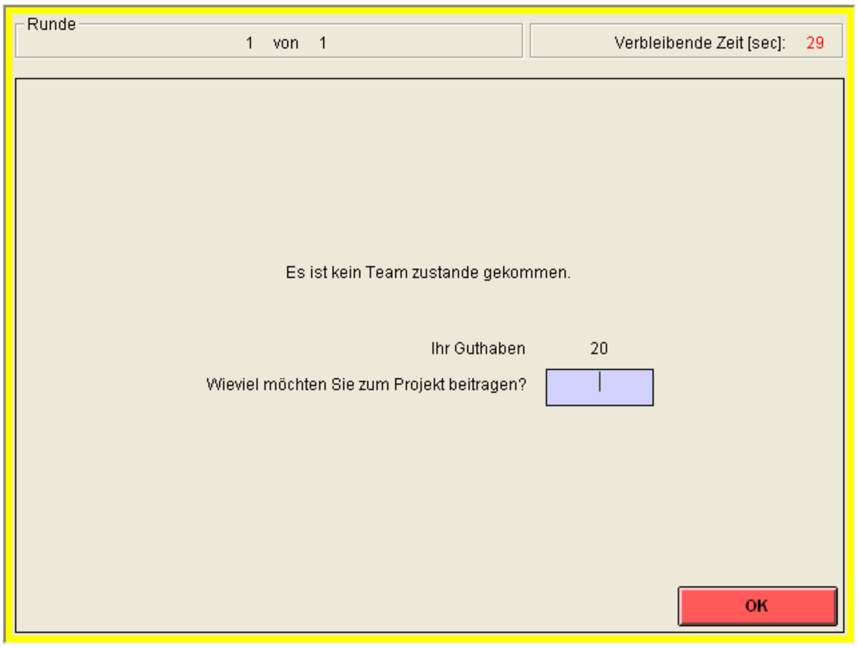
By joining the team you irrevocably commit yourself to invest your entire endowment of 20 tokens in the following investment phase, if a team is formed.

Investment phase:
If a team was formed, i.e., if all group members joined the team in joining phase 1 or at least three group members joined the team in joining phase 2, and you are in the team, then you do not take any decision in this phase. Your entire endowment of 20 tokens is automatically invested in the project and you see the screen:

Click on "OK" if you have read the information on the screen in order for the experiment to continue.
If a team was formed but you are not member of the team, then you can now decide how many tokens you want to invest in the project. On the screen you see:

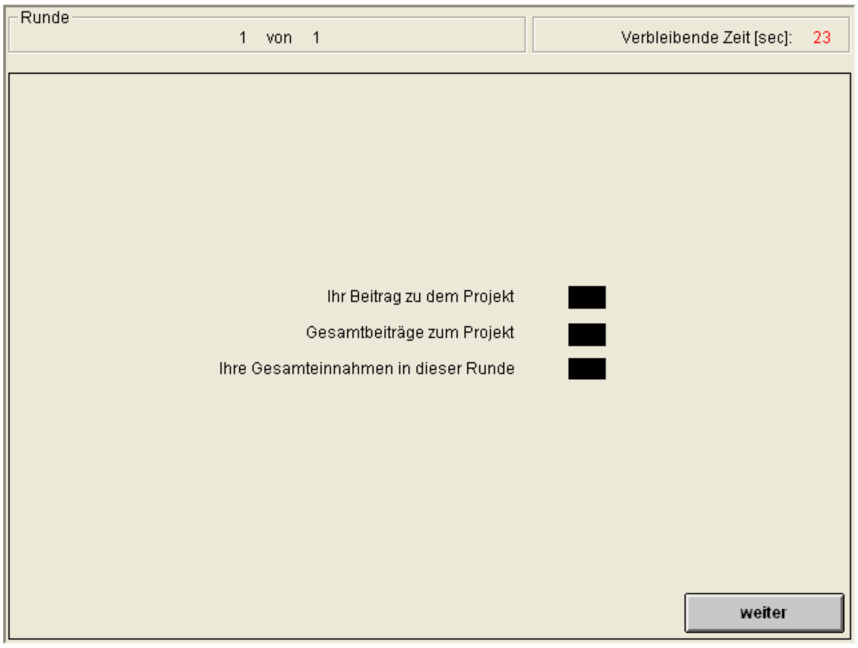


Enter your contribution to the project into the input box and click on "OK". Keep in mind that you can only enter integers 0, 1, 2, .., 20.
If no team was formed because not all group members joined the team in joining phase 1 and less than three group members joined the team in joining phase 2, then you can now decide how many tokens you want to invest in the project. On the screen you see:



Enter your contribution to the project into the input box and click on "OK". Keep in mind that you can only enter integers 0, 1, 2, .., 20.

At the end of the investment phase the round ends with a screen informing you about your contribution to the project, the total contributions to the project by the whole group and your total earnings in this round (numbers are blackened).



Click on "continue" if you have read the information on the screen in order for the experiment to continue with the next round which is identical to the previous round. The experiment ends after the 10th round.

Payment at the end of the experiment

At the end of the ten rounds of the experiment one round is randomly drawn. Your total earnings in that round are converted from tokens into Euro and are paid to you in cash. Independent of your earnings you additionally receive a fixed payment of 4 Euro for your participation in the experiment.

5.C.4 Control questions

In order to check if you have understood the experimental instructions correctly, we ask you to answer the following questions. All questions relate to one round of the experiment. The four group members are marked as A, B, C and D in the following. If you have a clarifying question, please raise your hand.

1. [ALL][19] Suppose that no team was formed and the group members A, B, C and D contribute as follows to the common project: A contributes 12 tokens, B 6 tokens, C 4 tokens and D 18 tokens.
   a) What are the project earnings of every group member? (Remember that 0.4*10 = 4)
   b) What are the total earnings of A?

2. [IF4] Suppose that all four group members joined the team in the joining phase. Please mark for each group member if the contribution to the project is fixed at 20 tokens or if the contribution may be freely chosen between 0 and 20 tokens.

Group member	Contribution fixed at 20 tokens	Contribution may be freely chosen between 0 and 20 tokens
A		
B		
C		
D		

3. [IF3] Suppose that all group members except for C joined the team in the joining phase. Please mark for each group member if the contribution to the project is fixed at 20 tokens or if the contribution may be freely chosen between 0 and 20 tokens.

Group member	Contribution fixed at 20 tokens	Contribution may be freely chosen between 0 and 20 tokens
A		
B		
C		
D		

[19] In the following "[ALL]" indicates that the corresponding question was used in all treatments, and "[IF4]" ("[IF3]", "[IF43]") indicates that the corresponding question was only used in IF4 (IF3, IF43).

4. [IF3] Suppose that all group members A, B, C and D joined the team in the joining phase. What are the total earnings of every group member? (Remember that 0.4*10 = 4)

5. [IF43] Suppose that not all four group members joined the team in the joining phase 1 and that all group members except for C joined the team in joining phase 2. Please mark for each group member if the contribution to the project is fixed at 20 tokens or if the contribution may be freely chosen between 0 and 20 tokens.

Group member	Contribution fixed at 20 tokens	Contribution may be freely chosen between 0 and 20 tokens
A		
B		
C		
D		

6. [IF43] Suppose that all group members A, B, C and D joined the team in joining phase 1. What are the total earnings of every group member? (Remember that 0.4*10 = 4)

Chapter 6
The Stability of Coalitions when Countries are Heterogeneous

Abstract I analyze the stability of coalitions in a global emissions game when countries are heterogeneous with respect to both costs and benefits of emissions. I consider the case where countries can generate either high or low benefits from individual emissions and either have high or low costs from global emissions. My analysis shows that large coalitions may be stable if heterogeneity in costs is large and only two countries with high costs join the coalition. A numerical example indicates that only in this case do the benefit types of coalition members also affect the size of the coalition.

Keywords Coalition Formation, Heterogeneous Countries, IEAs, Stability
JEL Classification C72, H41, H87, Q50

6.1 Introduction

The analysis of the impact of global climate change and its prevention has become increasingly important over the last two decades. It first became prominent in public after the signing of the Montreal Protocol in 1987 and drew increased attention with the Kyoto Protocol from 1997 and its failure to enter into force for almost ten years. Since there is no supernational power that can force signatories into any behavior, International Environmental Agreements (IEAs) need to be self-enforcing (cf. Barrett, 1994).

The literature on IEAs and its analysis may be divided into two strands; on the one hand the cooperative approach used by Chander and Tulkens (1995, 1997) and on the other hand the non-cooperative approach originating in Carraro and Siniscalco (1993) and Barrett (1994). The cooperative approach uses the stability concept of the core, whereas the predominant non-cooperative approach is based on the concept of internal and external stability introduced in the context of cartels by D'Aspremont et al. (1983). In his seminal paper, Barrett analyzes a model where all countries are identical and shows that self-enforcing IEAs with many signatories may exist, but that the maximal sizes of these IEAs are inversely related to the gains from cooperation. That is, if the difference between non-cooperative and full-cooperative global welfare is large, smaller IEAs will form than if the differences is low. He shows that for some simple combinations of cost and benefit functions the maximal size of an IEA may be as small as two or three.

Carraro and Siniscalco (1993) present the framework of how the formation of IEAs can be modeled as games between countries. Most of the time, the formation process of the IEA is modeled as a two-stage game where the first stage comprises the decisions of the countries whether or not to enter an agreement and in the second stage the countries choose their emission levels (dependent on the choices in the first stage). Researchers have examined the effects of transferability of utility and correspondingly of different transfer schemes in this setting. While it is agreed upon that transferability of utility can increase the size of an IEA, the rule according to which the payoff should be distributed is still under debate. Initially the Shapley-Value and the Nash bargaining solution was applied for the allocation of the coalition's payoff (Barrett, 1997b). Yet recently Eyckmans and Finus (2009) defined an *almost ideal sharing scheme* for games with positive externalities, Weikard (2009) developed an *optimal sharing rule* and McGinty (2011) proposed the so called *new allocation rule*. These rules all have in common that they use the free-rider payoff of a coalition member as the threat point and additionally assign a share of the remaining coalition's payoff to each member.

Participants in the global emissions game are often assumed to be homogeneous, which clearly does not reflect reality. However, heterogeneity in the model complicates the analysis and many authors have resorted to simulations and numerical analysis instead of analytical results. In one of the early contributions, Barrett (1997a) simulates a model with a total of 100 countries of two different types. McGinty (2007) later adds to this model by allowing for all countries to be asymmetric in simulations with seven and 20 countries.

Newer contributions to the literature use real world data and for the simulations pool several real countries into regions. Carraro et al. (2006) analyze a model with six world regions (USA, Japan, EU, China, Former Soviet Union and Rest of the World) and compare welfare and cumulative emissions for different stable coalitions. Other more recent contributions to the topic include Finus et al. (2008), who use a model with twelve world regions, and Osmani and Tol (2009), who apply a simulation model with 16 regions for their analysis. All these models face the criticism that the parameters have to be estimated and the quantitative results may be affected by small changes in parameter values.

Other related contributions come from Osmani and Tol (2010), who consider two types of countries and self-enforcing IEAs in settings with symmetric and asymmetric countries and compare the simulated outcomes in terms of welfare and total abatement, and from Biancardi and Villani (2010), who also use two types of countries and propose an algorithm and use a graphical solution method to determine stable agreements for different asymmetry levels.

The objective in this paper is to analyze a simple model of the global emission game where the countries are heterogeneous with respect to costs and benefits of emissions. I derive analytical results without having to resort to simulations or numerical analysis. In order to reduce the complexity of the model and thus keep it analytically tractable, I restrict to a limited amount of heterogeneity by only using two specifications, high and low costs and benefits respectively. I then identify maximally stable coalitions for models with four types of countries.

I know of two papers that have addressed a similar question. Fuentes-Albero and Rubio (2010) (corrected by Glanemann, 2012) focus on whether cooperation may be increased in a case of heterogeneous countries and when transfers are allowed or not. They find that the heterogeneity itself cannot increase the size of stable coalitions, but in conjunction with transfers it can. However, they only analyze the case when countries differ in either benefits or costs of emissions. My paper therefore extends their work on self-enforcing IEAs with transfers by allowing for simultaneous heterogeneity in both benefits and costs. Furthermore, there is the recent paper by Pavlova and de Zeeuw (2013) which comes closest to my model. In their paper they also allow for simultaneous heterogeneity in benefits and costs and find that, driven by the two-fold heterogeneity, transfers are not necessary in order to achieve larger stable coalitions. However, like Fuentes-Albero and Rubio (2010), they only consider two types of countries. That is, in their model the countries differ both in benefits and costs (or are identical), while in my model the countries differ from each other either only in costs or only in benefits or in both (or are identical).

In this paper, I propose maximally stable coalitions as solution concept for the coalition formation game. I show that for our model with four types of countries and heterogeneity in both benefits and costs, I can generalize the results from the models with two types of countries. My analysis shows that large coalitions may be maximally stable if heterogeneity in costs is large and only two countries with high costs join the coalition. Additionally, I provide a numerical example which indicates that the heterogeneity in benefits also has an impact on the maximally stable coalition size if there are two members with high costs.

6.2 Model and Approach

Suppose there is a set of $N = \{1, ..., n\}$ countries and let $e_i \geq 0$ be the emission of country i and $e = (e_1, e_2, ..., e_n)$ the vector of global emissions. I assume the payoff function of the countries to have the form

$$\pi_i(e) = b_i \cdot \left(ae_i - \frac{1}{2}(e_i)^2 \right) - c_i \sum_{j=1}^{n} e_j, \tag{6.1}$$

where $b_i > 0$ is called the benefit parameter and $c_i > 0$ the cost parameter of country i. $a > 0$ is the level of emissions at which the marginal benefit is zero and therefore constitutes a natural threshold for the level of emissions.[1]

I concentrate upon coalition structures that only consist of one non-singleton coalition. Thus, at first countries decide if they want to join the coalition or not, and countries who do not join the coalition remain singletons. Given that there is only one coali-

[1] The literature offers two general approaches to modeling payoffs from emissions. First, the model used here where countries receive benefits from individual emissions and suffer costs induced by global emissions. Second, a public good model where countries suffer from costs of controlling emissions and benefit from global abatement. However, both approaches are identical if cost and benefit functions are correctly translated into the corresponding functional form. The emission game used here with quadratic benefits and linear costs would translate into linear benefits and quadratic costs in the public goods game. In the related literature Pavlova and de Zeeuw (2013) apply the first approach while Fuentes-Albero and Rubio (2010) use the second. I choose this particular functional form for our analysis as it allows me to obtain an analytical solution.

tion possible, a coalition structure is then fully characterized by a coalition $S \subseteq N$. Thereafter, the countries choose their optimal emission level depending on the coalition structure, i.e., non-members choose an emission level that maximizes their individual payoffs, while coalition members choose a level that maximizes the coalition's payoff. Thus non-member countries choose the emission level $e_i^I = a - \frac{c_i}{b_i}$ and under full cooperation all countries reduce their emissions to $e_i^F = a - \frac{\sum_{j=1}^n c_j}{b_i}$, which is the socially optimal choice. It is straightforward to verify that members of a coalition S choose $e_i^C = a - \frac{\sum_{j \in S} c_j}{b_i}$, $i \in S$.

Moreover, it is apparent that for this particular payoff function the optimal emission choice of a country or coalition is independent of the emission levels of all the other countries. This results from the fact that the reaction functions are orthogonal due to the linearity of the cost function. Assume that all countries choose their emission levels as presented above. The emission levels of the countries are then determined by the coalition structure and consequently by the only non-singleton coalition S. Therefore, with a slight abuse of notation, I will denote the total equilibrium payoff of coalition S by $\pi_S(S)$ and the equilibrium payoff of a non-member j when the coalition S forms by $\pi_j(S)$.

I consider this to be a transferable utility game, hence the total payoff of the coalition may be redistributed in any way among its members. Thus there is a valuation function v which assigns to every possible coalition $S \subseteq N$ a vector $v(S) \in \mathbb{R}^n$ such that $\sum_{i \in S} v_i(S) = \pi_S(S)$ and $v_j(S) = \pi_j(S)\ \forall j \notin S$.

In order to discriminate among possible coalitions, I use the concept of stability according to D'Aspremont et al. (1983). A coalition S is stable with respect to the valuation v if it is internally stable, i.e., $v_i(S) \geq v_i(S \backslash \{i\})\ \forall\ i \in S$ and externally stable, i.e., $v_j(S) \geq v_j(S \cup \{j\})\ \forall\ j \notin S$. Hence, S is stable if no member wants to leave the coalition and if no non-member wants to join the coalition, given the valuations.

In the following I use a slightly modified version of the concept of stability in order to be independent of specific valuation functions. Moreover, as global welfare increases with the size of a coalition, I define maximally internally stable and maximally stable coalitions as the desirable outcomes of the game.

Definition 6.1 *A coalition S is **(internally) stable** if there exists a valuation function v for which S is (internally) stable.*

Definition 6.2 *A coalition S is **maximally (internally) stable** if S is (internally) stable and any coalition $T \supset S$ is not (internally) stable.*[2]

[2] Here and in the following, " $\supset$ " denotes a proper superset, hence $T \supset S \Leftrightarrow (T \supseteq S$ but $T \neq S)$.

I intend to use the surplus of a coalition S as a means to determine maximally stable coalitions.[3] The surplus of a coalition S is defined as the difference of the payoff of the coalition and the sum of all free-rider payoffs that the members.

Definition 6.3 $\Delta(S) \equiv \pi_S(S) - \sum_{i \in S} \pi_i(S \setminus \{i\})$ *is the* ***surplus of coalition*** S*.*

If there is a positive surplus of coalition S, then the payoffs of the coalition can be distributed in a way such that every member of the coalition has a higher valuation in the coalition than outside of the coalition. In other words, there exists a valuation function v for which S is internally stable. Eyckmans and Finus (2009) call a coalition S *potentially internally stable* if $\Delta(S) \geq 0$. Note that this implies that for a coalition S to be internally stable according to Definition 6.1, it has to be potentially internally stable, and vice versa. I provide no proof for this statement as it is straightforward.

Lemma 6.1 *S is internally stable if and only if $\Delta(S) \geq 0$*

I can use this property to show that maximal internal stability is equivalent to maximal stability, and that all maximally stable coalitions may be identified by the surplus.

Lemma 6.2 *A coalition S is maximally internally stable if and only if S is maximally stable.*[4]

Proof. Assume that S is maximally internally stable, i.e., by Lemma 6.1 $\Delta(S) \geq 0$ and $\Delta(T) < 0 \ \forall T \supset S$. We may then define a valuation function v that assigns $v_i(U) = \pi_i(U \setminus \{i\}) + \lambda_i(U) \cdot \Delta(U) \ \forall i \in U$ with $0 \leq \lambda_i(U) \leq 1$ and $\sum_{i \in U} \lambda_i(U) = 1$ to the members of any coalition U. Clearly, S is then stable with respect to valuation v and therefore also stable.

Now assume that there is a larger coalition T that is also stable. This would especially imply that it is internally stable, which is a contradiction to our initial assumption that S is the maximally internally stable coalition.

Now, for sufficiency, assume that S is maximally stable and but not maximally internally stable. Hence, there must be a coalition $T \supset S$ that is maximally internally stable. But by the first part of the proof, coalition T must then be maximally stable, which contradicts the assumption that S is maximally stable. $\square$

[3] I use the terminology from Eyckmans and Finus (2009) throughout the paper. In Weikard (2009) this property is called the Claim Rights Condition.

[4] This result is an extension of previous results concerning the relationship between external stability and potential internal stability (Lemma 2 in Eyckmans and Finus (2009) and Lemma 1 in Weikard (2009)), and closely related to the results regarding stability of internally stable coalitions with the highest worth (Proposition 4 in Eyckmans and Finus (2009)).

Summarizing, I defined maximally stable coalitions as the desirable outcomes of the game. I showed that any maximally stable coalition is also maximally internally stable. Thus, there is only need to identify the maximally internally stable coalitions. Since internally stable coalitions are fully characterized by the surplus, maximally internally stable coalitions are the largest coalitions with non-negative surpluses.

6.3 2×2-Heterogeneity: Four Types of Countries

It seems to be impossible to derive analytical results from the model when benefit and cost parameters are different across all countries and many authors have therefore applied numerical analysis. I try to derive general analytical results by only allowing for a certain degree of heterogeneity in between countries. In my model I assume that there are some countries that can generate more benefits from a given level of emissions than the other countries. That is represented by a benefit parameter b_h while the other countries have the parameter b_l. The absolute values of these parameters are not of any importance, but they have to be such that $b_h > b_l > 0$. Additionally, I assume that some countries suffer more from global emissions than others. That is represented by two levels of cost parameters such that $c_h > c_l > 0$. A country thus belongs to one of the four types (b_h, c_h), (b_h, c_l), (b_l, c_h) or (b_l, c_l). Let n_{hh}, n_{hl}, n_{lh} and n_{ll} be the number of countries of each type, respectively.

Let $\pi_{rt}(e) = b_r \cdot \left(ae_i - \frac{1}{2}(e_i)^2\right) - c_t \sum_{j=1}^{n} e_j$ be the payoff function of a country i with $r \in \{h, l\}$ and $t \in \{h, l\}$. Thus, country i emits $e^I_{rt} = a - \frac{c_t}{b_r}$ when not member of a coalition. I define $\alpha \equiv \frac{c_h}{b_h}$, $\beta \equiv \frac{c_l}{b_h}$, $\gamma \equiv \frac{c_h}{b_l}$, and $\delta \equiv \frac{c_l}{b_l}$ in order to simplify. Suppose a coalition S forms, then for each type of country, there is the possibility to be either a member of the coalition or not. Thus, each country may be classified into one of eight categories. The emission levels of the different categories are summarized in Table 6.1 to give a better overview. Note that the emission levels of coalition members depend on the composition of the coalition in terms of cost-types but also of one's own benefit-type. Hence coalition members with the same benefit parameter choose the same emission levels even if they have different cost types. Non-members however do not choose the same emission levels. The size of coalition S is denoted by $s = m^C_{hh} + m^C_{hl} + m^C_{lh} + m^C_{ll}$. I also use $s_{b_h} = m^C_{hh} + m^C_{hl}$, $s_{b_l} = m^C_{lh} + m^C_{ll}$, $s_{c_h} = m^C_{hh} + m^C_{lh}$ and $s_{c_l} = m^C_{hl} + m^C_{ll}$ to provide a more compact notation.

Table 6.1: The emission levels of the eight different country categories.

Type	Coalition member	Number of countries	Emission level
(b_h, c_h)	yes	m_{hh}^C	$e_{hh}^C = a - s_{c_h}\alpha - s_{c_l}\beta$
(b_h, c_l)	yes	m_{hl}^C	$e_{hl}^C = a - s_{c_h}\alpha - s_{c_l}\beta$
(b_l, c_h)	yes	m_{lh}^C	$e_{lh}^C = a - s_{c_h}\gamma - s_{c_l}\delta$
(b_l, c_l)	yes	m_{ll}^C	$e_{ll}^C = a - s_{c_h}\gamma - s_{c_l}\delta$
(b_h, c_h)	no	$m_{hh}^I = n_{hh} - m_{hh}^C$	$e_{hh}^I = a - \alpha$
(b_h, c_l)	no	$m_{hl}^I = n_{hl} - m_{hl}^C$	$e_{hl}^I = a - \beta$
(b_l, c_h)	no	$m_{lh}^I = n_{lh} - m_{lh}^C$	$e_{lh}^I = a - \gamma$
(b_l, c_l)	no	$m_{ll}^I = n_{ll} - m_{ll}^C$	$e_{ll}^I = a - \delta$

It is straightforward to show that the total payoff of the coalition S is

$$\begin{aligned}\pi_S(S) &= \frac{1}{2}b_h s_{b_h}\left(a^2 + (s_{c_h}\alpha + s_{c_l}\beta)^2\right) + \frac{1}{2}b_l s_{b_l}\left(a^2 + (s_{c_h}\gamma + s_{c_l}\delta)^2\right) \\ &\quad + (c_h s_{c_h} + c_l s_{c_l})\left(m_{hh}^I\alpha + m_{hl}^I\beta + m_{lh}^I\gamma + m_{ll}^I\delta - na\right).\end{aligned}$$

In order to determine the surplus $\Delta(S)$, the individual payoffs of non-members for coalition structure $S \setminus \{i\}$ are needed. For the different types, they result to

$$\begin{aligned}\pi_{hh}^I(S \setminus \{i\}) &= \frac{1}{2}b_h(a^2 + \alpha^2) + c_h\left[(s_{b_h} - 1)((s_{c_h} - 1)\alpha + s_{c_l}\beta)\right. \\ &\quad \left. + s_{b_l}((s_{c_h} - 1)\gamma + s_{c_l}\delta) + m_{hh}^I\alpha + m_{hl}^I\beta + m_{lh}^I\gamma + m_{ll}^I\delta - na\right]\end{aligned}$$

$$\begin{aligned}\pi_{hl}^I(S \setminus \{i\}) &= \frac{1}{2}b_h(a^2 + \beta^2) + c_l\left[(s_{b_h} - 1)(s_{c_h}\alpha + (s_{c_l} - 1)\beta)\right. \\ &\quad \left. + s_{b_l}(s_{c_h}\gamma + (s_{c_l} - 1)\delta) + m_{hh}^I\alpha + m_{hl}^I\beta + m_{lh}^I\gamma + m_{ll}^I\delta - na\right]\end{aligned}$$

$$\begin{aligned}\pi_{lh}^I(S \setminus \{i\}) &= \frac{1}{2}b_l(a^2 + \gamma^2) + c_h\left[(s_{b_l} - 1)((s_{c_h} - 1)\gamma + s_{c_l}\delta)\right. \\ &\quad \left. + s_{b_h}((s_{c_h} - 1)\alpha + s_{c_l}\beta) + m_{hh}^I\alpha + m_{hl}^I\beta + m_{lh}^I\gamma + m_{ll}^I\delta - na\right]\end{aligned}$$

$$\begin{aligned}\pi_{ll}^I(S \setminus \{i\}) &= \frac{1}{2}b_l(a^2 + \delta^2) + c_l\left[(s_{b_l} - 1)(s_{c_h}\gamma - (s_{c_l} - 1)\delta)\right. \\ &\quad \left. + s_{b_h}(s_{c_h}\alpha + (s_{c_l} - 1)\beta) + m_{hh}^I\alpha + m_{hl}^I\beta + m_{lh}^I\gamma + m_{ll}^I\delta - na\right].\end{aligned}$$

The surplus of coalition S then amounts to

$$\begin{aligned}\Delta(S) &= \pi_S(S) - \sum_{i\in S} \pi_i(S \setminus \{i\}) \\ &= b_h s_{b_h} \left(s_{c_h}\alpha^2 + s_{c_l}\beta^2\right) + b_l s_{b_l} \left(s_{c_h}\gamma^2 + s_{c_l}\delta^2\right) \\ &\quad + \left(c_h m_{hh}^C + c_l m_{hl}^C\right)(s_{c_h}\alpha + s_{c_l}\beta) + \left(c_h m_{lh}^C + c_l m_{ll}^C\right)(s_{c_h}\gamma + s_{c_l}\delta) \\ &\quad -\frac{1}{2}(s_{c_h}\alpha + s_{c_l}\beta)^2 b_h s_{b_h} - \frac{1}{2}(s_{c_h}\gamma + s_{c_l}\delta)^2 b_l s_{b_l} \\ &\quad -\frac{3}{2}\left(b_h m_{hh}^C\alpha^2 + b_h m_{hl}^C\beta^2 + b_l m_{lh}^C\gamma^2 + b_l m_{ll}^C\delta^2\right)\end{aligned}$$

Observe that the surplus only depends on the composition of the coalition S, i.e., the number of countries of each type that have joined the coalition. With a slight abuse of notation I will therefore use $S = \left(m_{hh}^C, m_{hl}^C, m_{lh}^C, m_{ll}^C\right)$ to specify a coalition. Similarly, I use $\Delta\left(m_{hh}^C, m_{hl}^C, m_{lh}^C, m_{ll}^C\right)$ for the coalition's surplus to capture differences in coalition composition.

I start the analysis of the surplus and maximally stable coalitions with a result that limits the number of c_h-type countries in a maximally stable coalition.

Proposition 6.1 *No coalition with $s_{c_h} > 3$ is stable. If $s_{c_h} = 3$, then coalition $S = \left(m_{hh}^C, m_{hl}^C, m_{lh}^C, m_{ll}^C\right)$ is maximally stable if and only if $s_{c_l} = 0$.*

Proof. I only provide the sketch of the proof here, the complete proof is shown in Appendix 6.A. Δ is a polynomial function of third degree with negative leading coefficients in each of the variables m_{hl}^C and m_{ll}^C. Then I can show that

(i) $\Delta(m_{hh}^C, 0, m_{lh}^C, 0) < 0$ if $s_{c_h} > 3$ and $\Delta(m_{hh}^C, 0, m_{lh}^C, 0) = 0$ if $s_{c_h} = 3$, i.e., if s_{c_h} is larger or equal to three, the surplus is less or equal to zero at $m_{ll}^C = m_{hl}^C = 0$.

(ii) $\frac{\partial\Delta}{\partial m_{hl}^C}(m_{hh}^C, 0, m_{lh}^C, m_{ll}^C) < 0$ $\forall m_{ll}^C \geq 0$, and $\frac{\partial\Delta}{\partial m_{ll}^C}(m_{hh}^C, 0, m_{lh}^C, 0) < 0$ if in each case $s_{c_h} \geq 3$, i.e., if s_{c_h} is larger or equal to three, then the function is decreasing in m_{hl}^C at $m_{hl}^C = 0$ and decreasing in m_{ll}^C at $m_{ll}^C = m_{hl}^C = 0$.

(iii) $\frac{\partial^2\Delta}{\partial(m_{hl}^C)^2}(m_{hh}^C, 0, m_{lh}^C, m_{ll}^C) < 0$ and $\frac{\partial^2\Delta}{\partial(m_{ll}^C)^2}(m_{hh}^C, 0, m_{lh}^C, 0) < 0$ if $s_{c_h} \geq 3$, i.e., the function is concave in both variables.

By (ii) and (iii) I may conclude that given a fixed m_{ll}^C (or $m_{hl}^C = 0$) and a fixed $s_{c_h} \geq 3$, at $m_{hl}^C = 0$ (or $m_{ll}^C = 0$ respectively) Δ is decreasing and concave in m_{hl}^C (m_{ll}^C) and thus in the section to the right of its local maximum. Hence, the maximum of Δ for any $m_{hl}^C \geq 0$, $m_{ll}^C \geq 0$ and $s_{c_h} \geq 3$ is at $m_{hl}^C = m_{ll}^C = 0$, which is less than (or equal to) zero by (i). $\square$

I thus know that no more than three countries of type c_h can be members in a maximally stable coalition, as that would lead to internal instability. This result resembles the results previously obtained by Pavlova and de Zeeuw (2013) and Fuentes-Albero and Rubio (2010). In both models, the authors also find that there is an upper limit of three countries with high costs in a stable coalition. I was thus able to extend these results to a more general and thus more realistic setting with more types of countries. Furthermore, the analysis allows me to generalize the most important of their findings: the maximal size of a stable coalition. This result is summarized as follows.

Proposition 6.2 *If $s_{c_h} = 2$, then for every $\underline{s} \geq 1$ there exist c_h, m_{hl}^C and m_{ll}^C with $m_{hl}^C + m_{ll}^C \geq \underline{s}$ such that the coalition $S = \left(m_{hh}^C, m_{hl}^C, m_{lh}^C, m_{ll}^C\right)$ is maximally stable.*

Proof. From Proposition 6.1 I already know that no c_h-type country has an incentive to join if $s_{c_h} = 2$ and $s_{c_l} \geq 1$. Thus, these coalitions are externally stable against entry of c_h-type countries. Therefore, I only need to consider the three possible cases in which $s_{c_h} = 2$. I find that in all of these cases Δ may be described as a function of the form

$$a \cdot (c_h)^2 - b \cdot c_h - c,$$

with b and c being functions in m_{hl}^C, m_{ll}^C, b_h, b_l and c_l. The factor a is a function in m_{hl}^C, m_{ll}^C, b_h and b_l and is greater than zero for all $s_{c_l} > 0$.[5] Hence, for any combination of m_{hl}^C, m_{ll}^C, b_h, b_l and c_l, there is a c_h such that $\Delta(S) \geq 0$ and thus potentially internally stable. Since the number of c_l-type countries is finite, there is always a maximally stable coalition with two high cost countries and at least $\underline{s}$ low cost countries. $\square$

Thus, by Proposition 6.2, I know that if the cost parameter c_h is sufficiently large, countries are able to form large stable coalitions. The largest possible stable coalition comprises two countries of type c_h and all countries of type c_l. This result shows that even full cooperation may be achieved if only two countries of type c_h exist.

For coalitions with only one c_h-type country, I am able to show that there is always a c_h which makes the coalition potentially internally stable. For this result presented in Proposition 6.3, which follows directly from the fact that Δ is a quadratic function in c_h, I provide no proof. See Appendix 6.A for the relevant functions.

Proposition 6.3 *If $s_{c_h} = 1$, then for every $\underline{s} \geq 2$ there exist c_h, m_{hl}^C and m_{ll}^C with $m_{hl}^C + m_{ll}^C \geq \underline{s}$ such that coalition the $S = \left(m_{hh}^C, m_{hl}^C, m_{lh}^C, m_{ll}^C\right)$ is potentially internally stable.*

[5] The exact functions are found in Appendix 6.A.

Example 6.1 shows that there also exist maximally stable coalitions with only one c_h-type country. However, due to the complexity of the model, I am not able to provide a general overview for which parameters and coalitions they may or may not emerge. The evidence of the example suggests that for given parameters c_h, c_l, b_h and b_l, a coalition with a larger size may be formed with only one country of c_h-type.[6] Thus, it is easier to form a larger coalition if only one c_h-type country joins than if two countries with high costs join.

The next result concerns the size of a coalition when only c_l-type countries are members. It is a direct corollary from previous results concerning only one type of member-country and Proposition 6.3.

Proposition 6.4 *No coalition with $s_{c_l} > 3$ and $s_{c_h} = 0$ is stable. Coalition $S = \left(m_{hh}^C, m_{hl}^C, m_{lh}^C, m_{ll}^C\right)$ with $s_{c_l} = 3$ and $s_{c_h} = 0$ is maximally stable if and only if c_h is sufficiently small.*

Proof. It is straightforward to show that the surplus is

$$\Delta(0, m_{hl}^C, 0, m_{ll}^C) = \frac{(c_l)^2}{2}\left(\frac{m_{hl}^C}{b_h} + \frac{m_{ll}^C}{b_l}\right)\left(1 - (s_{c_l} - 2)^2\right).$$

Thus, the root of the surplus is at $s_{c_l} = 3$. Any coalition with more members of type c_l has a negative surplus, which proves the first statement.

If $s_{c_l} = 3$ and $s_{c_h} = 0$, then the surplus is zero and thus the coalition potentially internally stable. It is maximally internally stable if and only if c_h is not large enough to produce a potentially internally stable coalition where $s_{c_l} = 3$ and $s_{c_h} = 1$ (see Proposition 6.3). It is straightforward to show that there always exists a c_h such that the surplus of any coalition $S = \left(m_{hh}^C, m_{hl}^C, m_{lh}^C, m_{ll}^C\right)$ with $s_{c_l} = 3$ and $s_{c_h} = 1$ is less than zero, i.e., $\exists c_h : \Delta\left(m_{hh}^C, m_{hl}^C, m_{lh}^C, m_{ll}^C\right) < 0$ where $s_{c_l} = 3$ and $s_{c_h} = 1$. That proves the second statement. □

Summarizing, I showed that in a model where there are four types of countries that are heterogeneous both in benefits and costs of emissions, the largest maximally stable coalition is the coalition that consists of two countries with high costs and all countries with low costs. However, the maximal size is highly dependent on the heterogeneity in costs. The higher the heterogeneity in costs, the larger is the largest maximally stable coalition. These findings are generalizations of the results obtained in models with only

[6] I find further support for this conjecture in the model by Fuentes-Albero and Rubio (2010) with only heterogeneity in costs.

two types of countries. Thus, I was able to extend the models of both Fuentes-Albero and Rubio (2010) and Pavlova and de Zeeuw (2013) to a more complex setting.

I conclude the section with an example. As already mentioned above, it shows that maximally stable coalitions are possible also with only one member of type c_h. Furthermore, Example 6.1 also indicates that the benefit types do play a significant role for the size of a maximally stable coalitions in all cases when there are two members of type c_h. This finding shows that, although it has minor impacts compared to heterogeneity in costs, it is not possible to completely ignore heterogeneity in benefits.

Example 6.1 Let $b_l = 1$, $b_h = 5$, $c_l = 0.1$ and $c_h = 10$ be the benefit and cost parameters. In Table 6.2 I depict a selection of maximally stable coalitions. In this example, I have chosen parameters that promote larger maximally stable coalitions. E.g., note that the cost ratio is 100. Throughout the example I assume that n is sufficiently large and that there are sufficiently many countries of each type in order to form the coalitions.

Table 6.2: Some maximally stable coalitions for $b_l = 1$, $b_h = 5$, $c_l = 0.1$ and $c_h = 10$.

m_{hh}^C	m_{lh}^C	m_{hl}^C	m_{ll}^C	**Size**
1	0	$0 \leq m_{hl}^C \leq 42$	$42 - m_{hl}^C$	**43**
0	1	$0 \leq m_{hl}^C \leq 42$	$42 - m_{hl}^C$	**43**
0	2	13	0	**15**
1	1	11	0	**13**
2	0	7	0	**9**
0	2	0	7	**9**
1	1	1	5	**8**
2	0	0	3	**5**

I obtain that if only one c_h-type country joins, the maximally stable coalition size amounts to 43. Furthermore, the largest IEA (in terms of most members) is obtained independently of the benefit types. Any combination of benefit types is a maximally stable coalition as long as there is one country with high costs from global emissions and 42 countries with low costs from global emissions. For all other cases, this property does not hold.

As may be seen in rows 3-5 and 6-8 of Table 6.2, the largest total size of the maximally stable coalitions decreases in the number of (b_h, c_h)-type countries. That is, when there are no (b_h, c_h)-type countries, then the largest total size is 15, when there is one (b_h, c_h)-type country, then the largest total size is 13 and when there are two (b_h, c_h)-type countries the largest total size is 9. Furthermore, the sizes of maximally stable coalitions with the same number of (b_h, c_h)-type members also depend on the benefit types of the c_l-type members. The more (b_l, c_l)-type countries are members, the smaller is the maximally stable coalition. E.g., the maximally stable coalition in row 6 with two (b_l, c_h)-type countries and 7 (b_l, c_l)-type countries has a total of only 9 members, compared to the 15 members of the coalition in row 3.

Generally I can conclude that if the set of countries is large, a very high degree of heterogeneity in costs is needed in order to form a stable grand coalition. Also, the composition in terms of benefit types does play a role for the size of the coalition if there is more than just one country of type c_h.

6.4 Conclusion

In this paper, I have derived analytical results for the formation of maximally stable coalitions when countries are heterogeneous both in the environmental costs induced by global emissions and the benefits obtained from local emissions. I relaxed the problem of complexity of the model by only allowing for two characteristics each. Countries either have high or low benefits from local emissions and either high or low costs from global emissions. Thus I created a model with four different types of countries.

My analysis may be summarized as follows: Stable coalitions with more than three members may only be achieved if only one or two countries with high costs are members. If that is the case, then the degree of heterogeneity in costs is the primary determinant of the size of the stable coalitions. The higher the degree of heterogeneity in costs, the larger stable coalitions may form. Thus, the largest possible coalition comprises two high cost countries and all low cost countries.

I was able to extend general findings of authors who employed models with only two types of countries to a more general model with four types of countries. This finding indicates that future researchers may settle for less complex models, as the additional effort need to solve the more complex ones might not be worthwhile. However, this reasoning is only true for the very specific model specifications of quadratic benefits and linear costs. It is not clear whether the general results hold for other functions, let alone for models where the reaction functions of the countries are not orthogonal and the players optimal emission choices are interdependent. This remains to be investigated.

Appendix

6.A Proofs

Proof of Proposition 6.1

(i) $\Delta(m_{hh}^C, 0, m_{lh}^C, 0) = -\frac{(c_h)^2\left(b_l m_{hh}^C + b_h m_{lh}^C\right)\left((s_{c_h}-2)^2-1\right)}{2b_h b_l}$ is trivially negative when $s_{c_h} > 3$ and zero if $s_{c_h} = 3$.

(ii) We have

$$\begin{aligned}
\frac{\partial \Delta}{\partial m_{hl}^C}(m_{hh}^C, 0, m_{lh}^C, m_{ll}^C) \quad = \quad & -\frac{1}{2b_h b_l}\left(2b_h c_l \left[c_h\left(s_{b_l}s_{c_h} - m_{lh}^C\right)\right.\right. \\
& \left. + c_l\left(s_{b_l}(m_{ll}^C - 1) - m_{ll}^C\right)\right] \\
& + b_l\left[(c_h)^2 s_{c_h}(s_{c_h} - 2)\right. \\
& + (c_l)^2(m_{ll}^C - 1)(2m_{hh}^C + m_{ll}^C - 3) \\
& \left.\left. + 2c_h c_l\left(s_{c_h}(m_{hh}^C + m_{ll}^C - 1) - m_{hh}^C\right)\right]\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Delta}{\partial m_{ll}^C}(m_{hh}^C, 0, m_{lh}^C, 0) \quad = \quad & -\frac{1}{2b_h b_l}\left(2b_l c_l m_{hh}^C\left[c_h\left(s_{c_h} - 1\right) - c_l\right]\right. \\
& + b_h\left[(c_h)^2 s_{c_h}(s_{c_h} - 2) - (c_l)^2(2m_{lh}^C - 3)\right. \\
& \left.\left. + 2c_h c_l\left(s_{c_h}(m_{lh}^C - 1) - m_{lh}^C\right)\right]\right)
\end{aligned}$$

Observe that in each derivative there are two parts (in brackets) being summed. We show that the whole functions are negative by showing that both parts of the sum never are. As the proof for $\frac{\partial \Delta}{\partial m_{ll}^C}(m_{hh}^C, 0, m_{lh}^C, 0)$ is analogous to the proof for $\frac{\partial \Delta}{\partial m_{hl}^C}(m_{hh}^C, 0, m_{lh}^C, m_{ll}^C)$ (by interchanging m_{lh}^C and m_{hh}^C), we only show the proof for the latter here.

The first part of $\frac{\partial \Delta}{\partial m_{hl}^C}(m_{hh}^C, 0, m_{lh}^C, m_{ll}^C)$ reads

$$c_h\left(s_{b_l}s_{c_h} - m_{lh}^C\right) + c_l\left(s_{b_l}(m_{ll}^C - 1) - m_{ll}^C\right).$$

This part is zero if $s_{b_l} = 0$ and strictly positive else. To see that, note that the first term $c_h\left(s_{b_l}s_{c_h} - m_{lh}^C\right) = c_h\left((m_{lh}^C + m_{ll}^C)s_{c_h} - m_{lh}^C\right)$ is zero if $m_{lh}^C = 0$ and $m_{ll}^C = 0$ and otherwise greater than (or equal to) $2c_h m_{lh}^C$ since $s_{c_h} \geq 3$.

The second term may be written as $c_l\left(\left(m_{lh}^C + m_{ll}^C\right)(m_{ll}^C - 1) - m_{ll}^C\right)$ and is also zero if $m^C_{lh} = 0$ and $m^C_{ll} = 0$ and negative in two cases. Either, when $m^C_{ll} = 0$ and $m^C_{lh} > 0$, it results to $-c_l m^C_{lh}$. But that is in absolute value less than the $2c_h m^C_{lh}$ from the first term and thus whole part would be positive. The second case in which the second term may become negative is when $m^C_{lh} = 0$ and $0 < m^C_{ll} < 2$. We calculate the value of the whole first part and get

$$c_h m^C_{ll} m^C_{hh} + c_l \left((m^C_{ll})^2 - 2m^C_{ll} \right).$$

Since $m^C_{hh} \geq 3$ in order to satisfy $s_{c_h} \geq 3$, it is always positive. Hence, the first part is non-negative.

The second part of $\frac{\partial \Delta}{\partial m^C_{hl}}(m^C_{hh}, 0, m^C_{lh}, m^C_{ll})$ reads

$$\begin{aligned}(c_h)^2 s_{c_h}(s_{c_h} - 2) + (c_l)^2(m^C_{ll} - 1)(2m^C_{hh} + m^C_{ll} - 3)& \\ +2c_h c_l \left(s_{c_h}(m^C_{hh} + m^C_{ll} - 1) - m^C_{hh} \right)&\end{aligned}$$

We show that this part is always strictly positive.

The first term is trivially positive since $s_{c_h} \geq 3$. Thus in the following we only consider the cases in which the other terms are strictly negative.

For the third term $2c_h c_l \left(s_{c_h}(m^C_{hh} + m^C_{ll} - 1) - m^C_{hh} \right)$, consider first the case when $m^C_{hh} \neq 0$. Then, it is obvious that it can only turn negative if $m^C_{ll} = 0$ and $m^C_{hh} = 1$. However, then the whole second part is larger or equal to

$$3(c_h)^2 + (c_l)^2 - 2c_h c_l > 0.$$

Consider now the case when $m^C_{hh} = 0$. Then again, let $m^C_{ll} = 0$ in order to get the minimum $-2c_h c_l m^C_{lh}$ and now the entire second part turns into

$$(c_h)^2 m^C_{lh}(m^C_{lh} - 2) + 3\,(c_l)^2 - 2c_h c_l m^C_{lh}.$$

This quadratic function in m^C_{lh} reaches its minimum at $m^C_{lh} = 1 + \frac{c_l}{c_h} < 2$, but as $m^C_{lh} \geq 3$ in order to satisfy the condition $s_{c_h} \geq 3$, we can calculate the lower bound, i.e., when $m^C_{lh} = 3$. We get

$$3\,(c_h)^2 + 3(c_l)^2 - 6c_h c_l = 3\,(c_h - c_l)^2 > 0.$$

Last, we investigate the second term $(c_l)^2(m^C_{ll} - 1)(2m^C_{hh} + m^C_{ll} - 3)$. Consider the case when $m^C_{ll} = 0$ and $m^C_{hh} > \frac{3}{2}$. The term reaches its minimum $-\left(c_l\right)^2 (2m^C_{hh} -$

3) when $m_{hh}^C \rightarrow \infty$. However, it is then dominated by the positive quadratic $(c_h)^2 s_{c_h}(s_{c_n} - 2)$ from the first term and the third term is also positive.

For the last possible combination, $m_{ll}^C = 2$ and $m_{hh}^C = 0$, we get $-(c_l)^2$. This can trivially not lead to a negative second part as the first term is always larger in absolute value. Hence, the second part is always positive.

Thus, Δ is decreasing in m_{hl}^C on the positive axis of m_{ll}^C (and decreasing in m_{ll}^C in the origins of m_{hl}^C and m_{ll}^C).

(iii) $\frac{\partial^2 \Delta}{\partial (m_{hl}^C)^2}(m_{hh}^C, 0, m_{lh}^C, m_{ll}^C) = -\frac{c_l\left[b_h c_l s_{b_l} + b_l\left(2c_h s_{c_h} + c_l(m_{hh}^C + 2m_{ll}^C - 4)\right)\right]}{b_h b_l}$ and $\frac{\partial^2 \Delta}{\partial (m_{ll}^C)^2}(m_{hh}^C, 0, m_{lh}^C, 0) = -\frac{c_l\left[b_l c_l m_{hh}^C + b_h\left(2c_h s_{c_h} + c_l(2m_{lh}^C - 4)\right)\right]}{b_h b_l}$ are both trivially negative for $s_{c_h} \geq 3$, i.e., the function is concave in both variables.

Functions for Propositions 6.2 and 6.3

$$\begin{aligned}
\Delta\left(1, m_{hl}^C, 0, m_{ll}^C\right) &= \frac{1}{2b_h b_l}\left[(c_h)^2\left(b_l m_{hl}^C + b_h m_{ll}^C\right)\right.\\
&\quad -2c_h c_l\left(s_{c_l}-1\right)\left(b_l m_{hl}^C + b_h m_{ll}^C\right)\\
&\quad -\left(c_l\right)^2\left(b_l\left[3m_{hl}^C + s_{c_l}\left(s_{c_l}-2+m_{hl}^C(s_{c_l}-4)\right)\right]\right.\\
&\quad \left.\left.+b_h m_{ll}^C\left[3+s_{c_l}\left(s_{c_l}-4\right)\right]\right)\right]\\
\Delta\left(0, m_{hl}^C, 1, m_{ll}^C\right) &= \frac{1}{2b_h b_l}\left[(c_h)^2\left(b_l m_{hl}^C + b_h m_{ll}^C\right)\right.\\
&\quad -2c_h c_l\left(s_{c_l}-1\right)\left(b_l m_{hl}^C + b_h m_{ll}^C\right)\\
&\quad -\left(c_l\right)^2\left(b_l m_{hl}^C\left[3+s_{c_l}\left(s_{c_l}-4\right)\right]\right.\\
&\quad \left.\left.+b_h\left[3m_{ll}^C + s_{c_l}\left(s_{c_l}-2+m_{ll}^C(s_{c_l}-4)\right)\right]\right)\right]\\
\Delta\left(2, m_{hl}^C, 0, m_{ll}^C\right) &= \frac{1}{2b_h b_l}\left[(c_h)^2 2b_l\right.\\
&\quad -4c_h c_l\left(b_l\left[\left(m_{hl}^C\right)^2 + m_{hl}^C m_{ll}^C + m_{ll}^C\right] + b_h m_{ll}^C\left[s_{c_l}-1\right]\right)\\
&\quad -\left(c_l\right)^2\left(b_l\left[3m_{hl}^C + s_{c_l}\left(2s_{c_l}-4+m_{hl}^C(s_{c_l}-4)\right)\right]\right.\\
&\quad \left.\left.+b_h m_{ll}^C\left[3+s_{c_l}\left(s_{c_l}-4\right)\right]\right)\right]\\
\Delta\left(0, m_{hl}^C, 2, m_{ll}^C\right) &= \frac{1}{2b_h b_l}\left[(c_h)^2 2b_h\right.\\
&\quad -4c_h c_l\left(b_l m_{hl}^C\left[s_{c_l}-1\right] + b_h\left[\left(m_{ll}^C\right)^2 + m_{hl}^C m_{ll}^C + m_{hl}^C\right]\right)\\
&\quad -\left(c_l\right)^2\left(b_l m_{hl}^C\left[3+s_{c_l}\left(s_{c_l}-4\right)\right]\right.\\
&\quad \left.\left.+b_h\left[3m_{ll}^C + s_{c_l}\left(2s_{c_l}-4+m_{ll}^C(s_{c_l}-4)\right)\right]\right)\right]\\
\Delta\left(1, m_{hl}^C, 1, m_{ll}^C\right) &= \frac{1}{2b_h b_l}\left[(c_h)^2\left(b_h+b_l\right)\right.\\
&\quad -2c_h c_l\left(b_l\left[\left(m_{hl}^C\right)^2 + m_{hl}^C m_{ll}^C + m_{ll}^C + m_{hl}^C\left(s_{c_l}-1\right)\right]\right.\\
&\quad \left.+b_h\left[\left(m_{ll}^C\right)^2 + m_{hl}^C m_{ll}^C + m_{hl}^C + m_{ll}^C\left(s_{c_l}-1\right)\right]\right)\\
&\quad -\left(c_l\right)^2\left(b_l\left[3m_{hl}^C + s_{c_l}\left(s_{c_l}-2+m_{hl}^C(s_{c_l}-4)\right)\right]\right.\\
&\quad \left.\left.+b_h\left[3m_{ll}^C + s_{c_l}\left(s_{c_l}-2+m_{ll}^C(s_{c_l}-4)\right)\right]\right)\right]
\end{aligned}$$

Bibliography

Abreu, D. and Sen, A. (1990). Subgame perfect implementation: A necessary and almost sufficient condition. *Journal of Economic Theory*, 50:285–299.

Alesina, A. and Drazen, A. (1991). Why are stabilizations delayed? *American Economic Review*, 81:1170–1188.

Alesina, A. and la Ferrara, E. (2000). Participation in heterogeneous communities. *The Quarterly Journal of Economics*, 115(3):847–904.

Anderson, L. R., Mellor, J. M., and Milyo, J. (2008). Inequality and public good provision: An experimental analysis. *Journal of Socio-Economics*, 37:1010–1028.

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

Andreoni, J. (1998). Toward a theory of charitable fund-raising. *Journal of Political Economy*, 106:1186–1213.

Aristotle (1984). Nicomachean ethics. In Barnes, J., editor, *The Complete Works of Aristotle*. Princeton University Press.

Babcock, L., Loewenstein, G., Issacharoff, S., and Camerer, C. (1995). Biased judgments of fairness in bargaining. *American Economic Review*, 85(5):1337–1343.

Babcock, L., Wang, X., and Loewenstein, G. (1996). Choosing the wrong pond: Social comparisons in negotiations that reject a self-serving bias. *The Quarterly Journal of Economics*, 111:1–19.

Bagnoli, M. and Lipman, B. (1989). Provision of public goods: Fully implementing the core through private contributions. *The Review of Economic Studies*, 56:583–601.

Bagnoli, M. and McKee, M. (1991). Voluntary contribution games: Efficient private provision of public goods. *Economic Inquiry*, 29:351–366.

Ball, S., Eckel, C., Grossmann, P. J., and Zame, W. R. (2001). Status in markets. *Quarterly Journal of Economics*, 116:161–188.

Barrett, S. (1994). Self-enforcing international environmental agreements. *Oxford Economic Papers*, 46:878–894.

Barrett, S. (1997a). Heterogeneous international environmental agreements. In Carraro, C., editor, *International Environmental Negotiations: Strategic Policy Issues*. Edward Elgar.

Barrett, S. (1997b). Towards a theory of international environmental cooperation. In Carraro, C. and Siniscalco, D., editors, *New direction in the economic theory of the environment*. Cambridge Univ Press.

Bergstrom, T., Blume, L., and Varian, H. R. (1986). On the private provision of public goods. *Journal of Public Economics*, 29:25–49.

Biancardi, M. and Villani, G. (2010). International environmental agreements with asymmetric countries. *Computational Economics*, 36:69–92.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bock, O., Nicklisch, A., and Baetge, I. (2012). hroot - hamburg registration and organization online tool. H-Lab Working Paper No. 1.

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

Bracha, A., Menietti, M., and Vesterlund, L. (2011). Seeds to succeed? sequential giving to public projects. *Journal of Public Economics*, 95:416–427.

Buckley, E. and Croson, R. (2006). Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics*, 90(4-5):935–955.

Cadsby, C. and Maynes, E. (1999). Voluntary provision of threshold public goods with continuous contributions: Experimental evidence. *Journal of Public Economics*, 71:53–73.

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

Carraro, C., Eyckmans, J., and Finus, M. (2006). Optimal transfers and participation decisions in international environmental agreements. *The Review of International Organizations*, 1:379–396.

Carraro, C. and Marchiori, C. (2003). Stable coalitions. In Carraro, C., editor, *The Endogenous Formation of Economic Coalitions*, chapter 5, pages 156–198. Edward Elgar, Cheltenham, UK.

Carraro, C., Marchiori, C., and Oreffice, S. (2009). Endogenous minimum participation in international environmental treaties. *Environmental Resource Economics*, 42:411–425.

Carraro, C. and Siniscalco, D. (1993). Strategies for the international protection of the environment. *Journal of Public Economics*, 52:309–328.

Chan, K. S., Mestelman, S., Moir, R., and Muller, R. A. (1996). The voluntary provision of public goods under varying income distributions. *The Canadian Journal of Economics*, 29:54–59.

Chan, K. S., Mestelman, S., Moir, R., and Muller, R. A. (1999). Heterogeneity and the voluntary provision of public goods. *Experimental Economics*, 2:5–30.

Chander, P. and Tulkens, H. (1995). A core-theoretic solution for the design of cooperative agreements on transfrontier pollution. *International Tax and Public Finance*, 2:279–293.

Chander, P. and Tulkens, H. (1997). The core of an economy with multilateral environmental externalities. *International Journal of Game Theory*, 26:379–401.

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

Chaudhuri, A. (2011). Substaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14:47–83.

Chaudhuri, A., Graziano, S., and Maitra, P. (2006). Social learning and norms in a public goods experiment with intergenerational advice. *Review of Economic Studies*, 73(2):357–380.

Cherry, T. L., Kroll, S., and Shogren, J. (2005). The impact of endowment heterogeneity and origin on public good contributions: Evidence from the lab. *Journal of Economic Behavior & Organization*, 57(2):357–365.

Clarke, E. H. (1971). Multipart pricing of public goods. *Public Choice*, 11(2):17–33.

Cox, J. C., Friedman, D., and Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59:17–45.

Croson, R. and Marks, M. (2000). Step returns in threshold public goods: A meta- and experimental analysis. *Experimental Economics*, 2:239–259.

Dahl, G. B. and Ransom, M. R. (1999). Does where you stand depend on where you sit? tithing donations and self-serving beliefs. *American Economic Review*, 89:703–727.

Dannenberg, A. (2012). Coalition formation and voting in public goods games. *Strategic Behavior and the Environment*, 2:83–105.

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

D'Aspremont, C., Jacquemin, A., Gabszewicz, J. J., and Weymark, J. A. (1983). On the stability of collusive price leadership. *The Canadian Journal of Economics*, 16(1):17–25.

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

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117.

Engel, C. (2011). Dictator games: a meta study. *Experimental Economics*, 14:583–610.

Erkal, N., Gangadharan, L., and Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. *American Economic Review*, 101(7):3330–3348.

Ertan, A., Page, T., and Putterman, L. (2009). Who to punish? individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53:495–511.

Eyckmans, J. and Finus, M. (2009). An almost ideal sharing scheme for coalition games with externalities. *Stirling Discussion Paper Series.*

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

Falkinger, J. (1996). Efficient private provision of public goods by rewarding deviations from average. *Journal of Public Economics*, 62(3):413–422.

Fehr, E. and Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4):185–190.

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

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

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

Finus, M. and Rundshagen, B. (2003). Endogenous coalition formation in global pollution control: a partition function approach. In Carraro, C., editor, *The endogenous formation of economic coalitions*. Edward Elgar Publishing.

Finus, M., Saiz, M. E., and Hendrix, E. M. (2008). An empirical test of new developments in coalition theory for the design of international environmental agreements. *Environment and Development Economics*, 1:1–21.

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

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Forsythe, R., Horowitz, J. L., Savin, N., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Frey, B. S. (1997). *Not Just for the Money: An Economic Theory of Personal Motivation*. Edward Elgar.

Frey, B. S. and Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15:589–611.

Fuentes-Albero, C. and Rubio, S. J. (2010). Can international environmental cooperation be bought? *European Journal of Operational Research*, 202:255–264.

Gächter, S., Renner, E., and Sefton, M. (2008). The long-rund benefits of punishment. *Science*, 322:1510.

Gerber, A. and Wichardt, P. (2009). Providing public goods in the absence of strong institutions. *Journal of Public Economics*, 93(4):429–439.

Gerber, A. and Wichardt, P. C. (2013). On the private provision of intertemporal public goods with stock effects. *Environmental and Resource Economics*, 55:245–255.

Glanemann, N. (2012). Can international environmental cooperation be bought: Comment. *European Journal of Operational Research*, 216(3):697–699.

Greiner, B. (2004). An online recruitment system for economic experiments. In Kremer, K. and Macho, V., editors, *Forschung und wissenschaftliches Rechnen 2003*, volume 63, pages 79–93. GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung.

Groves, T. (1973). Incentives in teams. *Econometrica*, 41(4):617–631.

Groves, T. and Ledyard, J. (1977). Optimal allocation of public goods: A solution to the "free rider" problem. *Econometrica*, 45(4):783–809.

Gürerk, O., Irlenbusch, B., and Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312:108–111.

Güth, W., Levati, M. V., Sutter, M., and van der Heijden, E. (2007). Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics*, 91:1023–1042.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3:367–388.

Güth, W. and Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology*, 11(3):417–449.

Hamman, J., Weber, R., and Woon, J. (2011). An experimental investigation of electoral delegation and the provision of public goods. *American Journal of Political Science*, 5:737–751.

Hofmeyr, A., Burns, J., and Visser, M. (2007). Income inequality, reciprocity and public good provision: An experimental analysis. *South African Journal of Economics*, 75(3):508–520.

Jackson, M. (1992). Implementation in undominated strategies: A look at bounded mechanisms. *Review of Economic Studies*, 59:757–775.

Keser, C., Markstädter, A., Schmidt, M., and Schnitzler, C. (2011). Rich man and lazarus - asymmetric endowments in public-goods experiments. *Working Paper*, Göttingen University.

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

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

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

Kroll, S., Cherry, T. L., and Shogren, J. F. (2007). Voting, punishment and public goods. *Economic Inquiry*, 45:557–570.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Kumru, C. S. and Vesterlund, L. (2010). The effect of status on charitable giving. *Journal of Public Economic Theory*, 12(4):709–735.

Lange, A., Löschel, A., Vogt, C., and Ziegler, A. (2010). On the self-interested use of equity in international climate negotiations. *European Economic Review*, 54:359–375.

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In Roth, A. E. and Kagel, J. H., editors, *The Handbook of Experimental Economics*. Princeton University Press.

Levati, M. V., Sutter, M., and van der Heijden, E. (2007). Leading by example in a public goods experiment with heterogeneity and incomplete information. *Journal of Conflict Resolution*, 51(5):793–818.

Lindroos, M. (2008). Coalitions in international fisheries management. *Natural Resource Modeling*, 21:366–384.

List, J. and Lucking-Reiley, D. (2002). The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign. *The Journal of Political Economy*, 110:215–233.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, 64:237–267.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press.

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

McGinty, M. (2007). International environmental agreements among asymmetric nations. *Oxford Economic Papers*, 59:45–62.

McGinty, M. (2011). A risk-dominant allocation: Maximizing coalition stability. *Journal of Public Economic Theory*, 13(2):311–325.

Meltzer, A. H. and Richard, S. F. (1981). A rational theory of the size of government. *Journal of Political Economy*, 89:914–27.

Moore, J. and Repullo, R. (1988). Subgame perfect implementation. *Econometrica*, 56:1191–1220.

Moxnes, E. and van der Heijden, E. (2003). The effect of leaderlead in a public bad experiment. *Journal of Conflict Resolution*, 47(6):773–795.

Neitzel, J. and Sääksvuori, L. (2014a). Normative conflict and cooperation in sequential social dilemmas. *mimeo*.

Neitzel, J. and Sääksvuori, L. (2014b). Wealth inequality, perceptions of normatively fair behavior and cooperation. *mimeo*.

Nikiforakis, N., Noussair, C., and Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics*, 96(9):797–807.

Osmani, D. and Tol, R. S. (2009). Toward farsightedly stable international environmental agreements. *Journal of Public Economic Theory*, 11(3):455–492.

Osmani, D. and Tol, R. S. (2010). The case of two self-enforcing international agreements for environmental protection with asymmetric countries. *Computational Economics*, 36:93–119.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3):137–158.

Ostrom, E., Gardner, R., and Walker, J. (1994). *Rules, Games, and Common-Pool Resources*. University of Michigan Press, Ann Arbor.

Palfrey, T. and Srivastava, S. (1991). Nash implementation using undominated strategies. *Econometrica*, 59:479–502.

Pavlova, Y. and de Zeeuw, A. (2013). Asymmetries in international environmental agreements. *Environment and Development Economics*, 18(Special Issue 01):51–68.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Harvard University Press.

Potters, J., Sefton, M., and Vesterlund, L. (2005). After you - endogenous sequencing in voluntary contribution games. *Journal of Public Economics*, 89:1399–1419.

Rege, M. and Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics*, 88:1625–1644.

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

Rivas, M. F. and Sutter, M. (2011). The benefits of voluntary leadership in experimental public goods games. *Economics Letters*, 112:176–178.

Rutz, S. (2001). Minimum participation rules and the effectiveness of multilateral environmental agreements. Working Paper No. 01/22, Centre for Economic Research, ETHZ, Zurich.

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

Samuelson, P. A. (1954). The pure theory of public expenditure. *The Review of Economics and Statistics*, 36:387–389.

Sefton, M., Shupp, R., and Walker, J. M. (2007). The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, 45:671–690.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-based choice. *Cognition*, 49:11–36.

Spiegler, R. (2002). Equilibrium in justifiable strategies: A model of reason-based choice in extensive-form games. *Riview of Economic Studies*, 69(3):691–706.

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

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

van Dijk, F., Sonnemans, J., and van Winden, F. (2002). Social ties in a public good experiment. *Journal of Public Economics*, 85:275–299.

Varian, H. R. (1992). *Microeconomic Analysis.* W.W. Norton & Company, Inc.

Varian, H. R. (1994a). Sequential contributions to public goods. *Journal of Public Economics*, 53:165–186.

Varian, H. R. (1994b). A solution to the problem of externalities when agents are well-informed. *American Economic Review*, 84(5):1278–93.

Vesterlund, L. (2003). The informational value of sequential fundraising. *Journal of Public Economics*, 87:627–657.

Walker, J. M., Gardner, R., Herr, A., and Ostrom, E. (2000). Collective choice in the commons: Experimental results on proposed allocation rules and votes. *Economic Journal*, 110:212–234.

Walker, M. (1981). A simple incentive compatible scheme for attaining lindahl allocations. *Econometrica*, 49:65–71.

Warr, P. G. (1983). The private provision of a public good is independent of the distribution of income. *Economics Letters*, 13:207–211.

Weikard, H.-P. (2009). Cartel stability under an optimal sharing rule. *The Manchester School*, 77:599–616.

Winter, F., Rauhut, H., and Helbing, D. (2012). How norms can generate conflict: An experiment on the failure of cooperative micro-motives on the macro-level. *Social Forces*, 90:919–946.

Eidesstattliche Versicherung

Hiermit erkläre ich, Jakob Neitzel, an Eides statt, dass ich die Dissertation mit dem Titel

Essays on Public Good Provision: Fair Contribution Rules and Institution Formation

selbständig und ohne fremde Hilfe verfasst habe.

Andere als die von mir angegebenen Quellen und Hilfsmittel habe ich nicht benutzt. Die den herangezogenen Werken wörtlich oder sinngemäß entnommenen Stellen sind als solche gekennzeichnet.