Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Working Paper No. 299

Institution Formation in Public Goods Games

Michael Kosfeld, Akira Okada and Arno Riedl

August 2006

Institution Formation in Public Goods Games*

Michael Kosfeld
University of Zurich & IZA

Akira Okada
Hitotsubashi University

Arno Riedl
Maastricht University, CESifo & IZA

August 2006

Abstract

Centralized sanctioning institutions are of utmost importance for overcoming free-riding tendencies and enforcing outcomes that maximize group welfare in social dilemma situations. However, little is known about how such institutions come into existence. In this paper we investigate, both theoretically and experimentally, the endogenous formation of institutions in a public goods game. Our theoretical analysis shows that players may form sanctioning institutions in equilibrium, including those where institutions govern only a subset of players. The experiment confirms that institutions are formed frequently as well as that institution formation has a positive impact on cooperation rates and group welfare. However, the data clearly reveal that players are unwilling to implement institutions in which some players have the opportunity to free ride. In sum, our results show that individuals are willing and able to create sanctioning institutions, but that the institution formation process is guided by behavioral principles not taken into account by standard theory.

JEL classification: C72, C92, D72
Keywords: public goods, institutions, sanctions, cooperation

*Michael 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). Akira Okada gratefully acknowledges financial support from the Japan Society for the Promotion of Science under grant No.(A)16203011 and the Matsushita International Foundation. Corresponding author: Arno Riedl, Department of Economics, Maastricht University, P.O.Box 616, NL-6200 MD Masstricht, e-mail: A.Riedl@algec.unimaas.nl

1 Introduction

When markets fail, the design of appropriate institutions is a key issue for economic analysis and policy. In social dilemmas the maximization of social welfare typically conflicts with individual payoff maximization. In such situations, the implementation of a sanctioning institution that castigates individual behavior if it deviates from the welfare maximizing action is a widely used solution. A classic example of such an institution is the constitutional state, which attains cooperation from its citizens through enforcement by central authorities (police, courts). Others include trade unions and employers' associations, which often have an arbitration board monitoring and enforcing the compliance of their members. Examples of institutional arrangements using sanctioning mechanisms can also be found in the international arena. 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, the institutional arrangements have two important elements in common. Firstly, they 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 arrangement. Secondly, sanctioning applies only to the 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. It is exactly this dilemma of institution formation that we address in this paper.[1]

The social dilemma situation we consider is a linear n-player public goods game. The institution we analyze is a centralized sanctioning institution, in which sanctions are imposed by a central authority, for example, a policeman, a court, or an arbitration board.[2] We model the 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

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

[2] We thereby abstract from possible enforcement problems that might arise if the players themselves had 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 in fact rather general since most of our results can be extended to other (non-centralized) institutional arrangements (see below).

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, whereby the minimum size depends on the payoffs in the public goods game and the cost of the organization. Furthermore, using strictness in each 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^* is strictly smaller than n, the unique strict subgame perfect equilibrium organization has a proper subset of players who voluntarily commit themselves to cooperation. Although each individual member of the organization 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 is the case in the status-quo equilibrium, in which no organization is implemented and no players contribute to the public good.

Whether or not the unique strict subgame perfect equilibrium prediction is valid is, of course, an empirical question. Since the game has multiple other (non-strict) subgame perfect Nash equilibria, theory alone can give only limited guidance with regard to the expected outcome of institution formation. In the second part of the paper, we therefore present the results of a laboratory experiment designed to investigate the process of institution formation described above in a 4-player public goods game. The 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. Based on the theoretical analysis, we implemented two experimental treatments, in which the strict subgame perfect equilibrium size of the organization s^* was varied. In the first experimental treatment (IF40), the equilibrium size s^*=3; in the second (IF65), s^*=2. We also conducted two corresponding control treatments (PG40, PG65), in which no institution could be formed and subjects only played the public goods game. In each treatment, subjects played 20 rounds of the relevant stage game; the composition of groups remained constant over rounds.

Our main experimental findings are as follows. Subjects successfully establish organizations and the number of organizations increases over time. In both experimental treatments, from 70 to 100 percent of all groups implement an organization by the final rounds. However, contrary to the strict subgame perfect equilibrium prediction, most (on

average, around 75 percent) of the organizations implemented are of maximum size, i.e., *all* players participate. The likelihood with which an organization is implemented greatly increase as the number of players willing to participate increases. If all players are willing to participate, the average implementation rate amounts to 69 and 91 percent in treatment IF40 and IF65, respectively. In contrast, when one or two players say they are not willing to participate, the likelihood with which an organization is implemented in the two experimental treatments falls to below 23 and 38 percent, respectively. A comparison of the experimental with the control treatments confirms that the opportunity to form institutions results in higher and more stable total contributions to the public good. Overall, institution formation enhances group welfare, despite the fact that it is costly.

Our results have important implications for public policy. First, since sanctioning institutions are 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. The result may 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.[3] The consequence of the observed behavior is 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 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 unique strict equilibrium prediction of 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, since institutions are rejected that from an individual as well as a social 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 are focusing on a particular institutional solution in this paper, our analysis can be extended to other institutional arrangements, including alternative centralized policy instruments, such as the mechanisms proposed by Groves and Ledyard (1977) and Falkinger (1996), but also non-centralized solutions, such as repeated-game trigger strate-

[3]To give two examples — in its latest report on climate strategies, the Dutch Scientific Council for Government Policy 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 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).

gies. 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 the 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 individual incentive to free ride. This second-order free-rider problem (or, as we call it, the dilemma of endogenous institution formation) applies to any mechanism that solves the first-order free-rider problem in social dilemma situations. The contribution of our paper is to 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 2 provides an overview of the related literature. Section 3 theoretically analyzes the institution formation game, characterizing subgame perfect Nash equilibria in the one-shot and the finitely repeated setting. Section 4 describes and analyzes the experiment. Finally, Section 5 concludes.

2 Related Literature

Our study is related to two strands of, mainly experimental, research. The first strand examines *exogenously imposed* institutions whereas the second encompasses studies concerning *endogenously formed* institutions.

Chen and Plott (1996) investigate the classic Groves-Ledyard mechanism, which theoretically can solve the free-rider problem in certain environments. The authors find that the mechanism only results in higher provision of the public good and higher efficiency if punishment is sufficiently high. Falkinger et al. (2000) examine a simple mechanism for public good provision developed by Falkinger (1996). This mechanism rewards and sanctions players who contribute more and less, respectively, than the average to the public good. The authors find that the Falkinger mechanism works very well in the experimental laboratory and implements an efficient contribution level most of the time.[4] Andreoni (1993) and Chan et al. (2002) investigate the effect of tax-financed government subsidies on private contributions. They are mainly interested in the so-called "crowding-out" hypothesis which states that, under some constellations, government contribution will crowd out private contributions completely.

[4] A recent paper by Falkinger (2004) analyzes private incentives to invest in an enforcement technology based on the Falkinger mechanism.

In the above-mentioned literature, exogenously imposed centralized mechanisms are investigated. Recently, the decentralized enforcement of public good provision has been examined in a series of studies. Such enforcement depends on the individual willingness to sanction (or reward) others' contribution behavior. As in the above papers, the institutions in these studies are exogenously imposed by the experimenter. The first papers to investigate such enforcement are Ostrom et al. (1992) and Fehr and Gächter (2000a, 2002). In these studies, players not only contribute to a public good, but also have the opportunity to costly punish other players, after having seen the amount of their contributions. The main result is that such decentralized enforcement can be very effective and often increases private contributions considerably. Subsequent studies (Anderson and Putterman, 2005; Carpenter, 2004, 2006; Masclet et al., 2003; Sefton et al., 2002) largely confirm this finding under different environmental conditions. Some papers, however, also point out the limits of enforcement by individual punishment (Carpenter, 2004, 2006; Egas and Riedl, 2005; Nikiforakis and Normann, 2005). Yamagishi (1986, 1988) investigates a sanctioning institution that lies somewhere between centralized and decentralized enforcement. In his experiments, subjects are able to make individual contributions to a common sanctioning mechanism that punishes free riders in a centralized way.

Studies more similar to ours are those that allow, at least to some extent, for an endogenous choice of a sanctioning or rewarding mechanism. A recent study by Gürerk et al. (2006) investigates "voting by feet" — giving subjects in a public goods experiment the option of playing the game under a regime either with or without individual sanctioning opportunities. They find that over time most subjects choose the sanctioning regime and contribution levels increase to full contribution. Walker et al. (2000) analyze a multi-level common-pool resource game in which participants are able to propose and vote on different allocation rules. Their major finding is that the use of voting substantially increases efficiency relative to a baseline with no opportunity for collective choice. Sutter and Weck-Hannemann (2004) let people vote on taxes that are to be spent on public good provision. Tyran and Feld (2006) let participants vote on the implementation of laws that impose punishment on free riders. Their results show that exogenously implemented mild laws do not result in compliance, whereas compliance improves substantially if the laws are chosen endogenously. Finally, Sutter et al. (2006) analyze an experimental public goods game in which group members are able to endogenously determine whether they want to supplement a standard voluntary contribution mechanism with the opportunity to reward or punish other group members. They find that, as compared to exogenously implemented institutions, endogenous institutional choice has a large and positive effect on the level of cooperation.

The key difference between the above-mentioned studies and ours is that, in the other studies, individual players do not have the opportunity to free ride on the participation

of others in the institution formation process. Free riders are either excluded from the production of the public good, or the mechanism, once implemented, applies to all players in the original social dilemma game. This, however, eliminates the second-order free-rider problem from the start, thus making the results less applicable to a number of institution formation problems, including those mentioned at the beginning of this paper. In the case of the Kyoto Protocol, for example, it would either mean that non-participating countries would somehow be excluded from the benefits of reduced global greenhouse gas emissions or that they would be forced to join the agreement. Both seem unrealistic. In contrast, our approach explicitly takes the second-order free-rider problem into account, so that we can analyze how individuals overcome this problem in an institution formation process.

3 Institution Formation: Theory

3.1 Model

Consider the following n-player public goods game. Every player $i = 1, \ldots, n$ has an endowment w and can determine the amount of his contribution g_i to a public good where $0 \leq g_i \leq w$. For a given contribution profile $(g_1, \ldots, g_n)$, the payoff to player i is given by

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

The parameter a determines the marginal per capita return from a contribution to the public good. The assumption $a < 1$ means that zero contribution is the dominant action for every player, that is, he is better off by choosing zero contribution than by choosing any positive contribution regardless of the contributions of the other players. In consequence, the strategy profile $g = (0, \ldots, 0)$ is the unique Nash equilibrium of the public goods game. The assumption $na > 1$ means that all players are better off if each player makes a full contribution to the public good. In particular, $g = (w, \ldots, w)$ is the welfare maximizing strategy profile in the game.

The institution formation process we consider in this paper is an extension of the idea laid out in Okada (1993) and takes the form of the following three-stage game:

1. *Participation stage:* Players simultaneously and independently decide whether or not they are willing to participate in an organization that punishes all organization members who do not contribute 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:* Participants negotiate about whether or not to actually implement an organization. If they agree to implement an organization, all participants

become members of that organization. Non-participants cannot become members. All players know that, once an organization is implemented, any organization member who does not contribute the full endowment to the public good in the contribution stage (see below) will be punished. Non-members are never punished, i.e., they can free ride on members' contributions. Specifically, in this stage, participants simultaneously and independently either accept or reject the implementation of an organization. An organization is implemented if and only if all participants accept (unanimity rule). The organization is costly; once the organization is implemented, costs are shared equally by all members.

3. *Contribution stage:* All players simultaneously and independently determine the amount of their own contribution to the public good. If an organization has been implemented, members who do not contribute fully are subject to punishment. Non-members are not punished. If no organization was implemented, no punishment is executed.

Let Γ denote the institution formation game as described above. Formally, Γ is a 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. A player's payoff u_i is defined 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. Then, for $i = 1, \ldots, n$,

(i) if an organization is implemented:

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

where $p(g_i)$ is the punishment imposed on member i satisfying[5]

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

(ii) if no organization is implemented:

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

In the following, we first characterize the set of subgame perfect equilibria of the institution formation game Γ if the game is played one shot. We then analyze equilibria of the finitely repeated game.

[5]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.

3.2 One-shot game

In a subgame perfect equilibrium, players decide on their actions in every stage of Γ, rationally anticipating the outcome of future stages. We analyze the Nash equilibria of the final contribution stage first.

Lemma 1 *The contribution stage has a unique Nash equilibrium. If an organization is implemented, all members contribute fully and all non-members contribute nothing. If no organization is established, no players contribute anything.*

The proof of Lemma 1 follows from the rules of the contribution stage at which punishment of members is high enough to deter free riding. We next analyze the implementation stage. Suppose that s players participate. Although each individual player has an incentive to free ride in the public goods game, it is important to note that participants are better off if they coordinate their contributions in the framework of an organization if

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

Definition 1 *The threshold s^* of an organization is the minimum non-negative integer s satisfying condition (5).*

The threshold s^* gives the minimum size of an organization such that participants have an incentive to actually establish it. Note that $s^* \geq 2$ because $a < 1$ and $c \geq 0$. Since the left side of condition (5) is strictly increasing in s, $s^* \leq n$ exists uniquely if

$$(an - 1)nw > c. \tag{6}$$

If (6) does not hold, an organization is never beneficial to the players, i.e., no group of players would ever have an incentive to implement it. In the following, we therefore assume that (6) holds. Our next result characterizes the set of Nash equilibria of the implementation stage.

Lemma 2 *If the number of participants is greater than or equal to the threshold s^*, there exist two types of Nash equilibria in the implementation stage: one with and one without an organization. If $s < s^*$, the organization is not implemented in all Nash equilibria.*

Proof. Note first that non-participants do not make any decision in the implementation stage. Let s be the number of participants. If $s \geq s^*$, it can be shown that both the action profile where all participants agree to implement the organization and the action profile where they all reject the organization are Nash equilibria. Consider first that all participants agree. Equation (2) and Lemma 1 imply that each participant earns $asw - c/s$. A unilateral deviation, i.e., one player rejecting the implementation, results in

the deviating player (as well as all other players) earning w (equation (4) and Lemma 1). The definition of s^* guarantees that such a deviation is strictly worse than playing the equilibrium strategy. Now consider that all participants reject the implementation. The unanimity rule then guarantees that unilateral acceptance by a participant is not profitable because it does not alter the outcome. If $s < s^*$, the action profile where all participants agree to implement the organization is not a Nash equilibrium since, by definition of s^*, each participant is better off if he rejects the implementation.[6] Q.E.D.

Intuitively, the implementation stage is an unanimous voting game in which all participants collectively decide between two alternatives: the status quo or the organization. If the number of participants is greater than or equal to the threshold s^*, the organization Pareto-dominates the status quo. While both the organization and the status quo can be supported by a Nash equilibrium, the equilibrium with an organization can be selected by applying refinements of the Nash equilibrium concept such as strictness and undominatedness. We will discuss this issue further at the end of this section.

We are now ready to characterize a subgame perfect equilibrium of Γ. For convenience, we introduce the following classification of subgame perfect equilibria.

Definition 2 *A subgame perfect equilibrium of* Γ *is called an "organizational equilibrium" if an organization is formed on the equilibrium path. A subgame perfect equilibrium of* Γ *is called a "status-quo equilibrium" if an organization is never formed independent of the number of participants.*

In an organizational equilibrium, an organization is formed and all members contribute their full endowment. By Lemma 2 the size of the organization is greater than or equal to the threshold s^*. In contrast, in the status-quo equilibrium, no organization is implemented regardless of how many players participate in the first stage. The status quo of the public goods game, when no player contributes, prevails.

Proposition 1 *A subgame perfect equilibrium of the institution formation game* Γ *is characterized as follows.*

(1) There exists an organizational equilibrium if and only if the number s *of participants is greater than or equal to the threshold* s^**.*

(2) For any number of participants $s = 1, \ldots, n$ *there exists a status-quo equilibrium.*

[6] Note that if the equality $a(s^*-1)w - \frac{c}{s^*-1} = w$ happens to hold, every participant is, in fact, indifferent between accepting and rejecting the organization of size $s^* - 1$. In this case, all action profiles in the implementation stage with $s^* - 1$ participants are Nash equilibria resulting in the same payoffs w. We have excluded this degenerate case from the analysis. In the experiment, the inequality $a(s^* - 1)w - \frac{c}{s^*-1} < w$ always holds.

Proof. (1) The "only-if part" follows from Lemma 2, which states that no organizational equilibrium exists if $s < s^*$. To prove the "if part," suppose that $s \geq s^*$. Define players' strategies as follows:

- s players participate in an organization, and the others do not participate,
- If exactly s players participate in an organization, each participant accepts the implementation of the organization. Otherwise, each participant rejects it.
- All members of the organization contribute fully, and all non-members choose zero contribution.

By Lemma 1 and 2 these strategies induce Nash equilibria in the second and the third stage of Γ. Consider now the participation stage. If players follow the above strategies, an organization is established and, therefore, every participant receives a payoff $asw - c/s$, which is larger than w since $s \geq s^*$. By construction of the strategies, participants receive a payoff of w if they deviate, because no organization is implemented in this case. Hence, participating is a best response to the other players' strategies. Non-participants also have no incentive to deviate because by construction the organization is formed and non-participants can free ride on participants. Hence, the above strategies induce a Nash equilibrium in the participation stage.

(2) In the status-quo equilibrium, every organization is rejected in the implementation stage, independent of the number of participants. Consequently, every player is indifferent about whether or not to participate, no matter what actions the other players choose. Therefore, all action profiles are Nash equilibria in the participation stage. Q.E.D.

Proposition 1 shows that players can overcome the second-order free-rider problem. By successfully implementing an organization that punishes free riders, players in an organizational equilibrium bind themselves to contributing to the public good. Yet, as the result demonstrates, there exist multiple equilibria in the institution formation game. In particular, equilibrium organizations may be of any size from $0, s^*, s^* + 1, \ldots, n$. In the proof, we show that organizations of size $s > s^*$ can be supported in equilibrium by a disciplinary strategy, by which participating players reject any organization with strictly less than s participants (even if these participants might all be better off by implementing such an organization). The status-quo outcome is also possible in equilibrium in the following situations: Either no player participates, or some players participate but the organization is not implemented independent of how many players participate. In the latter case, the participation stage is "inessential" in the sense that the course of the game always leads to the same outcome, namely the status quo.[7]

[7]Proposition 1 is 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

While there exist multiple subgame perfect equilibria of Γ, the next result shows that there is a unique equilibrium in terms of the organization size if we apply strictness as a refinement of Nash equilibrium at each stage. 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. A subgame perfect equilibrium of a multi-stage game with perfect information is called *strict* if it induces a strict Nash equilibrium in every stage game.

Proposition 2 *The institution formation game Γ has a unique strict subgame perfect equilibrium in terms of the organization size. The unique organization size is equal to the threshold s^*.*

Proof. It follows from the proof of Lemma 2 that in any strict subgame perfect equilibrium of Γ, an organization is implemented whenever $s \geq s^*$. Given this outcome of the second stage, the payoff of every player i in the participation stage equals

$$u_i = \begin{cases} asw - \frac{c}{s} & \text{if} \quad i \in S \\ w + asw & \text{if} \quad i \notin S, \end{cases} \tag{7}$$

if $s \geq s^*$, and $u_i = w$ if $s < s^*$.

(i) Suppose that, in the first stage, the number of participants s is strictly larger than s^*. Then — because of the strict equilibrium requirement that an organization be implemented in the second stage whenever $s \geq s^*$ — an organization is also implemented if one participant deviates in the participation stage because the condition $s - 1 \geq s^*$ still holds for the remaining participants. Therefore, each participant can increase his payoff from $asw - c/s$ to $w + a(s-1)w$ by not participating. Thus, the situation is not supported by a Nash equilibrium. (ii) If $s = s^*$, no participant has an incentive to deviate to non-participation because by doing so, his payoff strictly decreases from $as^*w - c/s^*$ to w. Any non-participant deviating to participation would also strictly decrease his payoff from $w + as^*w$ to $a(s^* + 1)w - c/(s^* + 1)$. The arguments in (i) and (ii) imply that, whenever $s \geq s^*$ in the participation stage, only an action profile with exactly s^* participants is a strict Nash equilibrium. (iii) If the number of participants is $s = s^* - 1$, then any single non-participant has an incentive to participate since by doing so his payoff will increase from w to $as^*w - c/s^*$. Therefore, the situation is not supported by a Nash equilibrium. Finally, it can easily be seen that any action profile with $s^* - 2$ or less participants is a Nash equilibrium, but not a strict one. Q.E.D.

The refinement result in Proposition 2 yields a clear prediction regarding the size of the organization that is implemented by the n players: exactly the minimum number of

has been implemented or not. The subgame perfect equilibrium concept should then be replaced by a sequential equilibrium. This is because players' payoff functions depend on the number of participants (not their identity) and on whether or not an organization is established. The strategies constructed in the proof of Proposition 1 also only depend on these variables.

players s^* in order for the organization to be individually profitable form the organization, while the remaining players do not participate. Thus, unless the threshold $s^* = n$ (where all players form the organization), players are divided into two proper subsets: those who voluntarily implement the sanctioning institution, hence contributing to the public good, and those who do not participate, hence not contributing. Whether or not this prediction is valid is an empirical question which we will address in the following section. Before we come to the experiment, however, we will analyze the possible equilibria of the finitely repeated institution formation game.

3.3 Finitely repeated game

Suppose that the game Γ is played finitely many times among the same players. The number of repetitions T is common knowledge and players maximize the total payoff in the repeated game. Let Γ^T denote the T^{th} repetition of Γ in which every player has perfect information about the history of play in all previous rounds. Given the multiplicity of equilibria of Γ as shown in Proposition 1, the next result reveals that a variety of patterns may arise in an equilibrium of the finitely repeated game Γ^T.

Proposition 3 *Let s_t denote the size of the organization implemented in round t of Γ^T. Then, any sequence $s = (s_1, \ldots, s_T)$ where $s_t = 0, s^*, s^*+1, \ldots, n$ for all t can be supported by a subgame perfect equilibrium of the finitely repeated game Γ^T.*

Proof. Define the strategies of players in Γ^T as follows. In every round t, play the subgame perfect equilibrium of Γ such that exactly s_t players implement the organization, independent of the history of play. The existence of such an equilibrium is shown by Proposition 1. Together with the assumption that players maximize the total payoff in the repeated game, these strategies induce a subgame perfect equilibrium in Γ^T.[8] Q.E.D.

Proposition 3 is an example of the well-known fact that the repetition of Nash equilibria in a component game is a subgame perfect equilibrium of the finitely repeated game. Given the multiple equilibria of the component game Γ, there exist a plethora of equilibria once players interact repeatedly over several rounds. One possible prediction based on the strict subgame perfect equilibrium of Γ is that exactly s^* players implement an organization in each round, possibly alternating the roles between participating and non-participating players in different rounds. However, other equilibrium outcomes are also conceivable. For example, players may reject all organizations with less than n players in the implementation stage of each round and only form institutions in which all players

[8] Again, Proposition 3 also holds if players only know the number of participants and whether or not an organization is implemented in the current round. In that case, the subgame perfect equilibrium should be replaced by a sequential equilibrium together with any belief assigned to each information set in Γ^T.

participate in the organization. This equilibrium is favorable in terms of both equality and efficiency as it implements 100-percent contribution to the public good and all players share the cost of the organization equally. However, the equilibrium strategy requires participants to discipline non-participants by rejecting less than full-size organizations, which entails playing a weakly dominated action in case $s \geq s^*$ (cf. the argumentation following Lemma 2). It is unclear whether players are willing to do this. Theory alone gives little guidance regarding the expected outcome of institution formation in our set-up. In the following, we therefore present a laboratory experiment designed to investigate the process of institution formation in public goods games.

4 Institution Formation: Experiment

4.1 Procedural details

To simplify, we slightly modified the institution formation game introduced above. 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 in this paper rather than on the separate issue of institutional enforcement. The basic structure of the experimental game was, therefore, as follows.[9]

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*).[10] 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 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 whether or not an organization has been implemented, freely determine the amount of their contributions. If no organization is implemented, all players freely determine the amount of their contributions.

[9] Experimental instructions are available upon request from the corresponding author.

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

If an organization is implemented, organization members earn a payoff

$$20as + a\sum_{j\notin S} g_j - \frac{c}{s}. \tag{8}$$

Recall that S denotes the set and s the number of organization members, c is the cost of the organization, and g_j is player j's contribution to the public good. Non-members i earn a payoff

$$20 - g_i + 20as + a\sum_{j\notin S} g_j. \tag{9}$$

If no organization is implemented, all players i earn the payoff $20 - g_i + a\sum_{j=1}^{4} g_j$. A comparison of equation (8) and (9) reveals once again the key difference between members and non-members of the organization. While the former are bound to contribute their full endowment to the public good and share the costs of the organization, the latter are free to contribute whatever they want and do not pay any portion of the costs of the organization.

Since the decision to participate may constitute a nontrivial coordination problem — in particular, if only a subset of the players wants to form an organization — we elicited players' beliefs in the participation stage. Precisely, after the players decided whether to participate in the organization, each player is asked to indicate his belief about the number of players s that decided to participate in the organization, where s ranges from one to four if the player himself decided to participate and from zero to three if the player himself decided not to participate. Players are rewarded for correct predictions according to the quadratic scoring rule.[11]

As for the institution formation game, we implemented two experimental treatments with different equilibrium predictions regarding the minimum organization size s^*. In both of these treatments, the cost of the organization was set to $c = 2$. In the first treatment (IF40), the marginal per capita return 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 also 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, IF60), 40 subjects participated in treatment PG40, and 36 participated in treatment PG65. No subject participated in more than one treatment. Each session

[11] Quadratic scoring rules are known to be incentive compatible and have successfully been used in a number of experiments, for example, by Offerman (1997), Sonnemans et al. (1998), Huck and Weizsäcker (2002), and Nyarko and Schotter (2002).

lasted, on average, 120 minutes. On average, a subject earned €23.90 (about $25) in the experiment.

4.2 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. Next, we analyze the overall impact of institution formation on the provision of the public good by comparing the average contribution in the institution formation treatments with the average contribution in the corresponding control treatments. Finally, we evaluate the achieved efficiency in all treatments.

Our first result shows that players almost always initiate an organization, and also implement the initiated organization in between 43 and 60 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 IF60, in 98 percent of the cases, at least one player per group initiates an organization and an organization is implemented in 60 percent of the cases.*

Support for Result 1 is presented in Table 1, which summarizes the absolute number and relative proportion 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 (less than 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, $p > .14$).[12]

Table 1: Absolute number and percentage of organizations initiated and implemented

	Treatment			
	IF40		IF65	
	number	percentage	number	percentage
initiated organizations	220	100	216	98
implemented organizations	95	43	132	60

[12]Statistical tests are based on group average as the unit of observation. We report the results of two-sided tests throughout the paper.

Result 1 shows that players overcome the second-order free-rider problem and successfully establish organizations. The key question is what organizations are implemented. The answer is given in Result 2.

Result 2 *In both IF treatments, the large majority of implemented organizations are of full size, i.e., all players become members. Organizations of size $s < s^*$ are very rarely observed.*

Figure 1 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 of full size (83 percent). In treatment IF65, 90 of 132 organizations are of full size (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 six organizations (one in IF40, five in IF65) are implemented that contain less than s^* players (less than 3 percent). Thus, threshold s^* serves as a good prediction of the minimum size of an implemented organization. However, in view of the strict subgame perfect equilibrium prediction (Proposition 2), it clearly fails to predict the maximum size of an organization. Rather than seeing only three or two players bind themselves, we observe that most of the time all four players establish an organization. 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, $p = .20$).

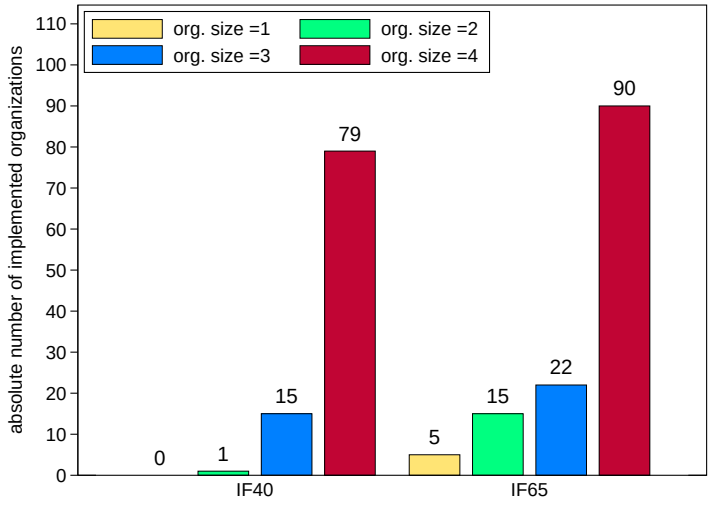

Figure 1: Distribution of implemented organizations in treatments IF40 and IF65.

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 next result shows that this is indeed the case.

Result 3 *The number of implemented organizations increase over rounds in both IF treatments. The overall rise is driven solely by an increasing number of implementations of full-size organizations.*

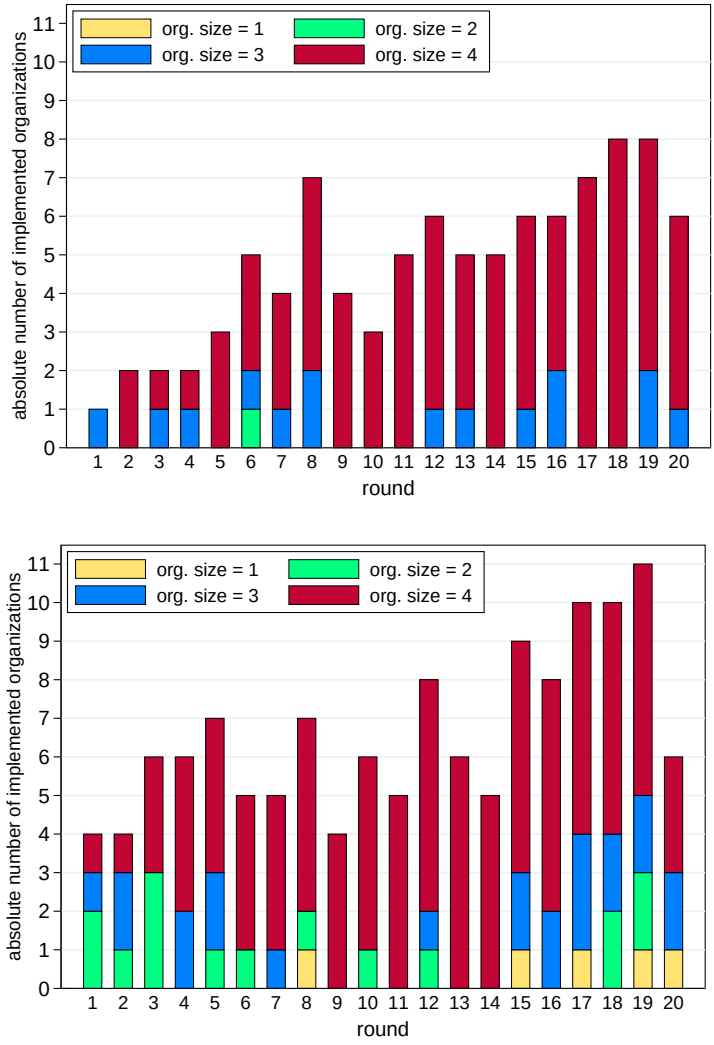

Figure 2: Distribution of implemented organizations over rounds (upper panel: IF40, lower panel: IF65).

Support for Result 3 is presented in Figure 2 which shows the number of implemented organizations in each round of the two IF treatments. This figure illustrates three interesting things. First, the number of implemented organizations increases over rounds in both treatments. In treatment IF40, for example, the number of organizations rises from one and two initially to a maximum of eight in rounds 18 and 19. Even more organizations

are implemented in treatment IF65 (cf. Result 1). While four of eleven groups implement an organization in the first two rounds of treatment IF65, almost all groups implement an organization in rounds 17 to 19.[13] Second, the overall rise in implemented organizations is exclusively driven by the implementation of full-size organizations. Initially, only one or two full-size organizations are implemented in the first few rounds of the two treatments. However, the number strongly increases to eight in treatment IF40 (round 18) and six in treatment IF65 (round 15 to 19).[14] Finally, while there is only a weak and insignificant endgame effect in treatment IF40, we observe a large and significant endgame effect in the final round of treatment IF65. Here, the number of implemented organizations decreases by almost half, decreasing from eleven in round 19 to six in round 20 (Mann-Whitney test, $p = .01$). Again, the effect is driven by a sharp decline in the number of full-size organizations (six organizations in round 19 vs. only three in round 20).

Why do we observe so many organizations larger than s^* in the experiment? Given the theoretical analysis from the previous section, each of the players should in principle have an interest in having the *other* players implement an organization. Is it that the players aim to implement the s^* organization but miscoordinate in the participation stage? Or, do participating players target a full-size organization and reject most 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 believe that the remaining players will participate as well. Secondly, we show that initiated organizations containing 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 and rather suggest that players target a full-size organization from the very beginning.

Result 4 *In both IF treatments, most players who participate in the organization in stage one believe that all other players will participate as well.*

Support for Result 4 is presented in Table 2, which shows a player's average probability belief about the number of other players willing to participate when the player himself decided to participate (left panel) and when the player himself decided not to

[13]The significant increase in organizations is confirmed by a Spearman rank order correlation between the number of implemented organizations and the variable "round" in both treatments (IF40: $\rho = .65, p = .00$; IF65 $\rho = .64, p = .00$).

[14]Spearman rank order correlations between the number of full-size organizations and variable "round" corroborate this finding (IF40: $\rho = .72, p = .00$; IF65: $\rho = .87, p = .00$). At the same time, the number of implemented organizations with less than four players does not change significantly over rounds (Spearman rank order correlation between the number of organizations less than full size and "round", $p > .41$ in both treatments).

participate (right panel) during stage one of the institution formation game. If the players' participation decision was mainly due to miscoordination, players who announce in stage one that they are willing to participate should expect with high probability that $s^* - 1$ players will participate. 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 $s^* - 1$ 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 "three (i.e, all) other participants". The average belief of a participant that three other 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 a full-size organization as documented above. Overall, these beliefs demonstrate that from early on players who are willing to participate rarely expect organizations of size s^* to be formed, but mostly believe that all of the players will participate in the organization. Thus, it seems unlikely that players' participation rates are driven by miscoordination.[15]

Table 2: Players' average probability belief (in percent) about how many of the other players will (also) participate

Treatment IF40										
	participant					non-participant				
		# of other participants					# of other participants			
round	# obs.	0	1	2	3	# obs.	0	1	2	3
first round	26	9.42	18.19	34.50	37.88	18	44.72	23.06	21.11	11.11
final round	35	5.29	4.83	15.86	74.03	9	5.56	3.33	38.89	52.22
all rounds	726	6.48	7.03	21.67	64.81	154	16.15	17.84	35.10	30.91
Treatment IF65										
	participant					non-participant				
		# of other participants					# of other participants			
round	# obs.	0	1	2	3	# obs.	0	1	2	3
first round	25	19.52	14.48	23.80	42.20	19	42.11	23.58	23.58	10.74
final round	32	2.34	5.47	21.28	70.91	12	26.33	22.42	31.58	19.67
all rounds	671	5.01	11.80	21.75	61.44	209	22.56	30.97	28.92	17.55

[15] As can be seen in Table 2, non-participants also do not strongly believe that organizations of size s^* are going to be formed. Over all rounds, in IF40 (IF65) the average belief that three (two) other players will participate is only 31 (29) percent. However, the non-participants' belief pattern differs remarkably from the participants'. Whereas participants learn over time that there is a high probability that all players will participate in the organization, the pattern of non-participants' beliefs remains more or less flat throughout the entire experiment.

Further evidence is presented in Table 3, 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 a few organizations smaller than s^* are observed in the experiment. When the number of participants exceeds 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 are organizations considerably likely 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.[16] Result 5 summarizes this finding.

Table 3: Likelihood of implementation (in percent) by organization size

	# of participating players			
	1	2	3	4
IF40	0.00 (7)	2.94 (34)	23.08 (65)	69.30 (114)
IF65	27.78 (18)	37.50 (40)	37.29 (59)	90.91 (99)

Note: Number of observations (i.e., initiated organizations) in parentheses.

Result 5 *Organizations with less than four participants have a high likelihood of being rejected in the implementation stage of both IF treatments. Only organizations in which all players participate have a substantial likelihood of being implemented.*

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 we saw above, there is always a possibility that implementation fails and players end up at 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 next result shows, the overall impact of institution formation on the average contribution is positive.

[16]The fact that 30 percent of the full-size organizations are not implemented in treatment IF40 may seem surprising, but is mainly 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 full-size 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 full-size organization is 93 percent (59 observations) in the final ten rounds and 90 percent (30 observations) in the final five rounds.

Result 6 *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.*

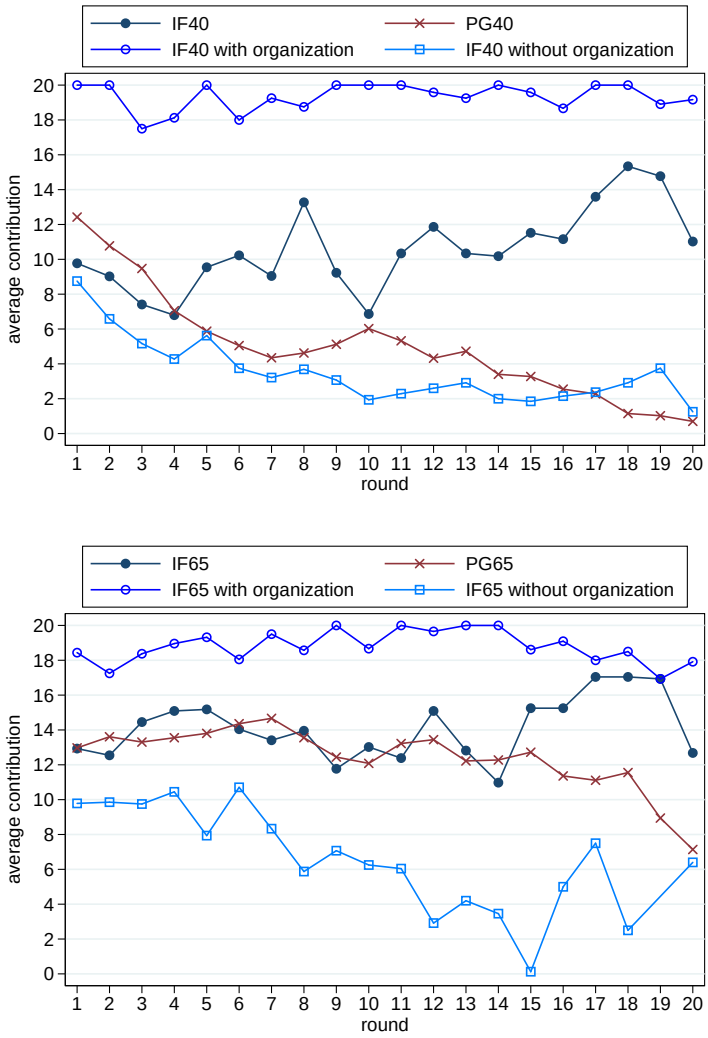

Figure 3: Average contribution to the public good with and without the possibility of institution formation (upper panel: IF40, PG40; lower panel: IF65, PG65).

Support for Result 6 is presented in Figure 3, which compares the average contributions to the public good in the institution formation treatments with those in the corresponding control treatments. Let us first consider treatment IF40 and the corresponding control treatment PG40 (upper 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, $p < .01$).

In the institution formation treatment IF65, a similar pattern emerges, yet on a slightly higher level. When the marginal per capita return of the public good is 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. Interestingly, the average contribution is also rather high in the control treatment PG65 (12.9 in round one and 14.7 in round seven). From this it follows that, up to round 14, the average contributions in treatments IF65 and PG65 do not differ statistically from each other. Contributions do not diverge clearly until the final rounds. While the average contribution falls to 7.1 in treatment PG65, it 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, $p = .48$). If we consider only the final six rounds, the difference becomes marginally significant (Mann-Whitney test, $p = .10$).[17]

Figure 3 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 3 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, $p = .23$). 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

[17]The non-members' average contribution to the public good is positive in both treatments, but far below the efficient level (average IF40: 4.0, IF65: 8.3). Interestingly, if other players are bound to contribute their full endowment, non-members contribute slightly more than they do when no organization exists (5.4 vs. 3.9 in IF40; 10.1 vs. 7.2 in IF65).

possible (Mann-Whitney test, $p = .07$). 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.

We finally analyze the level of efficiency in the different treatments. Note that group welfare is maximized when no organization is formed and each player still contributes his full endowment to the public good. To determine whether the possibility of institution formation increases group welfare, we calculate the actually achieved efficiency relative to the welfare maximum. That is, relative 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. Table 4 summarizes the results as well as Mann-Whitney test statistics comparing the efficiency in treatments IF40 and PG40 as well as in IF65 and PG65 for all rounds and for the final six rounds.

Table 4: Average relative efficiency in all treatments

	IF40	PG40	IF65	PG65
all rounds	0.51	0.25	0.70	0.62
	$p = .01$		$p = .47$	
final six rounds	0.62	0.09	0.77	0.52
	$p = .00$		$p = .10$	

Note: The table shows two-sided Mann-Whitney test statistics comparing the relative efficiency in the IF and PG treatments.

From the table it is clear that efficiency is higher in the IF than in the PG treatments, whereby the difference between treatments IF40 and PG40 is greater than that between IF65 and PG65. Taking all rounds into account, efficiency lies at 51 percent in IF40 and 25 percent in PG40. The efficiency gap between the two treatments widens strongly when only the final six rounds are taken into account (IF40: 62 percent, PG40: 9 percent). All differences are highly significant. For treatments IF65 and PG65, the observed pattern is similar: For all rounds and for the final six rounds, efficiency is greater in the institution formation treatments. However, due to the relatively high efficiency level in PG65, we fail to detect statistically significant differences. Importantly, in both PG treatments, efficiency decreases significantly over rounds (Spearman's $\rho = -.93, p = .00$ in PG40; $\rho = -.80, p = .00$ in PG65). In stark contrast, in the IF treatments, efficiency does not decrease but rather increases ($\rho = .73, p = .00$ in IF40; $\rho = .30, p = .19$ in IF65). We summarize these findings in the following result.

Result 7 *The possibility of institution formation has a positive effect on efficiency. The observed efficiency levels are higher when players are allowed to establish organizations than when they are not. Furthermore, in the institution formation treatments, efficiency remains constant or even increases over time, whereas it significantly decreases in the two treatments without institution formation.*

5 Discussion

We analyzed the endogenous formation of institutions in a linear public goods game in which players are allowed to establish an organization that punishes members who do not contribute the efficient amount to the public good. Our main results show that, despite the existence of a second-order free-rider problem, players overcome this problem and successfully form institutions. Importantly, however, the likelihood of effective implementation crucially depends on the number of participating players. In particular, we find that only full-size organizations, in which all players participate, have a reasonable chance of being implemented. The experimental results contrast considerably with the unique strict subgame perfect equilibrium in the one-shot game, which predicts that $s^* < n$ players will implement the organization. Nevertheless, the results are consistent with the theoretical model since subjects' behavior can be supported by non-strict Nash equilibria both in the one-shot and in the finitely repeated institution formation game. In particular, Proposition 1 shows that organizations of size $s > s^*$ can be supported by a disciplinary strategy whereby participating players reject any organization strictly smaller than s in the implementation stage. Notably, the data show that some groups use exactly this strategy to support the full-size organization, as illustrated by Figure 4.

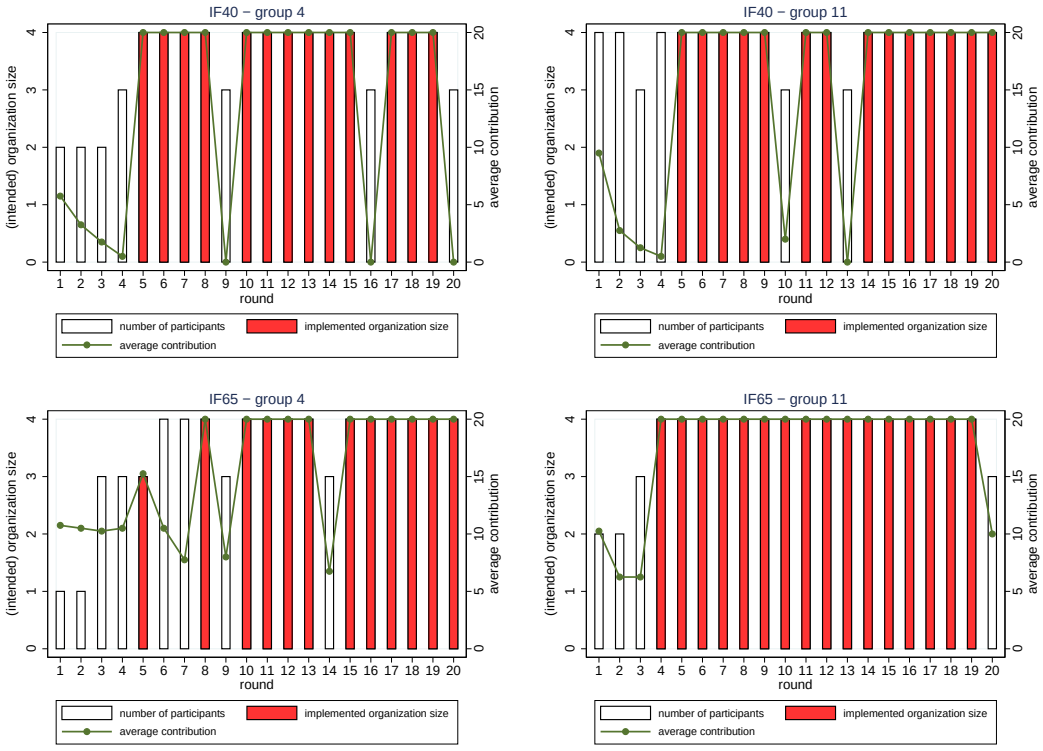

Figure 4: The dynamics of institution formation
(exemplary groups, upper panel: IF40; lower panel: IF65).

Figure 4 provides information about the dynamics of institution formation of four exemplary groups (two for each IF treatment) who implement the full-size organization most of the time. For each group, the figure shows for each round (i) the number of players who participate in the participation stage (light bars), (ii) the size of any organization established in the implementation stage (red bars), and (iii) the average contribution to the public good in the contribution stage (green line). As can be seen, none of these groups implements an organization in the first three rounds. There are often a few players who do not participate in the beginning. However, between rounds 4 and 8, players eventually start to implement full-size organizations and keep doing so for at least four rounds. Still, in each group, there is one player who deviates at least once by refusing to participate in the first stage of the game. The consequence is the same in all four groups: the remaining players reject the 3-player organization in the second stage and reduce the amount of their contribution significantly in the third stage. As a result, the deviating player starts participating again and a full-size organization is formed again in the subsequent round (except for the case of the second group in treatment IF65, in which the first deviation occurs in the final round).

While our main experimental findings can be explained by a model based on standard preferences, an interesting question is whether our results are also consistent with the assumption that players have some form of social preferences, such as a taste for fairness, equity, and efficiency. In fact, the question is of particular relevance in our set-up, since there is now considerable evidence that social preferences affect economic behavior in many important areas including the provision of public goods (for an overview see, e.g., Fehr and Gächter, 2000b; Camerer, 2003). A first indication of the role of social preferences comes from the observation in our experiment that non-participants in the contribution stage make positive contributions to the public good. In principle, this behavior can be explained by the model of Fehr and Schmidt (1999) if subjects are sufficiently inequity averse with regard to both disadvantageous and advantageous payoff allocations (i.e., using the Fehr-Schmidt notation, parameters α and β are sufficiently large). When β is large, a non-member incurs a significant utility loss if he free rides on the contributions of players who are members of the organization. If any other non-members are contributing less than he is, he is also concerned with disadvantageous inequality. Thus, when non-members are sufficiently inequity averse in both ways, the model predicts that these players will make positive contributions in equilibrium.[18] With regard to institution formation, the general principle governing participants' equilibrium behavior in the imple-

[18] More specifically, in the contribution stage of the institution formation game, all organization members are bound to contribute fully. Therefore, if players have a sufficiently large β and a relatively low α, then it may be optimal for non-members to contribute fully to avoid advantageous inequality. If players have a large α, a situation in which non-members make positive but less than full contributions can be sustained as an equilibrium. In this case, the large α prevents non-members from increasing their contributions.

mentation stage (Lemma 2) still holds. However, now the threshold for an organization to be implemented also depends on the parameters of the participants' inequity aversion and non-participants' contributions. If individuals are sufficiently disadvantageously inequity averse (with high α), then participating in any organization smaller than the largest one is not more profitable than the status quo with no organization. In this case, it can be shown that a 4-person organization can be attained as a unique strict equilibrium. Moreover, if α is such that the 3-person organization is not rejected by any of the participating players, the full-size organization may nevertheless be implemented in equilibrium. This is because if the remaining non-member is advantageously inequity averse (with high β), not participating makes him worse off; he therefore has an incentive to join the organization.

Our experimental results are also consistent with the existence of so-called conditional cooperators, who have a preference to cooperate if other players cooperate as well (see, e.g., Fischbacher et al., 2001). These individuals, contrary to selfish individuals, may make positive contributions to the public good even as non-members. Furthermore, if they anticipate that other players will participate in the organization, conditional cooperators are willing to participate in the organization as well.

Finally, an individual motive for efficiency would predict that the 4-person organization is formed and that both participants and non-participants contribute fully to the public good. Our experimental observations in the control treatments (PG40 and PG65), however, do not reveal evidence for a strong efficiency motive. Nevertheless, based on the observations in the institution formation game, one may conclude that a motive for efficiency strengthens the incentive to implement a full-size organization.

We employed the unanimous voting rule as a collective choice rule in the implementation stage. Our theoretical result for the institution formation game does not change in any crucial way if we employ majority voting as an alternative voting rule. To see this, suppose that an organization is implemented if at least $m \geq s/2+1$ participants accept the organization, where s is the number of participants. The first part of Lemma 2 still holds. That is, when s is greater than or equal to the threshold s^*, there exists two types of Nash equilibria with or without an organization. The second part of the lemma does not hold under the majority voting rule. Even when $s < s^*$, there exists a Nash equilibrium with an organization in the implementation stage. This type of equilibrium can be sustained if the number k of participants who accept an organization is strictly greater than the majority m. If this is the case, a unilateral deviation of any individual never affects the outcome of the implementation stage under the majority rule. Note, however, that every potential participant has an incentive to not participate in the first stage if he anticipates that such an equilibrium with an organization will prevail in the implementation stage with $s < s^*$. Therefore, Proposition 1 still holds under the majority rule.

References

Anderson, C. M. and Putterman, L. (2005). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, forthcoming.

Andreoni, J. (1993). An experimental test of the public-goods crowding-out hypothesis. *American Economic Review*, 83:1317–1327.

Camerer, C. (2003). *Behavioral game theory: experiments in strategic interaction*. Princeton University Press, New York and Princeton.

Carpenter, J. P. (2004). Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, forthcoming.

Carpenter, J. P. (2006). The demand for punishment. *Journal of Economic Behavior and Organization*, forthcoming.

Chan, K. S., Godby, R., Mestelman, S., and Muller, R. A. (2002). Crowding-out voluntary contributions to public goods. *Journal of Economic Behavior and Organization*, 48:305–317.

Chen, Y. and Plott, C. R. (1996). The groves-ledyard mechanism: An experimental study of institutional design. *Journal of Public Economics*, 59:335–364.

Egas, M. and Riedl, A. (2005). The economics of altruistic punishment and the demise of cooperation. Tinbergen Institute Discussion Paper TI 2005-065/1.

Falkinger, J. (1996). Efficient private provision of public goods by rewarding deviations from average. *Journal of Public Economics*, 62:413–422.

Falkinger, J. (2004). Noncooperative support of public norm enforcement in large societies. CESifo Working Paper No. 1368.

Falkinger, J., Fehr, E., Gächter, S., and Winter-Ebmer, R. (2000). A simple mechanism for the efficient provision of public goods: Experimental evidence. *American Economic Review*, 90:247–264.

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

Fehr, E. and Gächter, S. (2000b). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14:159–181.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415:980–994.

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

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71:397–404.

Groves, T. and Ledyard, J. (1977). Optimal allocation of public goods: a solution to the "free rider" problem. *Econometrica*, 45:783–810.

Gürerk, Özgür., Irlenbusch, B., and Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312:108–111.

Huck, S. and Weizsäcker, G. (2002). Do players correctly estimate what others do? evidence of conservatism in beliefs. *Journal of Economic Behavior and Organization*, 47:71–85.

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.

Nikiforakis, N. and Normann, H.-T. (2005). A comaprative statics analysis of punishment in public-good experiments. Technical report, Department of Economics, Royal Holloway, University of London.

Nyarko, Y. and Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, 70(3):971–1005.

Offerman, T., editor (1997). *Beliefs and Decision Rules in Public Good Games*. Kluwer, Dordrecht/Boston/London.

Okada, A. (1993). The possibility of cooperation in an n-person prisoners dilemma with institutional arrangements. *Public Choice*, 77:629–656.

Oliver, P. (1980). Rewards and punishments as selective incentives for collective action: theoretical investigations. *American Journal of Sociology*, 85:1356–1375.

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

Sefton, M., Shupp, R., and Walker, J. (2002). The effects of rewards and sanctions in provision of public goods. CeDEx Discussion Paper Series 2002-02.

Sonnemans, J., Schram, A., and Offerman, T. (1998). Public good provision and public bad prevention: the effect of framing. *Journal of Economic Behavior and Organization*, 34:143–161.

Sutter, M., Haigner, S., and Kocher, M. G. (2006). Choosing the stick or the carrot? endogenous institutional choice in social dilemma situations. mimeo.

Sutter, M. and Weck-Hannemann, H. (2004). An experimental test of the public-goods crowding-out hypothesis when taxation is endogenous. *Finanzarchiv*, 60:94–110.

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

Walker, J. M., Gardner, R., Herr, A., and Ostrom, E. (2000). Collective choice in the commons: Experimental results on proposed allocation rules and votes. *The Economic Journal*, 110:212–234.

WWR (2006). *Klimaatstrategie – tussen ambitie en realisme*. Amsterdam University Press, Amsterdam.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51:110–116.

Yamagishi, T. (1988). The provision of a sanctioning system in the united states and japan. *Social Psychology Quarterly*, 51:265–271.