Institution Formation in Public Goods Games

By MICHAEL KOSFELD, AKIRA OKADA, AND ARNO RIEDL*

Sanctioning institutions are of utmost importance for overcoming free-riding tendencies and enforcing outcomes that maximize group welfare in social dilemma situations. We investigate, theoretically and experimentally, the endogenous formation of institutions in public goods provision. Our theoretical analysis shows that players may form sanctioning institutions in equilibrium, including those governing only a subset of players. The experiment confirms that institutions are formed and that it positively affects cooperation and group welfare. However, the data also shows that success is not guaranteed. Players are unwilling to implement equilibrium institutions in which some players have the opportunity to free ride. Our results emphasize the role of fairness in the institution formation process. (JEL C72, C92, D72*)*

> "Persons agree to constraints on their own liberties in exchange for comparable constraints being imposed on the liberties of others."
>
> — James M. Buchanan and Roger D. Congleton (1998, p. 4)

When markets fail, the design of appropriate institutions is a key issue for economic analysis and policy. Social dilemma situations (e.g., public goods, common pool resources), in which the pursuit of individual interests conflicts with the maximization of social welfare, are a classic example. In such situations, the

* Kosfeld: Department of Economics and Business Administration, Johann Wolfgang Goethe-University Frankfurt, Dantestrasse 9, D-60325 Frankfurt am Main, Germany, and IZA (e-mail: kosfeld@econ.uni-frankfurt.de); Okada: Graduate School of Economics, Hitotsubashi University, Kunitachi, Tokyo 186-8601, Japan (e-mail: aokada@econ.hit-u.ac.jp); Riedl: Department of Economics, Faculty of Economics and Business Administration, Maastricht University, P.O.Box 616, NL-6200 MD Masstricht, the Netherlands, and CESifo and IZA (e-mail: a.riedl@algec.unimaas.nl). We thank Elinor Ostrom, Armin Falk, Ernst Fehr, Urs Fischbacher, Simon Gchter, Klaus Schmidt, Christian Zehnder, and three anonymous referees for their helpful comments. Kosfeld gratefully acknowledges financial support from the University of Zurich through the University Research Priority Program on "Foundations of Human Social Behavior: Altruism versus Egoism" and the Swiss State Secretariat for Education and Research through the EU-TMR Research Network ENABLE (MRTN_CT-2003-505223). Okada gratefully acknowledges financial support from the Japan Society for the Promotion of Science under grant No.(A)16203011 and the Matsushita International Foundation.

implementation of a sanctioning institution that castigates individual behavior if it deviates from the welfare maximizing action is a widely used solution. For example, many common pool resources regimes around the world rely on sanctions and there is unanimous agreement in the literature that an effective sanctioning system is a major determinant of the success of such regimes (Jean-Marie Baland and Jean-Philippe Platteau, 1996; Elinor Ostrom, 1999).[1] Similarly, trade unions and employers' associations often have arbitration boards monitoring and enforcing the compliance of their members. Even in the international arena famous examples exist. For instance, the EU Stability and Growth Pact was created to enforce budgetary discipline among EU member states, and the Kyoto Protocol aims to reduce global greenhouse gas emissions by implementing legally binding agreements.

As diverse as these examples are, structurally, they have two important elements in common. Firstly, the institutional arrangements are not imposed from without but are *formed from within* in the sense that, at some point in time, a set of agents voluntarily agreed to implement the particular arrangement. Secondly, sanctioning applies only to members of the institution; non-members remain free in their choices and, hence, are given a strong incentive to free ride. Together, these two elements constitute what we term a "dilemma of endogenous institution formation": jointly, everyone profits if a sanctioning institution is formed, but each individual profits more if only the others form the institution.[2] It is exactly this dilemma of institution formation that we address in this paper.

The social dilemma situation we consider is a linear n-player public goods game. The institutional arrangement we analyze is a sanctioning institution, in which sanctions are imposed by a central authority, for example, a policeman, a court, or an arbitration board.[3] We model the

[1]Examples include irrigation systems (Shui Yan Tang, 1992), forests (Arun Agrawal and Gautam N. Yadama, 1997), and fisheries (Edvard Hviding and Graham B.K. Baines, 1994).

[2]The dilemma is a particular type of the so-called "second-order free-rider problem" (cf. Pamela Oliver, 1980).

[3]We thereby abstract from possible enforcement problems that might arise if players themselves have to impose the sanctions. These problems represent an important research question in their own, but will not be the topic of this paper, which focuses on the *formation* of institutions. Given this, however, our analysis of institution formation is

process of institution formation by a three-stage non-cooperative game: In the first stage of the game, each player decides whether he wants to participate in an organization that, once implemented, exerts a punishment on each member who does not contribute his full endowment to the public good. The organization is costly and only players who are members of the organization can be punished. Thus, non-members can free ride on members' contributions. In the second stage, players learn how many of the other players are willing to participate. The organization is implemented if and only if all players willing to participate agree to its actual formation. In the final stage, the public goods game is played.

In the theory part of our paper, we show that two different types of subgame perfect Nash equilibria exist in this game, a so-called *organizational equilibrium*, where players successfully implement an organization, and a so-called *status-quo equilibrium*, where no organization is implemented. We prove that organizations in any organizational equilibrium are of a minimum size s^*, i.e., at least s^* players participate, where s^* depends on the payoffs in the public goods game and the cost of the organization. Furthermore, using strictness (in every subgame) as an equilibrium refinement, we show that a unique strict subgame perfect equilibrium exists in terms of the organization size. In this equilibrium, exactly s^* players implement the organization and consequently contribute to the public good, whereas the remaining $n - s^*$ players do not participate and free ride. Thus, if $s^* < n$, the organization has a proper subset of players who voluntarily commit themselves to cooperation. Although each organization member would be better off if someone else participated instead of him or if more players became members, the organization is nevertheless implemented because each individual member earns a higher payoff than in the status-quo equilibrium, in which no organization is implemented and no players contribute to the public good. This result is related to a similar theoretical finding in the environmental economics literature. There, the notion of so-called "internal stability", a concept originally developed for the analysis of cartels and coalition formation, implies that self-enforcing environmental agreements may support only a small subgroup of signatories (Scott Barrett, 1994).

---

in fact rather general since most of our results can be extended to other (non-centralized) institutional arrangements (see below).

While the strict subgame perfect equilibrium prediction is intuitive in terms of individual material payoff maximization, the equilibrium outcome is clearly unfavorable in terms of efficiency and equality. If players dislike payoff inequality, it seems plausible that other institutions, in particular the grand organization where all players are members, may be favored. We, therefore, also analyze the institution-formation game under the assumption that some of the players suffer from payoff inequality as is captured by the social-preference model of Ernst Fehr and Klaus Schmidt (1999). We show that inequality aversion may indeed select the grand organization either as a strict or even as the unique organizational equilibrium.

The theory part shows that equilibrium selection depends on refinements and assumptions on preferences. Hence, theory alone gives only limited guidance regarding institution formation. We, therefore, present the results of a laboratory experiment in the second part of the paper. Our experiment goes beyond the existing literature as it connects the classic social dilemma situation with an innovative element of political organization, i.e., the endogenous formation of institutions.[4] In each of our main treatments, subjects played 20 rounds of a 4-player institution-formation game as described above. We varied the marginal per capita return of the public good across treatments to yield different predictions regarding the minimum organization s^*. We also conducted two corresponding control treatments, in which no institution could be formed and subjects only played the public goods game.[5]

[4]Most of the related experimental work has focused on the effect of exogenously imposed institutions (e.g., Toshio Yamagishi, 1986, 1988; Ostrom, James Walker, and Roy Gardner, 1992; Yan Chen and Charles R. Plott, 1996; Josef Falkinger, Fehr, Simon Gächter, and Rudolf Winter-Ebmer, 2000; Fehr and Gächter, 2000a, 2002; David Masclet, Charles Noussair, Steve Tucker, and Marie-Claire Villeval, 2003; Christopher M. Anderson and Louis Putterman, 2006; Jeffrey P. Carpenter, 2007a, 2007b). Only a couple of recent papers allow, at least to some extent, for endogenous institutional choice (Walker, Gardner, Andrew Herr, and Ostrom, 2000; Özgür Gürerk, Bernd Irlenbusch, and Bettina Rockenbach, 2006; Matthias Sutter, Stefan Haigner, and Martin G. Kocher, 2006; Jean-Robert Tyran and Lars P. Feld, 2006; Stephan Kroll, Todd L. Cherry, and Jason F. Shogren, 2007). Different to our study, however, the latter models do not give players the opportunity to free ride on other players' participation in the institution, an assumption that basically eliminates the second-order free-rider problem.

[5]In addition, we conducted a third control treatment in order to check for possible experimental design effects (see details below).

Our main experimental findings are as follows. First, subjects successfully establish organizations. In both experimental treatments, from 70 to 100 percent of all groups implement an organization by the final rounds. Second and most importantly, the majority (on average, around 75 percent) of the organizations implemented are grand organizations, i.e., *all* players participate. This finding is consistent with the prediction based on social preferences and stands in stark contrast to the strict subgame perfect equilibrium prediction of the standard model. Further results on players' beliefs and rates of implementation show that the frequent observation of the grand organization is not driven by miscoordination among participating players. Instead, the data suggest that the institutional outcome is the result of (almost) equilibrium play. Finally, a comparison with our control treatments confirms that the opportunity to form institutions increases and stabilizes total contributions to the public good. Overall, institution formation enhances group welfare, despite the fact that it is costly.

Our theoretical and empirical results have important implications for public policy. First, since sanctioning institutions may be an effective solution in many social dilemma situations, the observation that subjects *voluntarily* implement such institutions can be taken as good news. However, subjects are very reluctant to implement (Nash equilibrium) institutions that govern only a subset of players. This is true even if participating players can earn a higher payoff compared to the non-production of the public good. Our result emphasizes the importance of fairness for the formation and stability of institutional arrangements, an issue which has frequently been documented, for example, in the common pool resource literature (Baland and Platteau, 1996; Margaret A. McKean, 2000) and has also been stressed by prominent public choice scholars (Buchanan and Congleton, 1998, see the quote at the beginning of the paper). As a famous instance, it may also bring to mind the discussion of the potential impact of the United States' withdrawal from the Kyoto Protocol on other nations' motivation to fulfill the agreement.[6] The

[6]To give two illustrative examples — in its 2006 report on climate strategies, the Dutch Scientific Council for Government Policy (WWR) advised the Dutch government not to stick (too tightly) to the Kyoto criteria, one reason being that large countries such as the U.S. did not ratify the agreement (WWR, 2006). Likewise, at the time of the U.S.'s withdrawal from the protocol in 2001, Australian Environment Minister Robert Hill declared that he did

consequences of the observed behavior are twofold. On the one hand, if the process of institution formation is successful, established institutions generally achieve a high level of efficiency because strictly more than the minimally required s^* players participate. In fact, in the most frequently observed organization in our experiment, 100 percent of the players participate and contribute to the public good. This is in stark contrast to the prediction based on s^*, yielding a participation (and cooperation) rate in the two treatments of only 50 and 67 percent, respectively. Yet, on the other hand, the risk for the process to fail is much higher than predicted as well. Institutions are rejected that from an individual as well as a social welfare perspective clearly represent a material improvement over the situation without an institution. Therefore, not taking this behavior into account not only yields misleading theoretical predictions, but may lead to the realization of highly inefficient outcomes.

While we focus on a particular institutional solution in this paper (i.e., centralized sanctioning), our analysis can easily be extended to other institutional arrangements, including alternative centralized policy instruments, such as the mechanisms proposed by Theodore Groves and John Ledyard (1977) and Falkinger (1996), but also non-centralized solutions, such as repeated-game trigger strategies. The only condition that must be fulfilled is that the particular institution "works," i.e., that participating players have an incentive to act in accordance with the institutional rules and contribute to the public good once the institution is formed. In game theoretic terms, the prescribed behavior must form a Nash equilibrium. This holds for the Grove-Ledyard and the Falkinger mechanisms given that parameters are chosen accordingly. It also holds for repeated-game trigger strategies, if players are sufficiently patient. The key question for any particular institutional solution, however, is whether players will actually agree to form the institution, the main problem being that each player has an incentive to free ride on others forming the institution. This second-order free-rider problem applies to any mechanism that solves the first-order free-rider problem in social dilemma situations. The contribution of our paper is to

not think "the Kyoto Protocol will succeed without the United States" (ABC, The World Today, March 30, 2001; http://www.abc.net.au/worldtoday/stories/s269266.htm).

show how individuals can overcome this problem, both from a theoretical and from an experimental viewpoint, and to point out behavioral regularities that govern and limit the process of endogenous institution formation.

The paper is organized as follows. Section I theoretically analyzes the institution formation game, characterizing subgame perfect Nash equilibria both if players have standard preferences and if players have fairness preferences. Section II describes and analyzes the experiment. Section III concludes.

I. Institution Formation: Theory

A. Model

Consider the following n-player public goods game. There are $n \geq 2$ players, each of whom has a private endowment $w > 0$ from which he can contribute $g_i \leq w$ to a public good. Given the contribution of all players $(g_1, \ldots, g_n)$, player i's material payoff is equal to

$$\pi_i(g_1, \ldots, g_n) = w - g_i + a \sum_{j=1}^{n} g_j, \tag{1}$$

where $0 < a < 1 < na$. Parameter a models the marginal per capita return (MPCR) from contributing to the public good. Assumption $a < 1$ implies that zero contribution is the dominant action for every player with standard preferences, i.e., each player's material welfare is maximized by contributing zero to the public good regardless of the other players' contributions. In consequence, the strategy profile $(0, \ldots, 0)$ is the unique Nash equilibrium. Assumption $na > 1$ implies that all players are better off if everyone contributes his full endowment to the public good. In particular, $(w, \ldots, w)$ is the welfare maximizing strategy profile.

Generally, institution formation is a complex process. Parties are typically involved in multi-stage bargaining with continual updates about other parties' behavior, goals, and expectations. Often, the process is little structured ex-ante and negotiations take the form of both bilateral and collective bargaining. The institution-formation game we analyze in this paper is necessarily simpler. It consists of a participation stage where players announce their (un)willingness to

form an institution and an implementation stage where they can actually form the institution. The advantage of this set-up is that we can provide precise game-theoretic predictions regarding institutional outcomes. At the same time, our model captures key elements of real-world negotiation processes. First, economic or political actors do not implement new institutions ad hoc but take decisions step-by-step. Second, players receive important information about other players' willingness to form an institution before they actually decide to implement it. The precise sequence of actions in our institution formation game is as follows:[7]

1. Participation stage: Players simultaneously and independently announce whether or not they are willing to participate in an organization that sanctions all organization members who do not contribute their full endowment to the public good. In the following, players who declare such a willingness are called *participants*; those who do not are called *non-participants*.

2. Implementation stage: After players are informed about the set of participants, all participants negotiate about whether or not to implement an organization. Negotiations take the form that all participants simultaneously and independently either accept or reject the implementation of the organization. The organization is implemented if and only if all participants accept (unanimity rule). In case an organization is implemented, all participants become *members* of the organization. Non-participants cannot become members. The organization is costly. Costs arise only if an organization is implemented and are shared equally among organization members.

3. Contribution stage: All players simultaneously and independently determine their contribution to the public good. If an organization has been implemented, organization members will be sanctioned for not contributing their full endowment to the public good. Importantly, non-members cannot be sanctioned. If no organization has been implemented, no player is sanctioned.

[7]The institution-formation game laid out in this paper is an extension of the model proposed in Akira Okada (1993).

A player's final payoff u_i in the institution-formation game is determined as follows. Let S be the set of players who are members of the organization with $s = |S|$, and let $c \geq 0$ be the cost of the organization. There are two cases.

Case 1: If $S \neq \emptyset$, i.e., an organization is implemented, then for every player i

$$u_i = \begin{cases} w - g_i + a\sum_{j=1}^{n} g_j - \frac{c}{s} - p(g_i) & \text{if} \quad i \in S \\ w - g_i + a\sum_{j=1}^{n} g_j & \text{if} \quad i \notin S, \end{cases} \tag{2}$$

where $p(g_i)$ is the sanction imposed on member i satisfying[8]

$$p(g_i) = \begin{cases} w - g_i & \text{if} \quad g_i < w \\ 0 & \text{if} \quad g_i = w. \end{cases} \tag{3}$$

Case 2: If $S = \emptyset$, i.e., no organization is implemented, then for every player i

$$u_i = w - g_i + a\sum_{j=1}^{n} g_j. \tag{4}$$

Equation (2) reveals the key difference between members and non-members of the organization. While the former are punished for free-riding and share the costs of the organization, the latter can freely decide whatever to contribute and do not pay any part of the organization costs. Formally, the institution-formation game is a n-player three-stage game with perfect information. In each stage, players choose their actions with perfect knowledge about the course of the game in previous stages.[9] In the following, we first characterize the set of subgame perfect equilibria of the institution formation game if players' preferences are captured by material payoffs u_i. We then analyze equilibria of the game if (some) players have social preferences with a taste for fairness.

[8]Note that $p(g_i)$ must be larger than $(1-a)(w-g_i)$ whenever $g_i < w$ for punishment to induce full contribution by organization members.

[9]All theoretical results are also valid if players in the implementation stage are only informed about the number of participants and if players in the contribution stage are only informed about whether the organization has been implemented or not. The subgame perfect equilibrium concept should then be replaced by a sequential equilibrium. The reason is — as will be shown below — that players' payoff functions depend on the number of participants (not their identity) and on whether or not an organization is established.

B. Standard Preferences

In a subgame perfect equilibrium, players decide on their actions in every stage rationally anticipating the outcome of future stages by applying backward induction. Consider first the contribution stage. If players' preferences are given by u_i, (2) and (3) imply that it is optimal for organization members to contribute w in the contribution stage once an organization has been implemented. Since non-members are not punished, they behave optimally by not contributing anything. Clearly, zero contribution is optimal for every player in the contribution stage if no organization has been implemented.

A key insight of this model is that, although players have an individual incentive to free ride in the public goods game, they might increase their payoff by coordinating their contributions in the framework of an organization. Note that in equilibrium organization members earn $asw - c/s$ if an organization has been implemented and that everybody earns w if no organization is established. Players are better off joining an organization compared to the zero-contribution outcome if the number of members s is such that

$$asw - \frac{c}{s} > w. \tag{5}$$

Let s^* denote the smallest non-negative integer s satisfying condition (5). The threshold s^* gives the minimum size of an organization such that participants in the implementation stage have an incentive to implement it. From $a < 1$ and $c \geq 0$ it follows that $s^* \geq 2$. Moreover, since the left-hand side of (5) is strictly increasing in s, $s^* \leq n$ exists uniquely if and only if $(an-1)nw > c$. If the latter condition does not hold, an organization is never beneficial to the players, i.e., no group would ever have an incentive to implement it. In the following, we therefore assume that the condition holds and hence a unique threshold $s^* \leq n$ exists.

We can now characterize the set of subgame perfect equilibria of the institution formation game. For convenience, we call a subgame perfect equilibrium an *organizational equilibrium* if an organization is formed on the equilibrium path. A subgame perfect equilibrium is called a *status-quo equilibrium* if no organization is formed.

Proposition 1 *If players have standard preferences, there exists an organizational equilibrium with s players being members of the organization if and only if* $s \geq s^*$. *For any number of participants s* $(1 \leq s \leq n)$ *there exists a status-quo equilibrium.*

All proofs are given in the supplementary material (Web Appendix A). Proposition 1 shows that players can, in principle, overcome the dilemma of endogenous institution formation. The only requirement is that at least s^* players participate. If this condition holds, there exists an organizational equilibrium for every $s \geq s^*$, in which players implement an organization of size s rejecting all organizations of different size (cf. the proof for details). However, success is not guaranteed. For any number of participants there always exists a status-quo equilibrium, in which players refrain from implementing an organization.[10] Moreover, if s^* is strictly smaller than n, the institution-formation game has multiple organizational equilibria, namely all organizations of size $s \in \{s^*, s^*+1, \ldots, n\}$.[11] In addition to agreeing on whether or not an organization shall be implemented, players then face two coordination problems. First, they have to solve the problem with regard to the organization size; second, if $s < n$, they also have to solve the problem with regard to who is going to become a member of the organization and who is going to stay out. Interestingly, as the following proposition shows, strictness as an equilibrium refinement might offer a possible solution to the first coordination problem. Generally, a Nash equilibrium of a strategic-form game is called *strict* if every player plays a unique best response to the other players' strategies, i.e., every player is strictly worse off by deviating from equilibrium play. A subgame perfect equilibrium of a multi-stage game is called strict if it induces a strict Nash equilibrium in every stage game.

[10] In a certain sense, the provision of the second-order public good (i.e., the sanctioning system) in equilibrium is thus characterized by a step-level technology with critical threshold s^*. Different from common step-level public goods (see, e.g., Mark Bagnoli and Barton L. Lipman, 1989; Rachel T. A. Croson and Melanie Beth Marks 2000), however, the step-level provision in our model is an equilibrium outcome. Furthermore, the threshold s^* is not given exogenously but is determined endogenously by players' incentives to form an institution.

[11] There also exists a mixed-strategy equilibrium in this case, in which players participate with positive probability and implement the organization whenever $s \geq s^*$.

Proposition 2 *If players have standard preferences, the institution formation game has a unique strict subgame perfect equilibrium in terms of the organization size. In this equilibrium exactly s^* players become members of the organization.*

Strictness of equilibrium yields a clear-cut prediction regarding organization size: exactly the minimum number of players s^* required for the organization to be individually profitable form the organization while the remaining players do not participate. Unless $s^* = n$, players are thus divided into two proper subsets: those who voluntarily implement the sanctioning institution, hence contributing to the public good, and those who do not participate and do not contribute. The intuition for this result is straightforward. The best that can happen in material terms to any player i is that *other* players implement an organization and contribute to the public good while i free-rides. As long as the number of participants s is strictly larger than s^* (and hence $s-1 \geq s^*$), there is always at least one participant who can successfully choose this option. The reason is that in a strict subgame perfect equilibrium, every organization with at least s^* participants is implemented in the implementation stage. Only if the organization is at its minimum size s^*, free-riding is no longer an option since the remaining $s^* - 1$ players will not implement the smaller organization.[12]

Whether or not the organizations form and whether or not they have the predicted size is of course an empirical question. An alternative equilibrium outcome, which always exists and which is favored in terms of efficiency, symmetry, and equality is the grand organization $s = n$. This organization is also consistent with the so-called *generality principle* of Buchanan and Congleton (1998) which asserts that political choices should be non-discriminatory and must involve equal treatment of all individuals. However, the grand organizational equilibrium requires players to reject all organizations with less than n participants. Such rejections are weakly dominated whenever $s \geq s^*$, since each participant is better off in material terms by implementing the orga-

[12]The strict subgame perfect equilibrium prediction nicely mirrors the concept of "internal stability" developed in the cartel and coalition formation literature (cf. Claude d'Aspremont, Alexis Jacquemin, Jean Jaskold Gabszewicz, and John A. Weymark, 1983; Carlo Carraro and Domenico Siniscalco, 1993; Barrett, 1994).

nization. If players maximize only their material payoff, it seems questionable whether they can credibly commit to reject all organizations smaller than n. Yet, intuition suggests that motives of fairness might induce players to reject organizations where non-members can free ride. In order to see whether fairness affects equilibrium outcomes we analyze the institution-formation game if (some) players have social preferences.

C. Social Preferences

There exists considerable evidence that social preferences, such as a taste for fairness, equity, and efficiency affect economic behavior in many important areas including the provision of public goods. For an overview see, e.g., Fehr and Gächter (2000b), Colin Camerer (2003). We analyze the impact of social preferences on institution formation using the inequity-aversion model suggested by Fehr and Schmidt (1999).[13] Suppose that players' material payoffs are given by the vector $u = (u_1, \ldots, u_n)$. Player i's utility U_i is then defined as

$$U_i = u_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{(u_j - u_i), 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{(u_i - u_j), 0\}. \tag{6}$$

The two parameters α_i and β_i measure player i's utility loss from disadvantageous inequality and from advantageous inequality, respectively. Following Fehr and Schmidt (1999), we assume that $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$ for all i. We first analyze the case when players' utility loss from advantageous inequality is small $(\beta_i < 1 - a)$ and behavior is driven by players' disutility from disadvantageous inequality. We then consider the case when some of the players suffer also strongly from advantageous inequality $(\beta_i > 1 - a)$.

Suppose that $\beta_i < 1 - a$ for all $i = 1, \ldots, n$. This implies that zero contribution to the public good is the dominant action for every player who is not a member of an organization (cf. Proposition 4 of Fehr and Schmidt (1999). In consequence, if an organization of size s is implemented,

[13]There exist other models of social preferences (e.g., Matthew Rabin, 1993; Gary Bolton and Axel Ockenfels, 2000; Gary Charness and Rabin, 2002; Martin Dufwenberg and Georg Kirchsteiger, 2004; Armin Falk and Urs Fischbacher, 2006; James C. Cox, Daniel Friedman, and Steven Gjerstad, 2007). We use the model of Fehr and Schmidt (1999) for reasons of simplicity and tractability.

organization member i's utility, is given by

$$asw - \frac{c}{s} - \frac{\alpha_i}{n-1}(n-s)\left(w+\frac{c}{s}\right). \tag{7}$$

The utility of an organization member is equal to his material payoff $asw - \frac{c}{s}$ minus some disutility that is generated by the difference between his payoff and the payoff of non-members. As in the case of standard preferences, we can calculate the minimum size s_i^+ of an organization to be profitable for player i, which is the smallest non-negative integer s satisfying

$$asw - \frac{c}{s} - \frac{\alpha_i}{n-1}(n-s)\left(w+\frac{c}{s}\right) > w. \tag{8}$$

Comparing (8) and (5) immediately reveals that the threshold s_i^+ for a player suffering from disadvantageous inequality ($\alpha_i > 0$) is larger than the one for a player with standard preferences ($\alpha_i = 0$). As players become more inequity averse, the threshold rises. When players have identical α_i's, basically all results from Section I.B carry over to the social-preference case with a new threshold $s^+ \equiv s_i^+$ defined by equation (8). It is easy to see that for sufficiently strong social preferences the grand organization is a unique organizational equilibrium ($s^+ = n$) when standard preferences predict multiple organizational equilibria ($s^* < n$).

If players differ in their concern for inequity, the situation is slightly more complex. First, Proposition 1 generalizes in the sense that an organizational equilibrium with participant set S exists if and only if equation (8) holds for every participant $i \in S$. Since there is no payoff inequality in the grand organization, the latter is always an equilibrium. The interesting question is, whether motives of fairness might induce this equilibrium to be strict (i.e., players are strictly worse off when deviating from the equilibrium) or even the unique organizational equilibrium.

Proposition 3 *Suppose that $\beta_i < 1 - a$ for all $i = 1, \ldots, n$. The grand organization is a strict organizational equilibrium if and only if there exist at least two players with $\alpha_i > \tilde{\alpha}$, where*

$$\tilde{\alpha} = \frac{(n-1)^2((n-1)a-1)w-(n-1)c}{(n-1)w+c}, \tag{9}$$

If at least $n-1$ players satisfy $\alpha_i > \tilde{\alpha}$, implementation of the grand organization is the unique organizational equilibrium.

The intuition for Proposition 3 is as follows. In Web Appendix A we show that an organizational equilibrium with participants S is a strict equilibrium if and only if each participant is pivotal, i.e., $S \setminus \{i\}$ is no organizational equilibrium for every $i \in S$. For the grand organization this means that no group of players is willing to implement an organization of size $n-1$. Threshold $\tilde{\alpha}$ guarantees that a participating player suffers sufficiently from disadvantageous inequality induced by the free-riding of the non-member to reject an organization of size $n-1$. At least two such players are required to ensure that organizations of size $n-1$ are never formed. Since a participating player's utility is increasing in s (cf. (7)), the grand organization is the unique organizational equilibrium if at least $n-1$ players are sufficiently inequality averse as this implies that no group of players is willing to implement any organization smaller than n.

Two things are worth mentioning at this point. First, the condition guaranteeing uniqueness of the grand organization equilibrium in Proposition 3 is sufficient but not necessary. This follows because the requirement for α_i rendering an organization of size s unattractive becomes less restrictive when s falls as participants' corresponding utility falls as well (cf. the left-hand side of (8)). For example, in one of our experimental treatments we present in the next section (IF40) the minimal degree of inequality aversion α_i to reject a three-player organization equals 0.48 whereas the minimal α_i to reject a two-player organization is even negative. There is no equilibrium in which players implement a two-player organization in this case. Thus, the grand (i.e., four-player) organization is the unique organizational equilibrium whenever condition (9) is satisfied for at least two players. Second, a *ceteris paribus* increase in the MPCR of the public good raises the threshold $\tilde{\alpha}$. The reason is that the material payoff of organization members rises, while the payoff inequality between members and non-members remains unchanged. In consequence, rejecting an organization becomes relatively more costly and players concern for inequality must be stronger to keep rejection a best response. In our experiment, for example, we implement treatments with $a = 0.4$ (IF40) and $a = 0.65$ (IF65). In the first treatment, $\tilde{\alpha} = 0.48$, in the second $\tilde{\alpha} = 2.23$.

Suppose now that there exist some players who suffer also strongly from advantageous inequality, i.e., $\beta_i > 1-a$ for some i. Let $B = \{i|\beta_i > 1-a\}$ denote the set of these players. Due

to their concern for advantageous inequality, players in B might contribute to the public good even when they are non-members of the organization. As such contributions increase the payoff of organization members (both by an increase in the material payoff and by a decrease in payoff inequality), condition (8) is now relaxed. In consequence, participating players who suffer from disadvantageous inequality are less likely to reject an organization given that non-members contribute to the public good, as well.

Nevertheless, organizations smaller than n may fail to be an equilibrium. The reason is that players with sufficiently large β_i prefer to be members of the grand organization. While uniqueness of the grand organization is more difficult to obtain, Proposition 4 shows that under comparable conditions as in Proposition 3 the grand organization is again a strict organizational equilibrium.

Proposition 4 *Suppose* $B = \{i|\beta_i > 1-a\} \neq \emptyset$. *The grand organization is a strict organizational equilibrium if (i) there exist at least two players with* $a_i > \tilde{\alpha}$ *and (ii) for each player* $i \in B, \beta_i > 1 - \frac{1}{n}$. *Only condition (ii) is relevant if* $B = \{1, \ldots, n\}$.

Proposition 4 can be illustrated by the following example. Suppose there exists one player i who suffers from advantageous inequality $\beta_i > 1-a$ while the remaining players suffer only (if at all) from disadvantageous inequality. Proposition 4 says that the grand organization is a strict organizational equilibrium if $\beta_i > 1 - \frac{1}{n}$ and at least two players (possibly including player i) have $\alpha_j > \tilde{\alpha}$. The latter condition is the same as in Proposition 3 and ensures that any organization of size $n-1$ is rejected by at least one participating player given that the single non-participant does not contribute to the public good. The condition on β_i additionally guarantees that in case player i does not participate, i himself is strictly worse off compared to joining the grand organization, and hence strictly prefers to participate. Obviously, if all players are sufficiently averse to advantageous inequality, condition (i) is of no relevance since condition (ii) already guarantees that each player strictly prefers to participate in the grand organization.

Let us summarize the theoretical analysis. If players have standard preferences, i.e., maximize their material payoff, the institution-formation game has multiple organizational equilibria

of size $s \in \{s^*, \ldots, n\}$. In addition, there exists a status-quo equilibrium for any number of participating players, in which no organization is formed. Strictness as an equilibrium refinement selects the smallest organizational equilibrium of size s^*. While this prediction is intuitive in terms of individual material payoff maximization, the resulting equilibrium outcome is unfavorable in terms of symmetry, equality and efficiency. If players dislike payoff inequality, the prediction changes in so far as smaller organizations are no longer an equilibrium. In fact, depending on players' degree of inequity aversion the grand organization, where all players participate, may be the unique or at least a strict organizational equilibrium.

The analysis shows that theory alone gives only limited guidance regarding the expected outcome of institution formation in our set-up. Due to the multiplicity of equilibria and different possibilities for refinement, several outcomes are possible and equally plausible.[14] In the following, we therefore present a laboratory experiment designed to investigate the process of institution formation in public goods games.

II. Institution Formation: Experiment

A. Procedural Details

In order to keep the complexity of the experiment low we slightly modified the institution formation game in the experiment. Once an organization was implemented, members of the organization did not make a decision in the contribution stage, but were bound to contribute their full endowment to the public good. Otherwise, everything else was the same as described above. The reason for this modification is that we want to focus on the problem of institution formation rather than on the separate issue of institutional enforcement.[15] The basic structure of the

[14]The multiplicity of equilibrium predictions becomes even larger once we consider a repeated-game set-up. In this case, any combination of Nash equilibria of the one-shot game is a subgame perfect equilibrium of the repeated game.

[15]To examine whether the modification has an effect on the experimental results, we conducted an additional control treatment (IF40$^+$), where all subjects were free to decide in the contribution stage but members of the organization were effectively punished if they did not contribute everything to the public good. In this treatment,

experimental game was as follows.[16]

At the beginning of each round of the experiment, each of four players receives an endowment of 20 points (i.e., $n = 4$ and $w = 20$). Each player then decides whether he wants to participate in an organization or not (*participation stage*).[17] After being informed about the number of players who want to participate, each participant decides whether or not he wants to implement the organization (*implementation stage*). The organization is implemented if and only if all participants decide to implement it. Non-participants do not make any decision in this stage and are only informed about the number of participants. Finally, players simultaneously determine the amount of their contributions to the public good (*contribution stage*). If an organization is implemented in the implementation stage, all members of the organization are bound to contribute their full endowment to the public good. Non-members, after being informed about the size of the implemented organization, freely determine the amount of their contributions. If no organization is implemented, all players freely determine the amount of their contributions.

Since the decision to participate may constitute a nontrivial coordination problem — in particular, if only a subset of players wants to form an organization — we elicited players' beliefs in the participation stage. Precisely, after players had decided whether to participate in the organization, each player was asked to indicate his expectation about the total number of players participating in the organization. Players were rewarded for correct predictions according to the quadratic scoring rule.[18]

We implemented two experimental treatments with different minimum size s^* for an organization to be materially profitable. In both treatments, the cost of the organization was set to

---

the somewhat more complex decision environment slowed down learning, otherwise no significant differences were found in comparison to our main treatment. Details are reported in the supplementary material (Web Appendix B).

[16]Experimental instructions are provided in the supplementary material (Web Appendix C).

[17]In the experimental instructions, we did not use the terms "organization" or "institution." Instead, subjects were asked if they were willing to bind themselves to contribute their full endowment.

[18]Quadratic scoring rules are known to be incentive compatible and have successfully been used in a number of experiments, for example, by Theo Offerman (1997) and Yaw Nyarko and Andrew Schotter (2002).

$c = 2$. In the first treatment (IF40), the MPCR of the public good $a = 0.4$ resulting in $s^* = 3$. In the second treatment (IF65), $a = 0.65$ yielding $s^* = 2$. In addition to these treatments, we implemented two control treatments (PG40 and PG65), in which players played the corresponding public goods game without the possibility of institution formation. Irrespective of the treatment, subjects played 20 rounds of the corresponding game with the same group of players (partner matching). All experiments were run at the CREED laboratory at the University of Amsterdam. In total, 164 subjects participated in the experiment, whereby 44 subjects participated in each of the institution formation treatments (IF40, IF65), 40 subjects participated in treatment PG40, and 36 participated in treatment PG65.[19] No subject participated in more than one treatment. Each session lasted about 120 minutes. On average, a subject earned €23.90 (about $25) in the experiment.

B. Results

In the results section we proceed as follows. We first analyze if subjects implement any organizations at all. Answering this question in the affirmative, we then study what kind of organizations are implemented. We also consider players' beliefs about the other players' participation decision and investigate the probability that an organization of a particular size is implemented in the implementation stage. Finally, we analyze the overall impact of institution formation on the provision of the public good by comparing average contributions and consequential levels of efficiency in the institution formation treatments and the corresponding control treatments.

Our first result shows that players almost always initiate an organization, and also implement the initiated organization in between 43 and 61 percent of the cases.

Result 1 *In treatment IF40, there is always at least one player per group who wants to establish an organization and an organization is implemented 43 percent of the time. In treatment IF65, in 98 percent of the cases, at least one player per group initiates an organization and an organization is implemented in 61 percent of the cases.*

[19] Another 52 subjects participated in the additional control treatment IF40$^+$.

Support for Result 1 is presented in Table 1, which summarizes the absolute and relative number of cases in which at least one player decides to participate in the participation stage (*initiated organizations*) and in which an organization is implemented by unanimous vote in the implementation stage (*implemented organizations*). While players always initiate an organization in treatment IF40, there are four cases in treatment IF65 in which a group of players does not initiate an organization (2 percent). At the same time, slightly more organizations are implemented in treatment IF65 than in treatment IF40 (132 vs. 95). None of these differences are statistically significant (Mann-Whitney test, $p > 0.14$).[20]

Table 1: Initiated and implemented organizations

	Treatment			
	IF40		IF65	
	Number	Percentage	Number	Percentage
Initiated organizations	220	100	216	98
Implemented organizations				
Total	95	43	132	61
One member	0	0	5	4
Two members	1	1	15	11
Three members	15	16	22	17
Four members	79	83	90	68

Note: The table presents the absolute and relative number of initiated and implemented organizations over all rounds. Relative numbers are calculated as follows: initiated organizations relative to all rounds, implemented organizations relative to all initiated organizations, different size of organizations relative to all implemented organizations.

Result 1 shows that players overcome the second-order free-rider problem and successfully establish organizations. Our next result reveals which organizations are implemented.

Result 2 *In both IF treatments, the large majority of implemented organizations are grand organizations, i.e., all players become members. Organizations of size $s < s^*$ are very rarely observed. Moreover, the observed increase of implemented organizations over time is solely due to an increase of grand organizations.*

[20]Statistical tests are based on group averages as units of observation. We report the results of two-sided tests throughout the paper.

Table 1 (lower part) shows the distribution of implemented organizations in the two IF treatments. The data speak clearly: Independent of treatment, the majority of organizations that are implemented include all four players. In treatment IF40, 79 of 95 organizations that are implemented are grand organizations (83 percent). In treatment IF65, 90 of 132 organizations are grand organizations (68 percent). In addition, players almost never implement organizations of less than s^* players. Recall that $s^* = 3$ in treatment IF40 and $s^* = 2$ in treatment IF65. Overall, only 1 (4) percent of the implemented organizations comprise less than s^* players in treatment IF40 (IF65). Thus, threshold s^* serves as a good prediction of the minimum size of an implemented organization. However, it clearly fails as a prediction of the maximum size of an organization (cf. Proposition 2). Rather than seeing only three or two players establishing an organization, we observe that most of the time an organization involves all four players. On average, the size of an implemented organization is slightly smaller in treatment IF65 than in treatment IF40 (3.49 vs. 3.82), but the difference is not significant (Mann-Whitney, $p = 0.20$).

The implementation of an organization is of course a rather complex process. It seems likely that players learn the benefits of establishing organizations in the course of the experiment and that the number of organizations implemented increases over time. Our data show that this is indeed the case. The Spearman rank order correlation between the number of implemented organizations and the round in the experiment is highly significant in both treatments (IF40: $\rho = 0.85, p = 0.00$; IF65: $\rho = 0.64, p = 0.00$). Moreover, the increase in implemented organizations is exclusively driven by the implementation of the grand organization. While the number of implemented grand organizations increases significantly over rounds (Spearman rank order correlation; IF40: $\rho = 0.87, p = 0.00$; IF65: $\rho = 0.72, p = 0.00$), the corresponding number of organizations with three or less members does not change significantly (Spearman rank order correlation; $p > 0.41$ in both treatments). Figure 1 illustrates the learning pattern by comparing the distribution of implemented organizations in early and in late rounds. As can be seen the share of the grand organization increases from 70 and 48 percent in rounds 1 to 5 of treatment IF40 and IF65, respectively, to 86 and 60 percent in rounds 16 to 20. At the same time, the relative number of smaller organization falls.

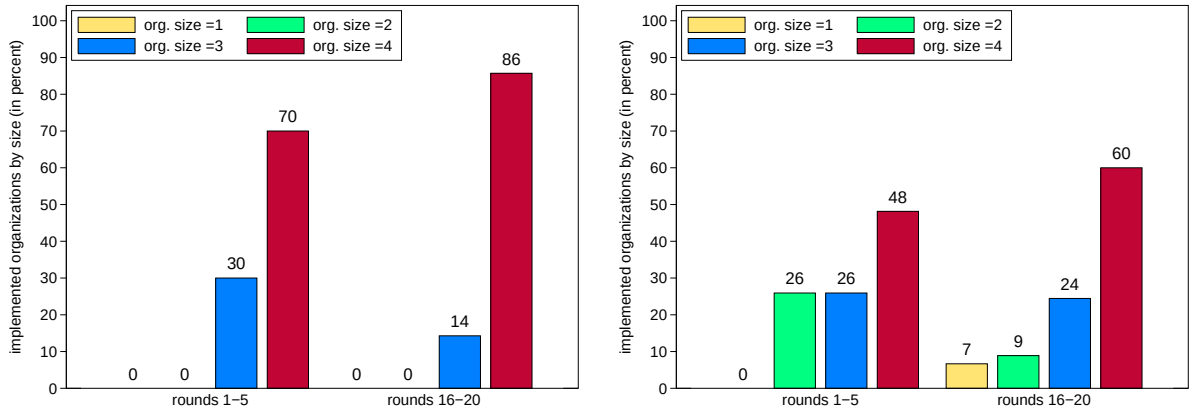

Figure 1: Distribution of implemented organizations in early and late rounds (left panel: IF40; right panel: IF65).

Why do we observe so many organizations larger than s^* in the experiment? Is it that players aim to implement the s^* organization but miscoordinate in the participation stage? Or, do participating players target at the grand organization and reject organizations that are smaller in the implementation stage? The following two results shed light on the driving forces of subjects' behavior. We first show that players who participate in the participation stage mostly expect that all other players will participate as well. Secondly, we show that initiated organizations comprising less than four participants are rejected with high probability in the implementation stage, even if no less than s^* players participate. These results cast doubt on the explanation that high participation rates are due to miscoordination. They rather suggest that players play the grand organizational equilibrium.

Result 3 *In both IF treatments, most players who participate in the organization in the participation stage expect that all other players will participate as well.*

Support for Result 3 is presented in the left panel of Table 2, which shows participating players' average probability belief about the total number of players willing to participate in the organization. If players' high participation rate was mainly due to miscoordination, players should expect with high probability that s^* players will participate in the organization. This is not what we find. In treatment IF40 and IF65, participants hold on average a belief of only

about 22 and 12 percent, respectively, that s^* players will participate in the organization. The belief is slightly higher in the first round of both treatments, but it decreases to 16 percent in the final round of treatment IF40 and even to 6 percent in the final round of treatment IF65. As can be seen, in both treatments, participants' average belief peaks at "four participants". The average belief that all players will participate amounts to 65 and 61 percent over all rounds in treatment IF40 and IF65, respectively. In fact, the belief is already quite high in the first round and increases to over 74 and 70 percent, respectively, in the final round of the two treatments. The increase clearly mirrors the corresponding increase in the implementation of the grand organization. Overall, beliefs demonstrate that from early on players who are willing to participate rarely expect organizations of size s^* to be formed, but mostly expect that all of the players will participate in the organization. Thus, it seems unlikely that high participation rates are driven by miscoordination.

Table 2: Beliefs and rate of implementation

	Treatment									
	IF40					IF65				
		# of participants					# of participants			
Belief	# obs.	1	2	3	4	# obs.	1	2	3	4
First round	26	9.42	18.19	34.50	37.88	25	19.52	14.48	23.80	42.20
Final round	35	5.29	4.83	15.86	74.03	32	2.34	5.47	21.28	70.91
All rounds	726	6.48	7.03	21.67	64.81	671	5.01	11.80	21.75	61.44

	# of participants				# of participants			
Implementation rate	1	2	3	4	1	2	3	4
All rounds	0.00	2.94	23.08	69.30	27.78	37.50	37.29	90.91
# obs.	7	34	65	114	18	40	59	99

Note: The upper panel of the table presents the average probability belief (in percent) of participating players in stage one of the game about the total number of participants in the organization. The lower panel presents the likelihood of implementation (in percent) of an organization depending on the number of participating players.

Further evidence is presented in the lower panel of Table 2, which shows the average likelihood over all rounds with which an organization is implemented depending on the number of

participating players. Note that there are many cases in both treatments in which from one to three players have to decide whether to implement an organization. As the data show, most of these organizations are not implemented. In fact, if less than s^* players participate, the likelihood of implementation lies below 3 percent in treatment IF40 and below 28 percent in treatment IF65. This finding corresponds to Result 2, which states that only few organizations smaller than s^* are observed in the experiment. When the number of participants hits the threshold s^*, the likelihood of implementation rises somewhat, but still remaines at a rather low level of 23 and 38 percent in treatments IF40 and IF65, respectively. Only if all players participate, do organizations have a high chance to be implemented. In this case, the likelihood of implementation rises to almost 70 percent in treatment IF40 and to over 90 percent in treatment IF65.[21] One may also hypothesize that the predominant implementation of grand organizations is a consequence of the coordination problem and focalness. Indeed, the result that most participating players expect all other players to participate as well is consistent with such a hypothesis. However, since organizations of size $s < n$ (but not smaller than s^*) are frequently rejected by players, the data suggest that the formation of grand organizations is not solely due to focalness but that subjects' behavior is driven by additional forces such as equality and fairness.

Table 2 shows that implementation rates are, *ceteris paribus*, higher in treatment IF65 than in IF40. Interestingly, this is exactly what social preferences would predict (cf. the discussion following Proposition 3). While according to standard theory any organization larger than s^* should be implemented with probability one, fairness predicts that players reject unequal organizations but that these rejections are more costly (and hence less likely to occur) the more

[21] The fact that 30 percent of the grand organizations are not implemented in treatment IF40 may seem surprising, but is only due to the different learning speeds in the two treatments. In IF40, the likelihood of implementation greatly increases over rounds, reaching levels similar to those in treatment IF65 in the second half of the experiment. In rounds 11 to 20 of treatment IF40, the likelihood of implementing a grand organization is 86 percent (63 observations) and even increases to 94 percent (32 observations) in the final five rounds of the experiment. As for treatment IF65, the likelihood of implementing a grand organization is 93 percent (59 observations) in the final ten rounds and 90 percent (30 observations) in the final five rounds.

productive the public good. Result 4 summarizes our findings.

Result 4 *Organizations with less than four participants have a high likelihood of being rejected in the implementation stage of both IF treatments. Only the grand organization has a substantial likelihood of being implemented. Ceteris paribus, implementation rates are higher the more productive the public good.*

Once an organization is implemented, its members are bound to contribute their full endowment to the public good. Thus, if all four players participate and the organization is implemented, contribution levels reach 100 percent. Yet, as shown above, there is always the possibility that implementation fails and players end up in the status quo. It is conceivable that the failure to implement an organization might have a negative effect on voluntary contributions to the public good. However, as our final result shows, the overall impact of institution formation on average contributions and on efficiency is clearly positive.

Result 5 *Overall, the possibility of institution formation has a positive effect on contributions to the public good. Contributions are both higher and more stable if players are allowed to form organizations than if they are not. In consequence, achieved efficiency levels are higher in the institution formation treatments than in the control treatments.*

Support for Result 5 is presented in Figure 2, which compares the average contributions to the public good in the institution formation treatments with those in the corresponding control treatments. Consider first treatments IF40 and PG40 (left panel). The data clearly show that in treatment PG40 the average contribution steadily declines from 12.4 in the first round to 0.7 in the last round. In stark contrast, in treatment IF40, the average contribution increases from 9.8 in round one to a maximum of 15.3 in round 18, at which point it falls to an average of 11 in the final round. Over all rounds, players contribute an average of 10.6 in treatment IF40 and an average of 5.0 in treatment PG40. Thus, the possibility of institution formation more than doubles players' average contribution to the public good. This difference is highly significant (Mann-Whitney test, $p < 0.01$).

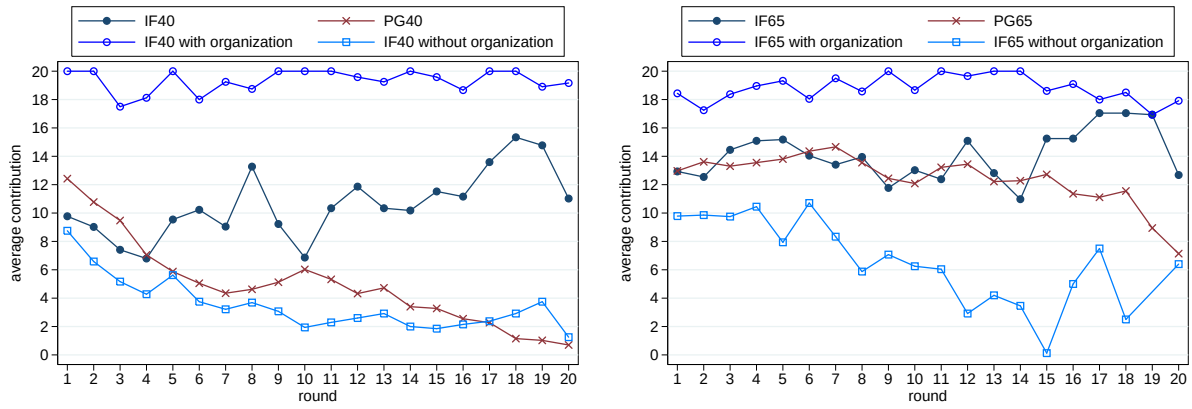

Figure 2: Average contribution to the public good with and without the possibility of institution formation (left panel: IF40, PG40; right panel: IF65, PG65).

In the institution formation treatment IF65, a similar pattern emerges, yet on a slightly higher level. When the MPCR equals 0.65, the average contribution is 12.9 in the first round, increases to a maximum of 17 in rounds 17 and 18, and falls to 12.7 in the final round. In line with previous studies showing a positive effect of an increase in the MPCR on contributions to the public good (see, e.g., Ledyard, 1995), the average contribution is also relatively high in the control treatment PG65. From this it follows that, up to round 14, the average contributions in treatments IF65 and PG65 do not differ statistically from each other. In the final rounds, however, the average contribution falls to 7.1 in treatment PG65, but rises and stays above 12.7 in treatment IF65. Over all rounds, the average contribution in treatment PG65 is 12.4, as compared to 14.1 in treatment IF65. Thus, when players have the possibility to form organizations, their contributions are again higher than when they do not have this possibility, but the difference is not statistically significant (Mann-Whitney test, $p = 0.48$). If we consider rounds 15 to 20, the difference becomes marginally significant (Mann-Whitney test, $p = 0.10$).

Figure 2 also shows the average contribution to the public good in treatments IF40 and IF65, both when an organization has been implemented and when no organization has been implemented. Contribution levels are close to 100 percent when an organization has been implemented since organization members are bound to contribute their full endowment and most

organizations that are implemented comprise all players. What is particularly interesting is the comparison of the average contribution in the IF treatment when no organization has been established with the average contribution in the corresponding PG treatment. In both cases, all subjects are able to freely decide how much they contribute to the public good. However, in the first case, this is because subjects have rejected the implementation of an organization, while in the second case subjects do not have the possibility of institution formation. By comparing the resulting contribution levels, we can thus determine whether the failure to implement an organization has a negative effect on players' voluntary contribution to the public good. As Figure 2 shows, implementation failure has basically no effect in treatment IF40. If players can form institutions but the implementation fails, the average contribution is 4.2 compared to 5.0 in treatment PG40 (Mann-Whitney test, $p = 0.23$). In treatments IF65 and PG65 the difference is larger and marginally significant. Subjects contribute on average 8.5 in treatment IF65 if no organization exists compared to 12.4 in treatment PG where no institution formation is possible (Mann-Whitney test, $p = 0.07$). Thus, institution formation failure has a negative effect on voluntary contributions in this treatment. Importantly, however, as we saw above the overall effect of institution formation on contribution levels is always positive.

Finally, to evaluate efficiency we calculate the average observed level of efficiency relative to the welfare maximum in all treatments. That is, efficiency is defined as $(\Pi_{observed} - \Pi_{min})/(\Pi_{max} - \Pi_{min})$, where $\Pi_{observed}$ denotes the observed group earnings, Π_{min} the theoretical minimum group earnings, and Π_{max} the theoretical maximum group earnings. Our data show that efficiency levels are higher in the IF treatments than in the corresponding PG treatments. In treatment IF40, the average efficiency over all rounds lies at 51 percent compared to 25 percent in PG40. This difference is significant (Mann-Whitney test, $p = 0.01$). Average efficiency in treatment IF65 is 70 percent compared to 62 percent in PG65. Due to the relatively high contributions in the PG treatment this difference fails to be significant (Mann-Whitney test, $p = 0.47$). Importantly, while efficiency decreases significantly over rounds in both PG treatments (Spearman rank order correlation; PG40: $\rho = -0.93, p = 0.00$; PG65: $\rho = -0.80, p = 0.00$ in PG65), this is

not the case in the IF treatments. Here, the Spearman rank order correlation between rounds and the achieved level of efficiency is positive (IF40: $\rho = 0.73, p = 0.00$; IF65: $\rho = 0.30, p = 0.19$).

III. Conclusion

We analyze the endogenous formation of institutions in a linear public goods game. Players in our model have the possibility to establish a sanctioning organization that punishes players who do not contribute the efficient amount to the public good. While the voluntary formation of such institutions typically suffers from a second-order free-rider problem — jointly, every player profits if an institution is formed, but each player profits even more if only the other players implement an institution — we show in the theory part of this paper that players can, in principle, overcome this problem. If institution formation is efficient, a non-empty set of subgame perfect Nash equilibria exists, in which s players $(s^* \leq s \leq n)$ implement a sanctioning organization. However, there also exist subgame perfect Nash equilibria, in which no institution is formed. Strictness as a refinement selects the minimally profitable organization of size s^* — a result which mirrors a similar prediction based on the notion of internal stability in the environmental economics literature (Barrett, 1994). Although this equilibrium prediction is intuitive in terms of individual material payoff maximization, it seems questionable if at least some players have a concern for equity and efficiency. Our analysis based on Fehr and Schmidt (1999) preferences confirms that if some players are inequity averse, the s^* organization may indeed no longer be an equilibrium and the grand organization becomes the expected equilibrium outcome.

In the empirical part we report the results of an experiment designed to investigate players' behavior in institution formation and its effect on public good provision. The obtained results show that organizations are formed and that the possibility of institution formation has a clear positive impact on contributions to the public good. Contributions are both higher and more stable if players are able to form institutions compared to if they are not able to do so. The results also support the theoretical prediction based on social preferences that mainly grand organizations will be formed. The likelihood of implementation of a sanctioning organization crucially

depends on the number of participating players. In particular, only the grand organization has a reasonable chance of being implemented. Smaller organizations, even if more than s^* players participate, are rejected most of the time.

The main message of our paper is twofold. First, our theoretical and empirical analysis shows that institution formation can be an important and effective solution in social dilemma situations. Despite the second-order free-rider problem the institution-formation process may be structured such that the implementation of a sanctioning organization is supported by a Nash equilibrium. The experimental data corroborate this result as players make frequent use of the possibility of institution formation and thereby substantially increase efficiency. Second and most importantly such success is not guaranteed, however. In particular, our results highlight the crucial role of fairness in the institution formation process. Individuals are very reluctant to comply with a sanctioning system that governs only a subset of individuals. This is true even if the subset of individuals can earn a higher material payoff than in the status quo with no sanctioning system and the implementation of the system, at least from a material standpoint, would thus be optimal. The importance of fairness for the governance of public goods has been documented in the field, e.g., in studies of real-world common property regimes (Baland and Platteau, 1996; McKean, 2000). As one of the researchers in this literature notes, the "[d]istribution of decision-making rights and use of rights to coowners of the commons need not be egalitarian but must be viewed as 'fair'. (...) If any subgroup feels cheated — denied 'adequate' access of 'fair' share — compared to another subgroup, the angry subgroup becomes unwilling to participate in decision making, unwilling to invest in maintaining or protecting the commons, and motivated to vandalize the commons" (McKean, 2000, p. 47-8). Notably, fairness arguments were also put forward in the political discussion about the Kyoto Protocol, in particular after the United States' withdrawal from the protocol (cf. Footnote 6 above). Buchanan and Congleton (1998) take such arguments — plus further efficiency considerations — to propose the so-called *generality principle* as a normative guideline for political action. The principle asserts that political choices should be general in nature, i.e., non-discriminatory and based on equal treatment of all

individuals. Our paper offers strong support for this principle both theoretically and empirically. Theoretically, because our formal results show that only those institutional rules that obey the generality principle are selected as equilibrium when players have social preferences with a taste for fairness. Empirically, because general rules are also the predominant outcome of institution formation in the experiment.

For economists interested in institutions, the role of fairness can be seen as good and bad news. Bad news, because fairness motives act as a constraint on equilibria which increases the risk of failure of the institution formation process (in particular, if these motives are not taken into account ex ante). Good news, because once the process is successful, efficiency levels will typically be higher compared to a world in which fairness motives do not play a role.

REFERENCES

Agrawal, Arun, and Gautam N. Yadama. 1997. "How do Local Institutions Mediate Market and Population Pressures on Resources? Forest Panchayats in Kumaon, India." *Development and Change,* 28(3): 435-465.

Anderson, Christopher M., and Louis Putterman. 2006. "Do Non-Strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism." *Games and Economic Behavior,* 54(1): 1-24.

Bagnoli, Mark, and Barton L. Lipman. 1989. "Provision of Public Goods: Fully Implementing the Core through Private Contributions." *Review of Economic Studies,* 56(188): 583-601.

Baland, Jean-Marie, and Jean-Philippe Platteau. 1996. *Halting Degragation of Natural Resources.* Oxford: Oxford University Press.

Barrett, Scott. 1994. "Self-Enforcing International Environmental Agreements." *Oxford Economic Papers - New Series,* 46(Suppl. S): 878-894.

Bolton, Gary, and Axel Ockenfels. 2000. "ERC - A Theory of Equity, Reciprocity, and Competition." *American Economic Review,* 90(1): 166-193.

Buchanan, James M., and Roger D. Congleton. 1998. *Politics by Principle, Not Interest.* Cambridge: Cambridge University Press.

Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction.* New York and Princeton: Princeton University Press.

Carpenter, Jeffrey P. 2007a. "The Demand for Punishment." *Journal of Economic Behavior and Organization,* 62(4): 522-542.

Carpenter, Jeffrey P. 2007b. "Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods." *Games and Economic Behavior,* 60(1): 31-51.

Carraro, Carlo, and Domenico Siniscalco 1993. "Strategies for the International Protection of the Environment." *Journal of Public Economics,* 52(3): 309-328.

Charness, Gary, and Matthew Rabin 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics,* 117(3): 817-869.

Chen, Yan, and Charles R. Plott 1996. "The Groves-Ledyard Mechanism: An Experimental Study of Institutional Design." *Journal of Public Economics,* 59(3): 335-364.

Cox, James C., Daniel Friedman, and Steven Gjerstad. 2007. "A Tractable Model of Reciprocity and Fairness." *Games and Economic Behavior,* 59(1): 17-45.

Croson, Rachel T. A., and Melanie Beth Marks. 2000. "Step Returns in Threshold Public Goods: A Meta- and Experimental Analysis." *Experimental Economics,* 2(3): 239-259.

d'Aspremont, Claude, Alexis Jacquemin, Jean Jaskold Gabszewicz, and John A. Weymark. 1983. "On the Stability of Collusive Price Leadership." *The Canadian Journal of Economics / Revue canadienne d'Economique,* 16(1): 17-25.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior,* 47(2): 268-298.

Falk, Armin, and Urs Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior,* 54(2): 293-315.

Falkinger, Josef. 1996. "Efficient Private Provision of Public Goods by Rewarding Deviations from Average." *Journal of Public Economics,* 62(3): 413-422.

Falkinger, Josef, Ernst Fehr, Simon Gächter, and Rudolf Winter-Ebmer. 2000. "A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence." *American Economic Review,* 90(1): 247-264.

Fehr, Ernst, and Simon Gächter. 2000a. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review,* 90(4): 980-994.

Fehr, Ernst, and Simon Gächter. 2000b. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives,* 14(3): 159-181.

Fehr, Ernst, and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature,* 415: 980-994.

Fehr, Ernst, and Klaus Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics,* 114(3): 817-868.

Groves, Theodore, and John Ledyard. 1977. "Optimal Allocation of Public Goods: A Solu-

tion to the 'Free Rider' Problem." *Econometrica,* 45(4): 783-810.

Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach. 2006. "The Competitive Advantage of Sanctioning Institutions." *Science,* 312: 108-111.

Hviding, Edvard, and Graham B. K. Baines. 1994. "Community-Based Fisheries Management, Tradition, and the Challenges of Development in Marovo, Solomon Islands." *Development and Change,* 25(1): 13-39.

Kroll, Stephan, Todd L. Cherry, and Jason F. Shogren. 2007. "Voting, Punishment, and Public Goods." *Economic Inquiry,* 45(3): 557-570.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *The Handbook of Experimental Economics,* ed. John H. Kagel and Alvin E. Roth, 111-194. Princeton, NJ: Princeton University Press.

Masclet, David, Charles Noussair, Steve Tucker, and Marie-Claire Villeval. 2003. "Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review,* 93(1): 366-380.

McKean, Margaret A. 2000. "Common Property: What Is It, What Is It Good For, and What Makes It Work?" In *People and Forests: Communities, Institutions, and Governance,* ed. Clark C. Gibson, Margaret A. McKean, and Elinor Ostrom, 27-55. Cambridge, MA: MIT Press.

Nyarko, Yaw, and Andrew Schotter. 2002. "An Experimental Study of Belief Learning Using Elicited Beliefs." *Econometrica,* 70(3): 971-1005.

Offerman, Theo. 1997. *Beliefs and Decision Rules in Public Good Games.* Dordrecht/Boston/London: Kluwer.

Okada, Akira. 1993. "The Possibility of Cooperation in an N-person Prisoners Dilemma with Institutional Arrangements." *Public Choice,* 77(3): 629-656.

Oliver, Pamela. 1980. "Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations." *American Journal of Sociology,* 85(6): 1356-1375.

Ostrom, Elinor. 1999. "Coping with the Tragedy of the Commons." *Annual Review of Political Science,* 2: 493-535.

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review,* 86(2): 404-417.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review,* 83(5): 1281-1302.

Sutter, Matthias, Stefan Haigner, and Martin G. Kocher. 2006. "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations." Center for Economic Policy Research Discussion Paper 5497.

Tang, Shui Yan. 1992. *Institutions and Collective Action.* San Francisco: ICS Press.

Tyran, Jean-Robert, and Lars P. Feld. 2006. "Achieving Compliance when Legal Sanctions are Non-Deterrent." *Scandinavian Journal of Economics,* 108(1): 135-156.

Walker, James M., Roy Gardner, Andrew Herr, and Elinor Ostrom. 2000. "Collective Choice in the Commons: Experimental Results on Proposed Allocation Rules and Votes." *The Economic Journal,* 110(460), 212-234.

WWR. 2006. *Klimaatstrategie – Tussen Ambitie en Realisme.* Amsterdam: Amsterdam University Press.

Yamagishi, Toshio. 1986. "The Provision of a Sanctioning System as a Public Good." *Journal of Personality and Social Psychology,* 51(1): 110-116.

Yamagishi, Toshio. 1988. "The Provision of a Sanctioning System in the United States and Japan." *Social Psychology Quarterly,* 51(3): 265-271.