On the Evolvement of Institutions in Social Dilemmas

ÖZGÜR GÜRERK
University of Erfurt

BERND IRLENBUSCH
London School of Economics

BETTINA ROCKENBACH
University of Erfurt

September 3, 2004

PRELIMINARY VERSION

Abstract

The question of how mechanisms sustaining cooperation in social dilemmas may evolve is still not answered satisfactorily. In this paper we experimentally investigate the performance and the reception of endogenously chosen institutions for enhancing cooperation in public good provision. We find that the initially low acceptance of the punishment institution steadily increases to its unambiguous prevalence over time. The institution choice option yields maximum efficiency in the punishment institution. Our results expand the previously observed superior performance of the punishment mechanism to cases in which its use is not obligatory. They strongly indicate that sanctioning may "evolutionary" prevail once it is invented. Interestingly, an endogenous reward institution is considerably less successful.

Keywords

Cooperation, social norms, endogenous institutions, punishment, rewards

JEL-Classification

C72, C92, H41

Acknowledgements

Financial support by the University of Erfurt is gratefully acknowledged.

Addresses

Özgür Gürerk	Bernd Irlenbusch	Bettina Rockenbach
Laboratory for Experimental Economics University of Erfurt Nordhaeuser Str. 63, D-99089 Erfurt, Germany e-mail: guererk@uni-erfurt.de www.uni-erfurt.de/elab	London School of Economics Interdisciplinary Institute of Management Houghton Street London WC2A 2AE e-mail: b.irlenbusch@lse.ac.uk http://www.lse.ac.uk/collections/iim	Chair in Microeconomics University of Erfurt Nordhaeuser Str. 63, D-99089 Erfurt, Germany e-mail: bettina.rockenbach@uni-erfurt.de www.uni-erfurt.de/mikrooekonomie

1. Introduction

In his anthropological classic *The Forest People* on the pygmies living in the Ituri forest (Congo), Turnbull (1987) describes in detail the experiences of Cephu, one of the male community members. Cephu's tribe had established a quite effective form of hunting by forming a circle around the envisaged prey. On the one hand it was in the collective interest that the circle remained closed so that no animals could escape. On the other hand each hunter had an individual incentive to jump in the direction of the prey and to catch it to the own benefit. This, however, would be disastrous for the group since an individual deviation opens the circle and would create a possibility for animals to escape. Some day it happened that Cephu opted out of the collective agreement to stay on the circle and he thereby gained from a considerable individual hunting success. Later, however, Cephu experienced the downside of his behavior: He was punished by almost all members of his tribe. For example others refused to speak with him; he was laughed at by women and children, and was completely ignored even by younger men. Turnbull argues that through this non-monetary but nevertheless effective punishment exerted by each single group member the pygmies were able to stabilize cooperation within their group.

Turnbull's story provides a neat impression of situations throughout the history of human societies in which cooperation was essential for survival. Quite often these situations have the structure of social dilemmas (Hardin 1968, Dawes 1980) which are characterized by conflicting individual and collective interests (examples include foraging, usage of common pool resources, predator avoidance, territory defense, parental care, food sharing, etc.). The collective interest demands a contribution of each member of society to the public good; however, it is in everyone's individual interest to free-ride on the contributions of the others. Hence, if everyone acts in his purely selfish interest the public good is not provided although it would be in the group's joint interest. The pygmies of the Ituri forest are only one example showing that human societies successfully established mechanisms for sustaining cooperation in social dilemma situations. An illustrative set of field studies is provided in Ostrom (1990), Bromley (1992), Ostrom, Gardner, Walker (1994).

To separate between different motives for cooperation and defection researchers have intensively studied behavior in dilemma situations in controlled laboratory experiments (for overviews see e.g. Ledyard 1995, Croson 1998, Ostrom 1998, Camerer 2003). In sequential social dilemmas remarkable levels of cooperation have been observed, even in interactions with completely anonymous strangers (Berg, Dickhaut, McCabe 1995, Fehr, Fischbacher, Gächter 2002, Fehr, Kirchsteiger, Riedl 1998). However, there is a considerable heterogeneity

among different cultures (Henrich et al. 2001). In repeated interactions cooperation is rarely stable and deteriorates to rather low levels towards the end of the interaction period. In a number of recent studies possibilities to punish norm violators have been identified as valuable means to sustain cooperation. Already relatively weak forms of punishment, like symbolic punishment (Masclet et al. 2003), have been shown to be quite effective. If the punishment allows to reduce the income of the punished subject, it is heavily used, even if punishing incurs costs for the punisher (Yamagishi 1986, Ostrom, Walker, Gardner 1992, Fehr and Gächter 2000). The lower an individual's contribution to the public good relative to the group average, the more the individual is punished. In repeated interactions punishment may be attributed to selfish incentives. In a recent study by Fehr and Gächter (2002) a potential motivation for own material gains from punishment is excluded because the experimental design guarantees that punished and punishing subjects never interact again. Nevertheless punishment is frequently observed and, moreover, the future partners of the punished subjects benefit from the punishment because the punished subjects typically increase their contribution. Hence, considerable evidence from different disciplines has accumulated showing that punishment is an effective mechanisms for norm enforcement. In addition it has been shown that the rewarding norm abiding behavior has a much weaker effect on cooperation (Sefton, Shupp, Walker 2002).

The question, however, of how mechanisms sustaining cooperation in social dilemma situations may evolve in human societies is still not answered in a satisfactory way.[1] Given the superior performance of the punishment mechanism it may seem obvious that it prevails even in a world with competing mechanisms to encourage cooperation. This, however, is not evident at all. If competing institutions are available people may well be reluctant to join the punishment institution due to the fear of being punished unjustified and the bad atmosphere caused by punishment execution. In the current paper we focus on the evolvement and the performance of different institutions of treating norm violation and norm abidance in a laboratory experiment. The novelty of our approach is that the subjects endogenously decide which mechanism for enhancing norm abiding behavior should govern their interaction. This allows us to study the initial and the over-time acceptance of different mechanisms and their ability to sustain cooperation in social dilemma situations. In this sense we model a partially

[1] The question of how altruistic punishment may survive evolutionary pressures even in relatively large groups has been addressed in recent models from evolutionary biologists (Bowles and Gintis 2004, Boyd et al. 2003). Henrich and Boyd (2001) present an evolutionary model in which certain types of behavior can establish norms of punishment in the population.

evolutionary process in which successful mechanisms may spread and unaccepted or ineffective mechanisms may be extinguished.

We extend previous studies by Fehr and Gächter (2000, 2002) by adding an institution choice stage prior to the public good provision game. At the beginning of each repetition each subject may choose to either join an institution with costly punishment possibilities – after being informed about the contributions of the others – or to enter an institution without any sanctioning options. In each round a subject then interacts in the public goods game with all other subjects which chose the same institution in that round.[2] Two remarkable results can be observed. First, initially, more than two-thirds of the subjects decide to interact in the regime without a punishment possibility, but over the rounds the proportion of the subjects in this institution steadily decreases with an almost complete extinction towards the end of the interaction. Second, the subjects that voluntarily choose to interact in the punishment institution heavily punished free-riders and achieved almost full cooperation. This cooperation is surprisingly stable and it even continues when the interacting group becomes large (because almost all subjects join this group) and the experiment approaches its end. In a control treatment in which subjects are exogenously allocated to the two institutions and have no possibility to change institutions the contribution levels in the punishment institution are significantly lower. This shows that it is not merely the punishment possibility but the option to select the institution and to interact only with those that have selected the same institution that leads to the efficient provision levels. Interestingly, a choice between an institution that allows costly rewarding of other subjects and a baseline institution without any sanctioning or rewarding technology, does not lead to such clear results. In both of these institutions a decay of cooperation over time is observed and subjects move back and forth between the two institutions with about rather stable 70 percent of the subjects opting for the reward institution.

[2] Ehrhart and Keser (1999), Hauk and Nagel (2001), Coricelli, Fehr, Fellner (2003) and Putterman, Page, and Unel (2003) also model a choice stage previous to the interaction in a social dilemma situation. In this choice stage, the participants select the partners with whom they are going to interact. In real-life, however, situations where someone can directly select the interaction partners in a social dilemma situation are limited (for example spouses, employees). Quite often it is only possible to select the rules that govern the interaction, but not the concrete partners themselves (for example corporate culture, political regime). Our study addresses the latter case in which partner choice only happens indirectly, because each subject interacts with the "like-minded" people who chose the same institution. Recently, the endogenous partner selection and its effects on behavior is also explored in other contexts, for example market interactions (Kirchsteiger, Niederle, and Potters, forthcoming, and Brown, Falk, and Fehr 2004), and networks (Riedl and Ule 2003).

2. A simple model of institution selection

In this section we present a simple model that allows us to study the endogenous evolvement of institutions in a public good provision dilemma. Prior to the public good provision each subject decides upon joining one of two institutions that differ with respect to the reaction possibilities to norm violating and norm abiding behavior. Each subject then interacts with all other subjects that chose the same institution. Three institutions are considered: the FREE institution resembles the standard voluntary public good provision mechanism in which subjects have no possibility of influencing the others' payoffs after having observed individual contributions to the public good; in the punishment (PUN) and the reward (REW) institution, subjects may punish or reward other players, respectively.

In the following, we first describe the FPN game which models the public goods provision with an endogenous choice between the FREE and PUN institutions. Then we introduce the FRN game where subjects may join either the FREE or the REW institution. We complete section 2 by introducing the games FPX and FRX. These two games serve as controls and differ from FPN and FRN in that the subjects are exogenously assigned to their institutions. Table 1 gives an overview over our 2x2 design.

Table 1: Overview over the institution characteristics of the studied games

		Available institutions	
		FREE and **P**UN	**F**REE and **R**EW
Institution assignment	e**N**dogenously	FPN game	FRN game
	e**X**ogenously	FPX game	FRX game

2.1. Institution Choice: Punishment

We consider a three stage game, consisting of an institution choice stage (S0), a voluntary contribution stage (S1) and a punishment stage (S2). With respect to stages S1 and S2 our game is analogous to the game considered in Fehr and Gächter (2000, 2002), however, we augment their design by the institution choice stage S0.

In S0, each of the $N = 12$ members of the population simultaneously chooses the institution she wants to join. The institution choice does not incur any cost. Each player may either join the FREE institution, in which case no decision has to be taken in S2. Or she may join the PUN institution, in which case players have the possibility to punish others in S2. Once all players finished their institution choice in S0, they are informed about the respective number

of players that chose each of the two institutions. The identity of the players, however, is not revealed.

In the contribution stage S1, a player interacts solely with those players that opted for the same institution. Each player *i* receives an endowment of 20 tokens from which she can contribute $0 \le g_i \le 20$ tokens to a public good to the benefit of all members of her institution. The remainder $20 - g_i$ is transferred to player *i*'s private account. If $n > 1$ players join an institution, one token invested in the public good of that institution yields a marginal per capita return (MPCR) of $a(n)$ with $0 < a(n) < 1 < na(n)$ for each member of that institution.[3] After all players have simultaneously made their contribution decision, each is informed about the contributions of each member in the own group.

At the beginning of the punishment stage S2 each player receives additional 20 tokens independent from the affiliation to an institution. In the PUN institution these tokens may be used to punish other members of their group.[4] For the institution members of FREE, the active part of the game ends here. The total payoff of player *i* in the FREE institution is

$$\pi_i^{FREE} = \left(20 - g_i + a(n)\sum_{j=1}^{n} g_j\right) + 20 . \qquad (1)$$

Obviously, for players in FREE contributing zero is the dominant strategy. This is true independent from the group size. If everybody applies this strategy, the individual earning is 40 tokens. However, the joint payoff of the group is maximized if each group member contributes her whole endowment. In this case each member's payoff rises to 52 tokens, independent from the group size.

In the PUN institution all players simultaneously decide whether or not to punish other members of their own group. Player *i* can punish group member *j* by assigning punishment tokens t_j^i to player *j*. Each token assigned by player *i* to player *j* incurs a cost of one token for

[3] If only one subject joins an institution no *public* good can be created and his total endowment is automatically transferred to his private account and he has no decision in stages S1 and S2. Note, that if $n > 1$ the marginal per capita return $a(n)$ varies with the number of subjects in the institution. In order to guarantee that the joint maximal payoff of the group is always the same independent from the group size, we adjust $a(n)$. This issue will be discussed in Section 3.

[4] Hence, in PUN all subjects are equipped with the same punishment possibilities, independent from their contributions in S1. Providing the subjects in the FREE institution with the same additional endowment eliminates the incentives to choose the PUN institution just for the extra "punishment tokens".

player *i* and reduces the payoff of player *j* by three tokens.[5] Each player may in total assign up to 20 punishment tokens. Let T^i denote the amount of punishment tokens that player *i* assigns and T^{-i} denote the amount of punishment tokens that player *i* receives from the other members of her group. The total payoff of player *i* in PUN results in

$$\pi_i^{PUN} = \left(20 - g_i + a(n)\sum_{j=1}^{n} g_j\right) + \left(20 - T^i - 3T^{-i}\right). \quad (2)$$

The expressions in parentheses represent the stage payoffs of S1 and S2 respectively. In equilibrium no monetary payoff-maximizing player will take the opportunity to punish since it is costly to do so. By applying backward induction rational players foresee this circumstance and they will not contribute in S1. Hence, independent from the group size in PUN, not to contribute and not to punish is the dominant strategy in the one-shot game. In this case the total payoff is 40 tokens, which is the same as myopically rational and purely money maximizing players obtain in FREE. Again, each player's payoff would be 52 tokens if all members of the group fully cooperate.

At the end of the game all players are informed about all other players' contributions, their punishment tokens assigned, their punishment tokens received and their resulting total payoff.

2.2. Institution Choice: Rewards

In S0 of the FRN game players have to choose between joining the FREE and the REW institutions. Players who join the FREE institution are confronted with an identical stage game as in FPN. Hence the predictions for the dominant strategy and payoffs are the same. Players who join the REW institution have the possibility to reward the members of their group after they are informed about the contributions of all group members. Player *i* can reward group member *j* by assigning reward tokens t_j^i to player *j*. Each token assigned by player *i* to player *j* costs one token for *i* and increases the payoff of *j* by one token. The reward mechanism is efficiency neutral, i.e. the tokens used for rewarding are redistributed between the sender and the receiver. The total payoff of player *i* from all stages of the game is

$$\pi_i^{REW} = \left(20 - g_i + a(n)\sum_{j=1}^{n} g_j\right) + \left(20 - T^i + T^{-i}\right). \quad (3)$$

[5] Our punishment mechanism is basically the same as applied elsewhere, for example by Abbink, Irlenbusch and Renner (2000), Fehr and Gächter (2002), or Andreoni, Harbaugh, Vesterlund (2003) and reflects the widely accepted assumption, that in general punishing someone is less costly than being punished. By this mechanism, the punisher can reduce an absolute "inequity" to his own disadvantage since the punishment action reduces the income of the punished player more than the own income is diminished. A variety of other possible punishment mechanisms can also be implemented in public goods (see e.g. Decker, Stiehler, Strobel 2003, Kosfeld and Riedl 2004).

The expressions in parentheses again represent the stage payoffs of S1 and S2 respectively. Again independent from the group size not to contribute and not to reward is the dominant strategy in the one-shot game. If everybody employs the equilibrium strategy each group member earns 40 tokens. However, the joint payoff of the group is maximized if each group member contributes the whole endowment. In this case each member's period payoff would be 52 tokens.

2.3. Exogenous Matching: Punishment / Rewards

In contrast to the games introduced so far, in the exogenous matching scheme the players are randomly assigned to two disjoint co-existing institutions of equal and constant size $n = 6$: FREE and PUN in the FPX game, and FREE and REW in the FRX game, respectively. Thus, in the exogenous games we eliminate the institution choice stage S0 and employ only stages S1 and S2. The information and feedback properties are the same as in the endogenous games. The payoff functions are also identical. As a consequence the game theoretical predictions for the one-shot game are exactly the same as in the endogenous matching scheme. Hence the dominant strategy for a rational player is neither to contribute nor to punish or to reward.

3. Hypotheses

The strategic situations induced by the four games presented in Section 2 constitute the framework for the research question we aim to analyze. We arrange our hypotheses corresponding to the three stages of the game. Applying the backward induction logic, we first set up the hypothesis for the final stage S2 where punishment and reward take place. Following this we present our hypotheses for contribution stage S1. Finally, our conjectures for the institution choice stage S0 are stated.

3.1. Punishment and Rewards

Dependent on the game in stage S2, subjects may punish (PUN) or reward (REW) each other. If agents are purely selfish and rational, no one will make use of either punishment or rewarding possibilities since it is costly to do so. However, several experimental studies show that people are willing to punish or reward others even if they forego substantial amounts of monetary payoff. For example Fehr and Gächter (2000, 2002) and Sefton, Shupp and Walker (2002) observe that subjects who contribute less than the group average are punished quite frequently. Moreover, the intensity of punishment increases with the negative deviation of the subjects' contribution from the group average. Analogously, recipients of rewards have contributed in general more than average (Sefton, Shupp and Walker, 2002). However, the

latter study also finds that subjects allocate rewards not so generously as punishments. For larger deviations below the group average the punishment is much higher than rewards for contributions considerably above the group average. In the light of these experimental findings we set up the following hypotheses for the punishment and reward behavior on stage S2 of the PUN and REW institutions respectively:

Hypothesis Punishment: *Free riding behavior is punished. We expect punishment exerted on free riders to increase with the negative deviation from the group average.*

Hypothesis Reward: *Cooperative behavior is rewarded. We expect rewards received by cooperators to increase with the positive deviation from the group average.*

3.2. Contributions

Rational and purely money-maximizing agents should not contribute to the public good in S1 because the marginal return from a token invested into the public good is lower than from a token transferred to the private account. A huge number of studies (see Ledyard 1995), however, show that experimental subjects contribute substantial amounts to the public good, especially in the beginning of the experiments. Moreover, punishment and rewards institutions lead even to higher contributions than in the pure voluntary contribution mechanism (Fehr and Gächter 2000, 2002, Sefton, Shupp and Walker 2002). Concerning the contributions level on S1 we set up the hypotheses:

Hypothesis Contributions in PUN: *Average contributions are expected to be higher in the PUN institution than in FREE.*

Hypothesis Contributions in REW: *Average contributions are expected to be higher in the REW institution than in FREE.*

3.3. The Impact of the Possibility to Choose on Contributions

How does the possibility to select the institution affect the subjects' contribution behavior? A purely money maximizing subject is indifferent between both available institutions and his behavior is not affected by the possibility to choose the institution. However, one may suppose that the fact that subjects deliberately choose the interaction regime evokes higher degrees of commitment and responsibility and thus a higher motivation for cooperation.[6]

[6] Experimental observations from Orbell and Dawes (1993), Bohnet and Kübler (forthcoming), and Maier-Rigaud and Apesteguia (2003) in Prisoner's Dilemma settings seem to point in this direction.

To control for the influence of the possibility to select institutions, we conduct the control treatments FPX and FRX that allow us to separate the pure selection effect from the effects of the institutions available.

3.4. Institution Choice

In a world of purely selfish and rational actors, an agent is indifferent between the three institutions, because in each of them the identical payoff of 40 will be achieved. Joining the FREE institution in the FPN game, however, is a weakly dominant strategy because there is - independent from the contribution – at least a possibility to be punished in PUN, while in FREE this is not the case. Similarly, a rational player in FRN should join the reward institution since there is a possibility to receive rewards.[7] From the previous experimental evidence, however, contributions can be expected to be higher in the PUN and the REW institutions compared to the FREE institution and that may attract more subjects to join these institutions rather than FREE. Hence, in the FRN game we unambiguously expect the agents to join the REW institution. In case of the FPN game, the situation is somewhat ambivalent. Subjects may choose PUN assuming that more cooperators joining this institution.[8] However, they risk to be punished. Therefore, we state our institution choice hypothesis for FPN and FRN respectively as follows:

Hypothesis Institution Choice FPN (i): *The subjects are expected to join the FREE institution rather than PUN.*

Or, alternatively:

Hypothesis Institution Choice FPN (ii): *The subjects are expected to join the PUN institution rather than FREE.*

Hypothesis Institution Choice FRN: *The subjects are expected to join the REW institution rather than FREE.*

3.5. The Evolvement of the Institutions

As argued above we unambiguously expect subjects to join the REW institution in FRN and contributions (and thus payoffs) in REW to be higher than in FREE. As a consequence we

[7] Recall that punishing or rewarding would constitute off-equilibrium behavior.

[8] This conjecture is supported by a similar effect found by Ehrhart and Keser (1999). In their public goods experiment without any punishment or reward mechanism, subjects may switch to other groups or create a new group by some cost. Ehrhart and Keser observe that groups with a higher level of cooperation are subsequently entered quite often by free riders. This discourages the cooperators and they "flee" from the free riders by forming new groups.

also expect that this institution will clearly "survive" and that the FREE institution will be extinguished. In FPN the picture is not so clear. We have conflicting hypotheses whether subjects may join the PUN institution or not and the question whether the expected higher contributions and the expected high punishment activity may yield a lower or a higher payoff compared to the alternative FREE institution cannot be answered unambiguously. Hence, the evolvement of the two institutions in FPN remains an open issue.

4. Experimental Design and Procedure

Following the model described in Section 2 our experimental design consists of four treatments (see Table 2), that differ with respect to the possibility to change institutions or not (endogenous versus exogenous) and the institutions available in the choice set.

Table 2: Experimental Treatments

Treatment Name	Co-existing Institutions	Institution Choice	# Members in an Institution	# Independent Observations
FRX	FREE and REW	exogenous	$n = 6$	6
FPX	FREE and PUN	exogenous	$n = 6$	6
FRN	FREE and REW	endogenous	$0 \le n \le 12$	8
FPN	FREE and PUN	endogenous	$0 \le n \le 12$	8

In each of our four treatments $N = 12$ subjects constitute a population that remains constant during the whole session.[9] In each session one of the four games described in Section 2 is repeated over 30 rounds.[10] In the exogenous matching treatments, the 12 members of the population are randomly assigned in two subsets of equal size of $n = 6$, each interacting in one of the institutions. Subjects' affiliations to an institution remain fixed during the whole session. In contrast, in the endogenous matching treatments the group size n may vary in each round. This circumstance raises a crucial experimental design question: Either the *marginal per capita return* (MPCR) from a contribution to the public good or the *productivity*, that is the total rate of return for the whole group from each token invested to the public good, varies with the group size. In order to give smaller groups the possibility to be as productive as larger groups we decided to keep the productivity constant with $R = na(n) = 1.6$. Thus, independent from the group size, the (maximal) return from the public good is always the same if all members of the group contribute their whole endowment. Consequently the MPCR

[9] For an investigation of the difference in public good behavior of strangers and partners see e.g. Croson (1996) or Keser, van Winden (2000).
[10] By randomly reshuffling the presentation ordering on the computer screens we made sure that the identity of the players could not be traced over rounds.

decreases with an increase in the group size.[11] Table 3 shows the MPCR dependent on the group size. In the exogenous matching treatments, the MPCR is constant and given by $a(6) = 0.27$. Note that the threat of punishment and the opportunities to be rewarded are greater in larger groups, as in total there are more punishment respective reward tokens available.

Table 3: Marginal per Capita Return $a(n)$

Group Size n	2	3	4	5	*6*	7	8	9	10	11	12
Marginal per Capita Return $a(n)$	0.80	0.53	0.40	0.32	*0.27*	0.23	0.20	0.18	0.16	0.15	0.13
Productivity R	1.6	1.6	1.6	1.6	*1.6*	1.6	1.6	1.6	1.6	1.6	1.6

The experiments were conducted in the computerized laboratory *eLab* of the University Erfurt. The subjects were recruited for voluntary participation on campus. All subjects were undergraduate students with majors predominant in economics, sociology, history, or educational science. Each subject was allowed to participate only once and none had participated in a similar experiment before. The experiment was programmed and conducted with the software *z-Tree* (Fischbacher 1999). Communication other than via the experimental software was not allowed.[12]

In the experiment a total of 336 subjects participated in 28 sessions with 12 subjects each. The sessions constitute the independent observations: eight for each of both endogenous treatments and six for each of the exogenous ones. Before starting the experiment the instructions were read aloud to all participants.[13] The subjects were informed about the experimental procedure as well as the respective matching protocol and the number of rounds. One session lasted for about 2 to 2.5 hours. The subjects earned between 15 and 25 Euros and payment was anonymously performed at the end of the experiment.

[11] Isaac and Walker (1988) examine the effects of different MPCRs in public good experiments. They find significantly more free-riding behavior in their low MPCR treatment (0.30) than in the high MPCR condition (0.75). Of course this observation emerges from a static comparison between treatments and not from a dynamic change of the MPCR within a population.

[12] On the influence of communication and (non-binding) promises on contribution behavior in public goods see Isaac and Walker (1988), Orbell, Van de Kragt, Dawes (1988), and Ostrom, Walker, Gardner (1992).

[13] A translation of the instruction sheet is given in Appendix. Original instructions were written in German. They are available upon request from the authors.

5. Results

In this section we report our experimental findings. First we address our main research question on the institution choice behavior and its evolvement over time. We continue by comparing the development of the contributions between treatments and institutions. Finally, the punishment and reward behavior is investigated which leads us to efficiency considerations. All reported non-parametrical statistical tests are based on the averages over the independent observations.

5.1. Institution Choice

In the first round of FPN 69 percent of participants join the FREE institution while only 31 percent opt for the PUN institution. In FRN, 76 percent choose REW whereas only 24 percent decide to join FREE.[14] These observations provide support for two of our hypotheses on institution choice: initially more than two-thirds of the participants seem to escape from the punishment threat but subjects show a clear affinity towards the reward mechanism.

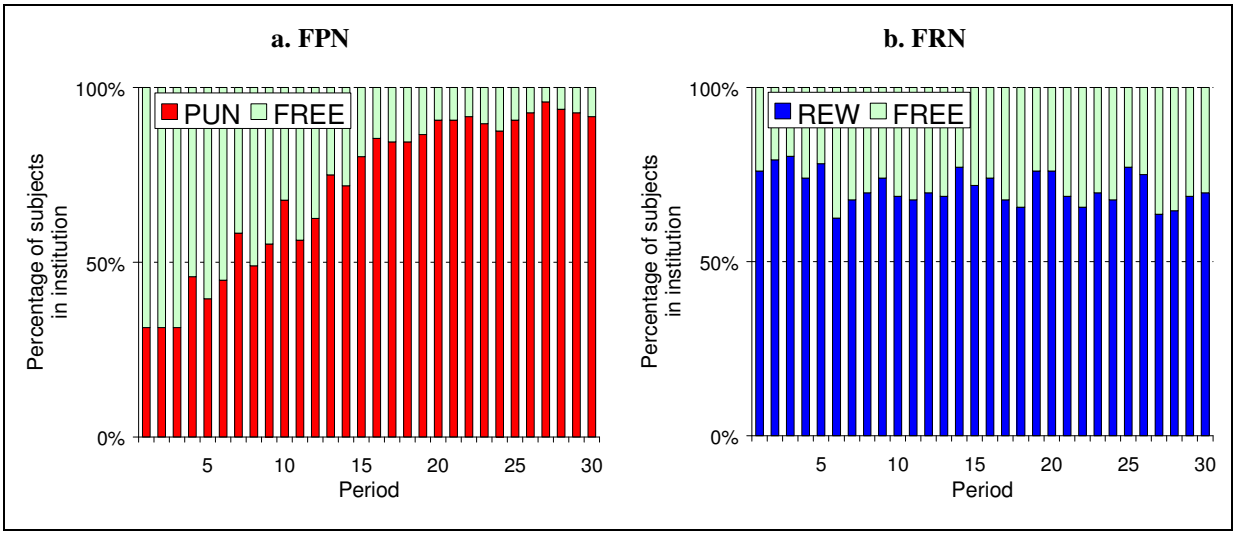
The development of the institution choices is displayed in Figure 1: Panel a) displays the percentage of subjects in the two institutions of treatment FPN in the course of the rounds and Panel b) displays this data for treatment FRN. The initial high acceptance of the FREE institution in FPN diminishes steadily with an almost complete extinction towards the end of the interaction. In the final rounds more than 90 percent have joint the PUN institution.[15] In FRN, in contrast, the number of subjects in both institutions is rather constant over the whole experiment.[16]

[14] One-sided Mann-Whitney U-Tests show that significantly more subjects choose the FREE institution than PUN ($p = 0.003$) and more subjects prefer REW to FREE ($p = 0.000$), respectively.

[15] In each of the eight independent observations of FPN the number of members in FREE has a negative trend over time (determined by Spearman-rank-correlation coefficients). Hence a two-sided binomial test with these correlation coefficients clearly rejects that a negative trend is as likely as a positive trend ($p = 0.016$).

[16] For three of the eight independent observations of FRN the number of members in FREE has a negative trend over time, in one observation the Spearman rank correlation test detects a positive trend, in the remaining four observations no trend can be found.

Figure 1: Institution Choice over Rounds



Fluctuation between the institutions

Since each subject was able to freely choose between the two institutions at the beginning of each round, the question concerning the fluctuation between the institutions arises. In the first five rounds of both treatments about 25 percent of the subjects change the institution, i.e. the institution chosen in the actual round is different from the institution chosen in the previous round. In rounds 6-10, the relative frequency of changes in the FPN treatment increases to almost 35 percent and from there on it steadily decreases with nearly no institution changes towards the end. In contrast, in FRN the fraction of institution changes remains constant between 20 and 25 percent. Hence, we observe a relatively stable institution composition after round 10 in FPN and a considerable fluctuation, even in later rounds, in FRN. This behavioral pattern is also supported by the individual numbers of institution changes: 75 percent of the subjects in FPN change the institution at most five times. However, in FRN, more than 50 percent of subjects switch even eight times or more.

5.2. Contributions

One of the most important questions regarding the provision of public goods concerns the induced contribution behavior. Table 4 summarizes the contribution levels in the different treatments and institutions, averaged over all 30 periods as well as separately for the first and the second half. Additionally average contributions of the first and the last round are shown. Figure 2 displays the development of the contributions over time in the different institutions of the four treatments.

The influence of the available institutions on the contributions

Averaged over all 30 periods, the contributions in the treatments that include a punishment institution (FPN and FPX) are significantly higher than in the treatments that include a reward institution (FRN and FRX).[17] This observation is well in line with the results of the study by Sefton, Shupp, and Walker (2002). In the last three rounds of their *sanction* treatment the contributions are significantly higher than in the *reward* treatment. In their additional experiment with an extended time horizon of 20 periods, the difference between sanction and rewards treatments is even bigger, even though the punishment mechanism they implemented is less effective, i.e. achieving the same level of punishment is more costly to the punisher than in our study.

Table 4: Average Contributions

Treatment	Periods 1-30		Periods 1-15		Periods 16-30		First Period		Last Period	
FPN	14.3 (8.2)		10.8 (8.8)		17.8 (5.9)		9.2 (6.6)		18.3 (5.5)	
FPX	9.8 (7.5)		9.3 (6.7)		10.2 (8.1)		8.6 (5.6)		9.0 (8.5)	
FRN	3.0 (4.7)		4.3 (5.3)		1.7 (3.6)		7.5 (5.4)		0.8 (2.4)	
FRX	6.8 (7.2)		8.5 (7.1)		5.2 (6.9)		9.6 (6.0)		3.0 (6.5)	
Institution	FREE	PUN	FREE	PUN	FREE	PUN	FREE	PUN	FREE	PUN
FPN	2.9 (4.9)	18.8 (3.6)	3.1 (4.9)	17.6 (5.0)	2.0 (4.8)	19.6 (2.7)	7.4 (5.9)	13.1 (6.5)	0.1 (0.3)	20.0 (0.1)
FPX	5.4 (6.9)	14.2 (5.0)	6.5 (7.0)	12.2 (5.0)	4.3 (6.7)	16.2 (4.1)	8.7 (5.5)	8.6 (5.7)	2.2 (5.5)	15.8 (4.8)
Institution	FREE	REW	FREE	REW	FREE	REW	FREE	REW	FREE	REW
FRN	2.3 (3.9)	3.3 (5.0)	3.2 (4.4)	4.7 (5.6)	1.5 (3.1)	1.8 (3.7)	5.4 (4.2)	8.1 (5.6)	0.3 (1.0)	1.0 (2.7)
FRX	4.5 (5.7)	9.2 (7.7)	5.8 (6.0)	11.2 (7.1)	3.2 (5.1)	7.1 (7.8)	9.1 (6.2)	10.2 (5.7)	2.0 (5.0)	4.1 (7.5)

Note: Numbers in parentheses are standard deviations.

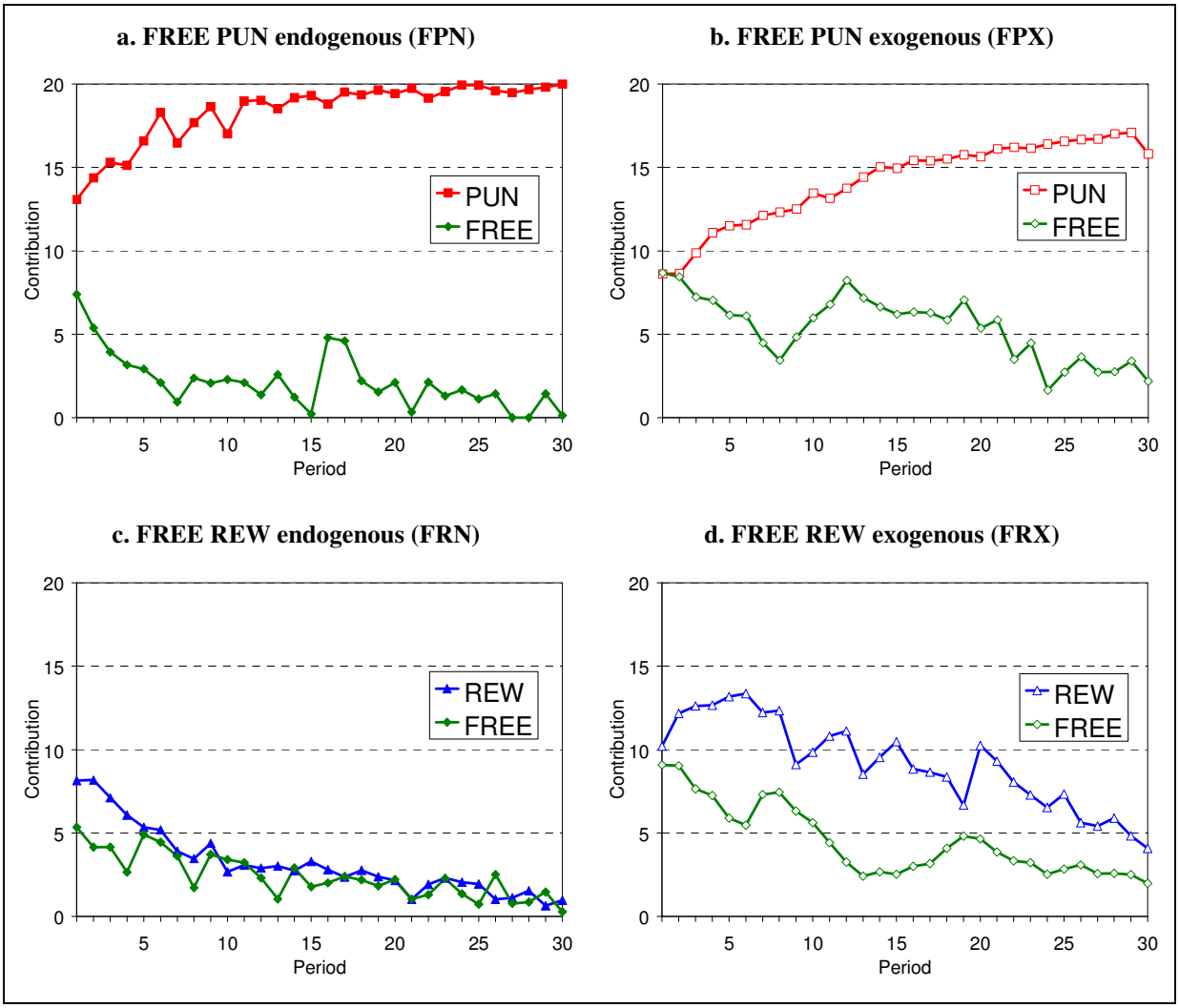
Within each treatment the FREE institution performs worst; it exhibits significantly lower contributions than the alternative institution that allows an influence of the other subjects'

[17] One-sided Mann-Whitney U-Tests reveal that contributions are significantly higher in FPN than in FRN and FRX at $p = 0.000$ and $p = 0.001$ respectively. Contributions observed in the FPX treatment are significantly higher than in the FRN ($p = 0.000$) and the FRX ($p = 0.066$) treatments.

[19] One-sided Wilcoxon matched-pairs tests show that in the endogenous matching scheme, contributions in PUN and in REW are significantly higher than in the respective FREE institution both at $p = 0.004$. In the exogenous scheme the contributions in PUN and REW are again significantly higher than in FREE at $p = 0.016$ and $p = 0.047$ respectively.

payoffs (PUN and REW, respectively).[19] This observation is also in line with the results of the studies by Fehr and Gächter (2000) and Sefton, Shupp and Walker (2002). A comparison of all our sessions shows that the highest contributions are observed in the PUN institution in the endogenous matching scheme, where, on average, subjects contribute more than 90 percent of their endowment.

Figure 2: Contributions over Rounds



The influence of the matching protocol on the contributions

A comparison of the two treatments that allow for punishment (FPN and FPX) shows that FPN exhibits significantly higher contribution levels than FPX (Mann-Whitney U-Test, one-sided, $p = 0.015$). As mentioned above the treatments with the exogenous matching protocol were run as control treatments that differ from our baseline treatments only with respect to the matching procedure. The extreme difference between the contribution levels in FPN and FPX

thus shows that it is not the punishment possibility alone that leads to these high contribution levels. The endogenous choice of the institution and the knowledge that all other interaction partners also deliberately chose that institution is decisive for the evolvement of the high cooperation. Compared to previous research (Fehr and Gächter 2000, Sefton, Shupp, Walker 2002) the contribution levels in the punishment institution of FPN are extremely high and as Figure 2, Panel a) shows they reach almost full cooperation towards the end of the interaction period. This is especially noteworthy since it is has been shown in previous studies (for example by Isaac and Walker 1988) that cooperative behavior is much more difficult to establish if the group size is higher and the MPCR is low.[21] As can be seen from Figure 1 our interaction groups are quite large in the final third (about 9 to 12 players) and as a consequence the MPCR is relatively low.

Interestingly, we also find a significant difference in the two treatments that allow for rewarding (FRX and FRN), however, in the opposite direction. If the interaction group is formed endogenously we observe lower contributions than if the group is composed exogenously (Mann-Whitney U-Test, one-sided, $p = 0.001$). This effect may be attributed to the considerable fluctuation in the endogenous REW treatment that makes it more difficult to establish a group norm than in the fixed composition of the exogenous matching.

First round contributions

In the exogenous matching scheme, the players contribute roughly 50 percent of the endowment in the first round. This result is in line with typical findings from previous public goods experiments (see Ledyard 1995). The contributions between the two respective institutions of the exogenous treatments are not statistically different. On the contrary, in the endogenous matching scheme, first round contributions vary from 27 percent in the FREE institution of the FRN treatment to 65 percent in the PUN institution of the FPN treatment. [22] The highest first round contribution is observed in the PUN institution of the endogenous matching scheme that is significantly higher than in all other institutions. This observation shows that the fact that the subjects were able to choose the institution and that they know that

[21] Carpenter (2004), however, observes that cooperation is not necessarily reduced with an increasing group size if mutual monitoring and punishment possibilities are available.

[22] One-sided Mann-Whitney U-Tests reveal that first round contributions are higher in PUN and REW than in the FREE institutions of the respective treatments at $p = 0.000$ level and $p = 0.021$.

they interact with subjects who also voluntarily chose this institution makes an important difference right from the beginning of the interaction.[23]

Development of the contributions over time

Over time, we observe a clear increase in contributions in the PUN institutions whereas there is a decreasing trend both in REW and FREE.[24] This supports the previously made observation that the punishment possibility is a more effective mechanism for achieving cooperation in social dilemmas than the pure reward mechanism (Sefton, Shupp, Walker 2002). However, the most striking result is that the members of the endogenous PUN institution establish and maintain an almost perfect cooperation. In the second half of the rounds, subjects of this institution contribute on average 98 percent of their endowment to the public good and this is significantly more than the members of the exogenous PUN institution (Mann-Whitney U-Test, one-sided, $p = 0.039$).

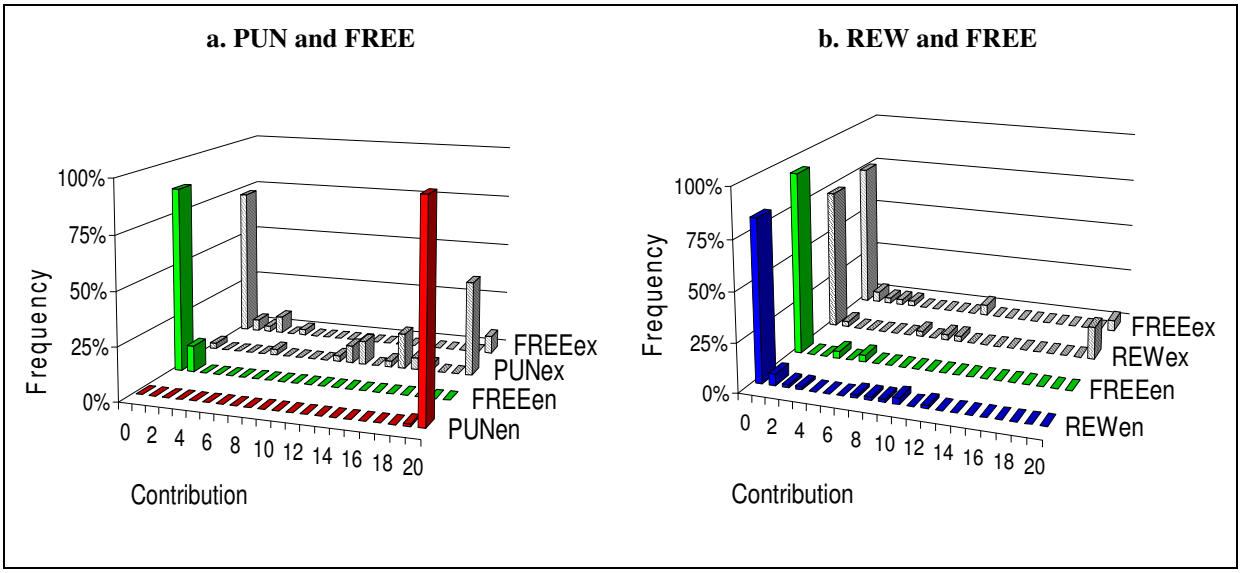
Last round contributions

Especially in the FPN treatment, the endogenous matching scheme leads to a surprisingly clear differentiation in last round contributions of the two institutions PUN and FREE. In the last round of the experiment, members of the PUN institution of FPN contribute almost their complete endowment (99.95 percent) to the public good whereas in the FREE institution of FPN almost everything is kept in the private account (99.36 percent). Hence, we do not observe an endgame effect in PUN; in contrast, the contributions even rise. Figure 3 shows the distribution of the last round contributions. In that round, 88 out of 96 subjects who participated in the FPN treatment join the PUN institution. 87 out of that 88 subjects contribute 20 in that round, the other one 19. On the contrary in the exogenous scheme with the same institutions available the contributions vary considerably. Only 44 percent contribute fully and 50 percent contribute between 10 and 20 tokens. In the FREE institutions of both matching schemes and in the REW institutions the vast majority of subjects free ride in the last round.

[23] In a recent study, Gächter and Thöni (2004) report the results of a public good provision experiment in which subjects were exogenously sorted according to their first round contribution. The three players with the three highest contributions formed an interaction group for a public goods game, etc. This exogenous sorting of with respect to contributions "like-minded" people had a strong positive influence on the contribution levels. In a similar study Gunnthorsdottir et al. (2001) sort the participants in groups, however without informing them exactly about the sorting rule.

[24] One-sided Wilcoxon matched-pairs tests reject the null hypotheses that the first and the second half of the rounds have the same mean for the PUN institution of the FPN treatment ($p = 0.004$), for the PUN institution of the FPX treatment ($p = 0.016$); for the REW institution of the FRN treatment ($p = 0.004$), for the REW institution of the FRX treatment ($p = 0.016$); for the FREE institution of the FPN treatment ($p = 0.109$), for the FREE institution in the FPX treatment ($p = 0.016$); for the FREE institution of the FRN treatment ($p = 0.004$), for the FREE institution in the FRX treatment ($p = 0.031$).

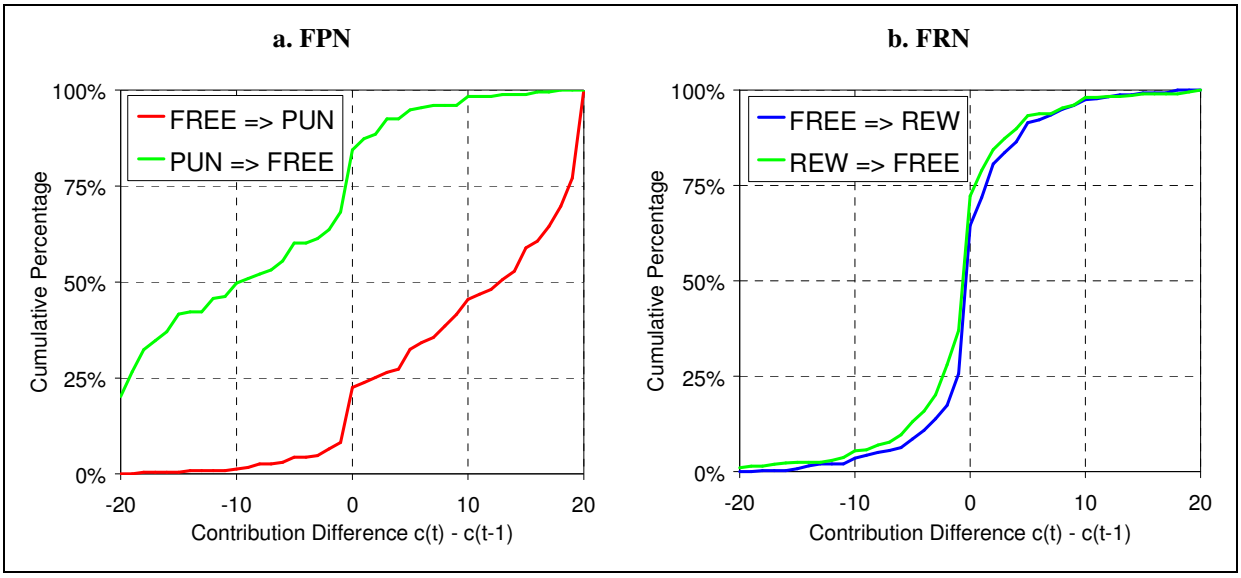
Figure 3: Distribution of the Last Round Contributions



Contributions and institution change

An interesting question is whether and how subjects change their contribution when they switch to another institution. The answer would provide insights whether there exists an implicit understanding of how to behave in a certain institution. Figure 4 illustrates the change in contribution immediately after a subject has moved to the other institution. We calculate the difference between the subject's contribution in round *t* and the contribution of that subject in round *t*-1 conditional upon the subject changed the institution from round *t*-1 to round *t*. The cumulative distributions over these contribution differences reveal a completely different pattern in the two endogenous treatments. In FPN, more than 75 percent of the subjects raise their contributions when they switch from FREE to PUN. About 50 percent contribute even at least 10 tokens more compared to the previous round in FREE. Moreover, roughly 25 percent switch from complete free riding (contribution 0) in FREE to full cooperation (contribute 20) after they change from FREE to PUN. In contrast, roughly 75 percent of subjects who switch from PUN to FREE reduce their contributions. Almost 25 percent allocate even nothing after a change. This observation is well in line with the experimental findings by Falk, Gächter, and Fischbacher (2003). They show that the contribution behavior of subjects is influenced by social interactions with their "neighbors". Subjects who simultaneously belong to two different groups with disjoint group compositions exhibit conditionally cooperative behavior, i.e. the same subject contributes more if she is in a group with high contributors and contributes less if she is in a group with low contributors.

Figure 4: Cumulative Distribution of Contribution Differences after a Change



In FRN, a change between institutions barely has an effect on subject's contribution behavior. About 50 percent of subjects do not change their contributions at all, independent from the change direction. In this treatment, virtually no one ever increases or decreases her contribution after a change by more than 10 tokens.

5.3. Punishment and Rewards

The exertion of punishment

The punishment behavior of the subjects in the different matching protocols is displayed in Panels a), c), and e) of Figure 5. Panel a) shows the average number of punishment tokens a subject allocates in stage S2 dependent on the round of the experiment. In both matching schemes punishment diminishes over time. Panels c) and e) address the question of who is punished. Panel c) displays the received punishment points dependent on the difference between the punisher's and the punished subject's contribution. For both matching protocols we observe that the lower the contribution of the other subject compared to the own contribution the more severely this subject is punished. Complete free riding behavior is punished most heavily by other group members.

Panel e) of Figure 5 displays the received punishment points dependent on the difference between the average contribution in the PUN institution and the punished subject's contribution. We find a positive correlation between the deviation and the points received.[25]

[25] In each of the eight independent observations of FPN and the six independent observations of FPX, subjects receive the more punishment tokens the more they deviate from the respective institution average (all Spearman-rank-correlation-coefficients computed for each independent observation are positive). Hence a one-sided binomial test with these correlation coefficients rejects that a negative correlation is as likely as a positive correlation for FPN ($p = 0.008$) and for FPX ($p = 0.031$).

Independent of the matching protocol, we see that subjects who contribute less than 14 tokens compared to the group average are on average punished with more than 30 punishment points, i.e. they have to suffer from an income reduction of more than 90 points. This means that they incur a true loss in that round since the saved contribution of 20 point is by far out weighted by the received punishment. Thus on average, any attempt to completely free ride is not tolerated by the other group members. This observation is in line with the results reported in Figure 5 by Fehr and Gächter (2000). They also find that the punishment of free riders increases with the negative deviation from the group average.

The decline of the amount of punishment (as shown in Panel a) may be attributed to the fact that due to the increasing contributions over time (see 5.1) punishing defectors become rarer. In the endogenous treatment the punishment of defectors can serve two purposes: it may either encourage the defectors to make higher contributions or it may encourage them to leave this institution. This twofold argument in favor of punishment may partly explain its more heavily usage in the endogenous matching than in the exogenous one as seen in Panel a) of Figure 5.[26] It can also be shown that the vast majority of subjects who punish in the endogenous matching scheme are the high contributors who punish low contributors much more heavily than in the exogenous matching scheme.[27]

The exertion of rewards

The reward behavior of the subjects in the different matching protocols is displayed in Panels b), d), and f) of Figure 5, analogous to panels displaying the punishment behavior. Rewards are typically exerted by high contributors to other high contributors.[28] Hence, rewarding is a pure transfer within the group of cooperative subjects. This exhibits an essential structural difference between punishment and rewards. Punishment is addressed to norm violators in order to encourage them to obey the social norm of cooperation. Rewarding, however, is addressed to the subjects already obeying the cooperation and has no disciplining instrument on norm violators. They may only be indirectly "convinced" to cooperate by observing that

[26] The punishment intensity in the endogenous matching scheme is higher than in the exogenous one: On average in the endogenous matching scheme 2.82 tokens are assigned to one punishment instance which is significantly more than in the exogenous matching where on average only 2.08 tokens are assigned to each instance (Mann-Whitney U-Test, one-sided, $p = 0.01$).

[27] In FPN, 85 percent of all subjects who punish contribute at least as much as the institution average; in FPX, only 64 percent of punishers contribute at least the average. Mann-Whitney U-Test (one-sided, $p = 0.005$) reveals that in the endogenous punishment institution, the ratio of punishers who are below average contributors is significantly lower than in the exogenous institution. In FPN, on average, below average contributors are punished with 10.35 (in FPX 2.58 tokens) tokens whereas others receive 0.33 tokens (in FPX 0.52 tokens).

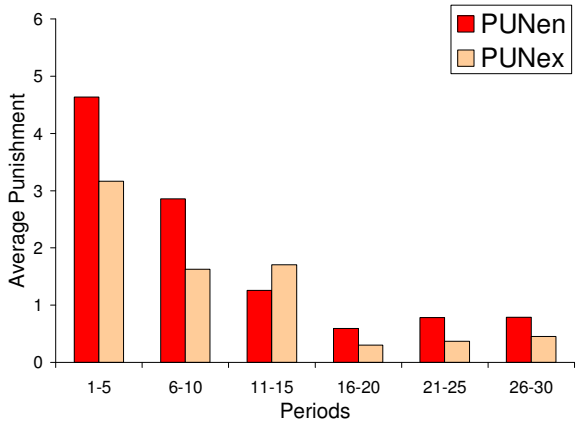
[28] In the FRN treatment, the below average contributors receive 0.31 whereas others 2.67 (in FRX 0.59/3.62). In FRN, below average contributors send on average 0.72 reward tokens to other institution members whereas the group of participants who contribute at least as the group average send 1.93 tokens (in FRX 2.56/1.85).

cooperators receive additional rewards, but not by directly addressing them. This indirect mechanism does not seem to be as effective as the direct mechanism of punishment. Hence, the decline in rewarding may be a direct consequence of the decline in the contributions. In contrast to punishment, rewards are more heavily used in the exogenous than in the endogenous matching scheme.[29] The stable group composition in the exogenous matching may yield a higher incentive to demonstrate that norm abiding behavior may pay.

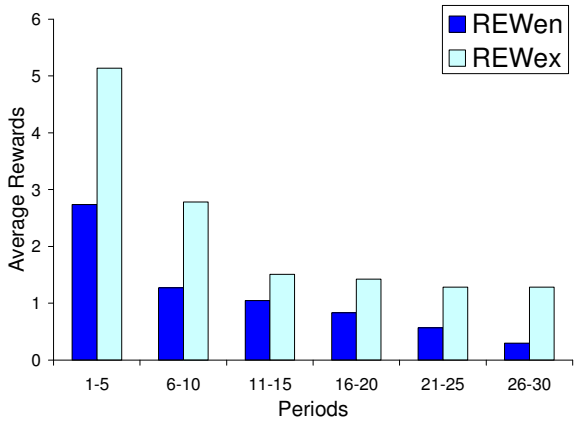
[29] The average number of reward tokens exerted per player is significantly higher in the exogenous matching scheme than in the endogenous one (Mann-Whitney U-Test, one-sided, $p = 0.006$). Also the reward intensity is significantly higher in FRX than in FRN: averaged over all 30 periods, 1.87 tokens are assigned to one reward instance in FRX whereas there are only 1.55 in FRN (Mann-Whitney U-Test, one-sided, $p = 0.041$).

Figure 5: Punishment and Rewards

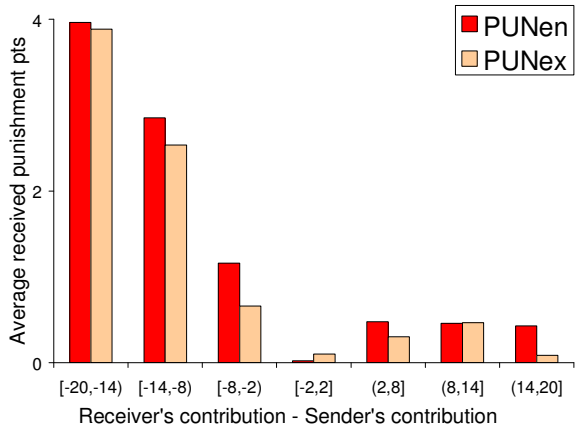
a. Average allocated punishment tokens per player over rounds



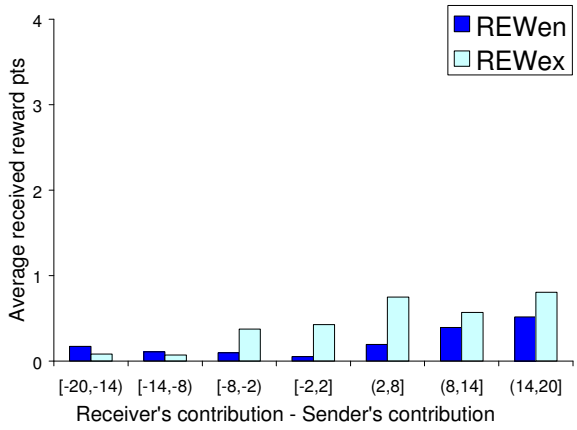
b. Average allocated reward tokens per player over rounds



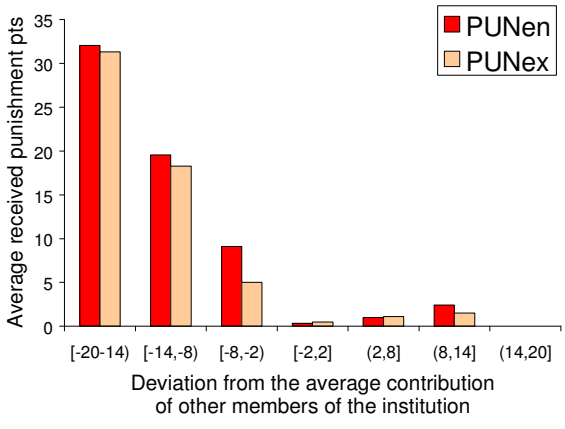
c. Average received punishment tokens, dependent on contribution difference between receiver and sender



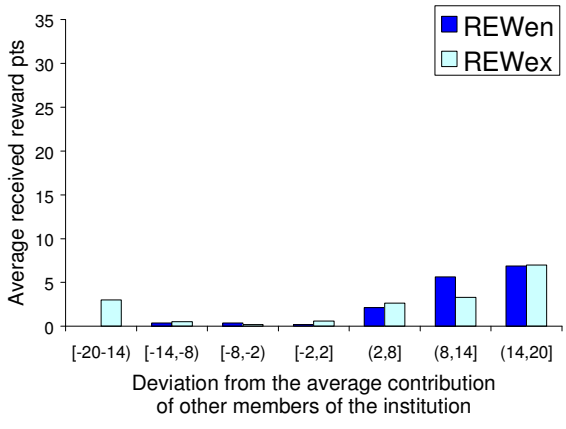
d. Average received reward tokens, dependent on contribution difference between receiver and sender



e. Average received punishment tokens, dependent on the deviation from the average contribution in the institution



f. Average received reward tokens, dependent on the deviation from the average contribution in the institution



5.4. Payoffs and Efficiency

In the social optimum, each subject fully contributes to the public good and refrains from punishing. This results in a payoff of 52 for each subject, independent of the size of the interaction group, the institution, and the matching protocol. Full contribution, however, does not constitute an equilibrium; in the Nash equilibrium each subject completely free rides and refrains from punishing and rewarding. This results in a payoff of 40 independent from the group size, the matching and the institution. Efficiency is measured as the ratio of the actually achieved payoffs and the socially optimal payoff of 52 and hence directly correlated to payoffs. Table 5 depicts the achieved average payoffs over all rounds, over the first and the second half of the rounds, as well as over the first and the last round. Figure 6 displays the development of the payoffs over the rounds for each of the four treatments and for reasons of comparison indicates the two benchmarks social optimum and Nash equilibrium.

Table 5: Average Payoffs

Treatment	Periods 1-30		Periods 1-15		Periods 16-30		First Period		Last Period	
FPN	44.5 (12.5)		41.0 (13.4)		48.1 (10.3)		37.8 (15.3)		50.4 (3.9)	
FPX	43.3 (9.0)		41.3 (10.9)		45.4 (6.1)		36.1 (12.5)		42.5 (4.5)	
FRN	41.8 (4.2)		42.6 (4.9)		41.0 (3.3)		44.5 (6.4)		40.5 (2.3)	
FRX	44.1 (5.8)		45.1 (5.9)		43.1 (5.6)		45.8 (5.3)		41.8 (6.0)	
Institution	FREE	PUN	FREE	PUN	FREE	PUN	FREE	PUN	FREE	PUN
FPN	41.7 (4.4)	45.6 (14.3)	41.9 (4.4)	40.3 (17.8)	41.2 (4.1)	48.8 (10.5)	44.4 (5.7)	23.0 (19.0)	40.1 (0.3)	51.3 (2.4)
FPX	43.2 (6.0)	43.4 (11.3)	43.9 (5.9)	38.6 (13.7)	42.5 (5.9)	48.2 (4.8)	45.2 (5.2)	27.1 (11.0)	41.3 (4.6)	43.7 (4.0)
Institution	FREE	REW	FREE	REW	FREE	REW	FREE	REW	FREE	REW
FRN	41.4 (3.3)	42.0 (4.6)	41.9 (3.7)	42.8 (5.3)	40.9 (2.8)	41.1 (3.5)	43.2 (3.9)	44.9 (7.0)	40.2 (1.0)	40.6 (2.7)
FRX	42.7 (4.6)	45.5 (6.5)	43.4 (5.1)	46.7 (6.1)	41.9 (4.0)	44.3 (6.6)	45.4 (6.0)	46.1 (4.5)	41.2 (4.2)	42.4 (7.3)

Note: Numbers in parentheses are standard deviations.

The most striking observation is that in the first half of the experiment the payoffs in the punishment institutions are substantially lower than in the FREE institutions of the respective treatments.[34] The higher initial contributions in the punishment institutions are "eaten up" by

[34] One-sided Wilcoxon matched-pairs test backs this observation for FPN ($p = 0.059$), for FPX ($p = 0.031$)

the higher expenses for punishment. This observation is in accordance to previous research (Sefton, Shupp and Walker 2002, Fehr and Gächter 2002). Hence, changing to the PUN institution in FPN cannot be attributed to choosing the institution that generates the highest average (short-run) payoff. Actually we observe that in the first half of the experiment almost 40% of all changes from FREE to PUN are made although the last round average payoff in PUN is lower than in FREE. A change to the less profitable PUN institution cannot be explained by the prospect of higher future payoffs, because subjects are free to change at any time in the game and they might also wait until the free riding diminishes. Joining the PUN institution in the early stages with high contributions and high punishment expenses thus means the contribution to two public goods: to the public good provided in stage S1 of that period and to the (second order) public good of enforcing the norm of cooperation via punishment in the long run of the game (for the later see also Yamagishi 1986).

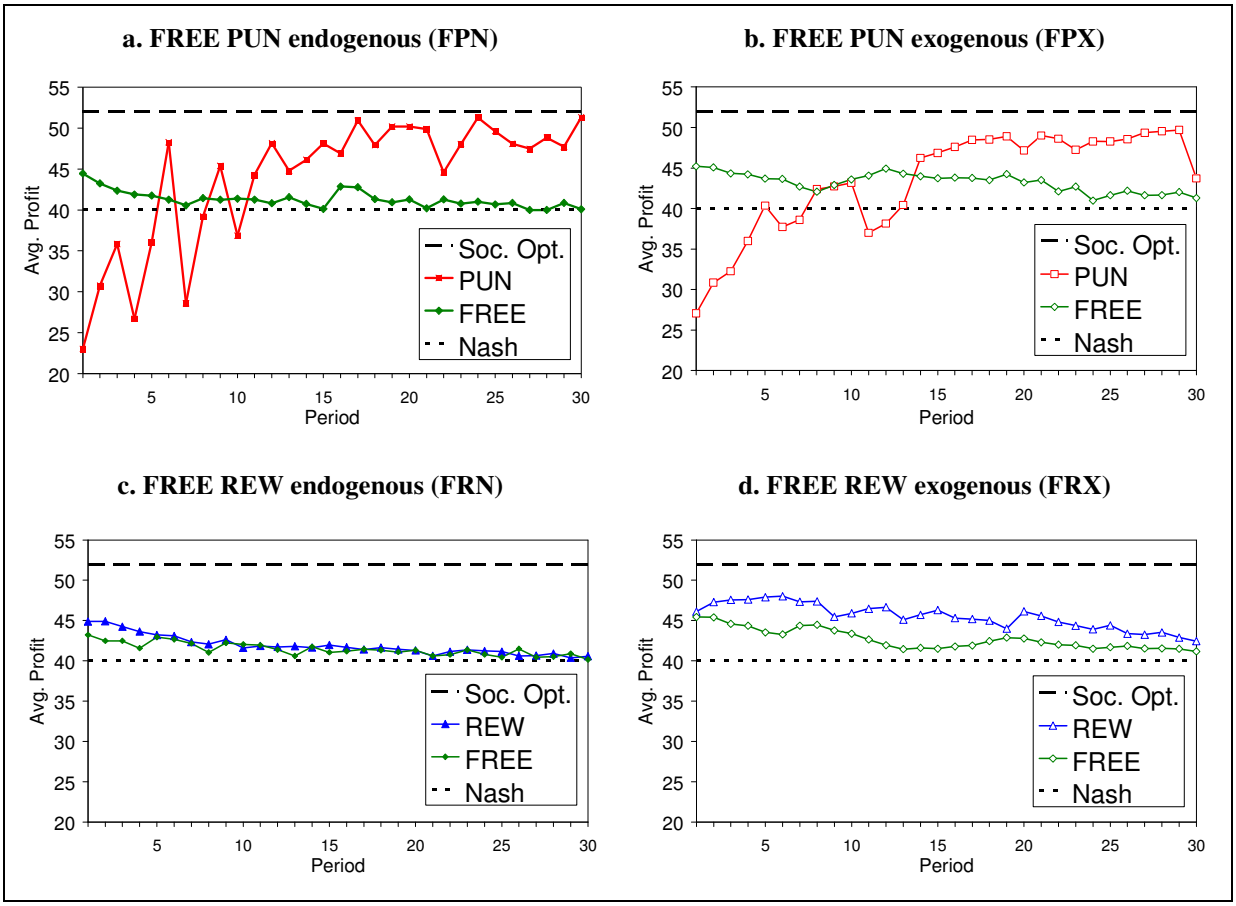
However, in the second half of the experiment the payoffs in the punishment institutions steadily increase towards the social optimum. The payoffs in the corresponding FREE institutions, in contrast, follow a decreasing trend and approach the Nash equilibrium. Similarly the payoffs in all four institutions in the reward treatments (FREE and REW in FRN and FREE and REW in FRX) follow a decreasing trend.[35]

A social planner may be interested in the question which of the proposed treatments generates the highest social welfare. The FPN treatment clearly wins this contest, because the long run payoffs achieved in FPN are significantly higher than the payoffs in each of the other treatments.[36] Moreover, providing the subjects with the endogenous choice of the punishment and the standard institution (as FPN does) generates a higher welfare than the possibility of punishment alone.

[35] Wilcoxon matched-pairs tests reveal that the payoffs in the second half of the experiment are significantly higher than in the first half of the experiment in the following institutions: PUN in FPN ($p = 0.004$), PUN in FPX ($p = 0.016$). On contrary, payoffs fall significantly in FREE in FPN ($p = 0.109$), FREE in FPX ($p = 0.016$), REW in FRN ($p = 0.004$), REW in FRX ($p = 0.016$), FREE in FRN ($p = 0.004$), FREE in FRX ($p = 0.031$).

[36] In the second half of the experiment, one-sided Mann Whitney U-Test shows significant differences between FPN and FPX, ($p = 0.054$), FPN and FRN ($p = 0.000$), FPN and FRX respectively ($p = 0.004$).

Figure 6: Payoffs over Rounds



6. Conclusion

In this study we experimentally investigate the evolvement of institutions in a public good provision setting. In general we find that the possibility to punish norm abiding behavior is heavily used and leads to higher levels of cooperation than in the absence of this possibility. It turns out, however, that the degree of cooperation is even stronger and reaches full cooperation if the subjects endogenously decide to interact in a punishment regime. Initially the alternative standard institution without any possibilities to influence the others' payoff is mostly chosen, but over time the punishment institution becomes selected with an increasing frequency. In the final rounds almost all subjects join the punishment institution. Despite severe punishment in the beginning, the punishment institution leads to the highest efficiency levels with full contribution of all participants and no punishment in the end. A control treatment with an exogenous allocation of the subjects to the institutions leads to significantly lower contributions. This demonstrates the importance of the endogenous selection process: subjects voluntarily choose the institution and know that the others with whom they interact

have also intentionally chosen that institution. An endogenous rewarding institution, however, does not lead to a comparable effect.

With this study we provide a first step to the understanding of the evolvement of institutions in social dilemmas. We show that a punishment mechanism "survives" in direct comparison to an alternative mechanism without any sanctioning possibilities. This result expands the previous analyses that have demonstrated the superior performance of punishment mechanisms in case their use was obligatory and it strongly indicates that punishment mechanisms (like social sanctions) may "evolutionary" prevail once they are invented. To some extent this might explain why, as in Cephu's example, sanctioning mechanisms are well established to treat norm violators.

References

Abbink, Klaus; Irlenbusch, Bernd, and Renner, Elke. "The Moonlighting Game – An Experimental Study of Reciprocity and Retribution." *Journal of Economic Behavior and Organization*, 2000, *42*, pp. 265-277.

Andreoni, James; Harbaugh, William, and Vesterlund, Lise. "The Carrot or the Stick: Rewards, Punishment, and Cooperation." *American Economic Review*, 2003, *93*(3), pp. 893-902.

Berg, Joyce; Dickhaut, John, and McCabe, Kevin. "Trust, Reciprocity and Social History." *Games and Economic Behavior*, 1995, *10*(1), pp. 122-42.

Bowles, Samuel and Gintis, Herbert. "The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations." *Theoretical Population Biology*, 2004, *65*(1), pp. 17-28.

Bohnet, Iris and Kübler, Dorothea. "Compensating the Cooperators: Is Sorting in the Prisoner's Dilemma Possible?" forthcoming in: *Journal of Economic Behavior and Organization*.

Boyd, Robert; Gintis, Herbert, Bowles, Samuel, and Richerson, Peter J. "The Evolution of Altruistic Punishment." *Proceedings of the National Academy of Sciences of the United States of America*, 2003, *100*(6), pp. 3531-3535.

Boyd, Robert and Richerson, Peter J. "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups." *Ethology and Sociobiology*, 1992, *13*(3), pp. 171-195.

Bromley, Daniel W.. (Ed.) "Making the Commons Work: Theory, Practice and Policy." *San Francisco, Institute for Contemporary Studies*, 1992

Brown, Martin; Falk, Armin, and Fehr, Ernst. "Relational Contracts and the Nature of Market Interactions." *Econometrica*, 2004, *72*, pp. 747-780.

Carpenter, Jeffrey. "Punishing Free-Riders: how group size affects mutual monitoring and the provision of public goods." Forthcoming in *Games and Economic Behavior*, 2004

Croson, Rachel T.A. "Partners and strangers revisited." *Economics Letters*, 1996, *53*, pp. 25-32.

Croson, Rachel T.A. "Theories of Altruism and Reciprocity: Evidence from Linear Public Goods Games." *Working Paper*, 1998, *98-11-04*, The Wharton School of the University of Pennsylvania.

Dawes, Robyn M.. "Social Dilemmas." *Annual Review of Psychology*, 1980, *5*, pp. 163-193.

Decker, Torsten, Stiehler, Andreas and Strobel, Martin. "A Comparison of Punishment Rules in Repeated Public Good Games." *Journal of Conflict Resolution*, 2003, *47*(6), pp. 751-772.

Ehrhart, Karl-Martin and Keser, Claudia. "Mobility and Cooperation: On the Run." *Working Paper 99s-24*, 1999, CIRANO, Montreal.

Falk, Armin; Gächter, Simon, and Fischbacher, Urs. "Living in Two Neighborhoods - Social Interactions in the Lab.", *Working Paper*, 2003, *No. 150*, University of Zurich.

Fehr, Ernst; Fischbacher, Urs, and Gächter, Simon. "Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms." *Human Nature-an Interdisciplinary Biosocial Perspective*, 2002, *13*(1), pp. 1-25.

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

Fehr, Ernst and Gächter, Simon. "Altruistic Punishment in Humans." *Nature*, 2002, *415*, pp. 137-140.

Fehr, Ernst; Kirchsteiger, Georg, and Riedl, Arno. "Gift Exchange and Reciprocity in Competitive Experimental Markets." *European Economic Review*, 1998, *42*(1), pp. 1-34.

Fischbacher, Urs. "z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual." *Working Paper*, 1999, *Nr. 21*, Institute for Empirical Research in Economics, University of Zurich.

Gächter, Simon and Thöni, Christian. "Voluntary Cooperation among Like-minded People." Mimeo, University of St. Gallen, 2004

Gunnthorsdottir, Anna; Houser, Daniel, McCabe, Kevin, and Ameden, Holly. "Disposition, history and contributions in public goods experiments." *mimeo*, 2001

Hardin, Garrett. "The Tragedy of the Commons." *Science*, 1968, *162*, pp. 1243-1248.

Henrich, Joseph; Boyd, Robert, Bowles, Samuel, Camerer, Colin, Fehr, Ernst, Gintis, Herbert, and McElreath, Richard. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *American Economic Review*, 2002, *91*(2), pp. 73-78.

Henrich, Joseph and Boyd, Robert. "Why People Punish Defectors - Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas." *Journal of Theoretical Biology*, 2001, *208*(1), pp. 79-89.

Isaac, Mark and Walker, James. "Communication and Free Riding Behavior: The Voluntary Contributions Mechanism." *Economic Inquiry*, 1988, *26*(2), pp. 585-608.

Isaac, Mark and Walker, James. "Group Size Hypotheses of Public Goods Provision: An Experimental Examination." *Quarterly Journal of Economics*, 1988, *103*, pp. 179-199.

Keser, Claudia and van Winden, Frans. "Conditional Cooperation and Voluntary Contributions to Public Goods." *Scandinavian Journal of Economics*, 2000, *102* (1), pp. 23-39.

Kirchsteiger, Georg, Niederle, Muriel and Potters, Jan. "Endogenizing Market Institutions: An Experimental Approach)." Forthcoming in: *European Economic Review*,

Kosfeld, Michael and Riedl, Arno. "The Design of (De)centralized Punishment Institutions for Sustaining Cooperation." *mimeo*, 2004.

Ledyard, John. "Public Goods: A Survey of Experimental Research." J. Kagel and A. Roth, *Handbook of Experimental Economics*, 1995, Princeton University Press, pp. 111-194.

Maier-Rigaud, Frank P. and Apesteguia, Jose. "The Role of Choice in Social Dilemma Experiments." Preprint *Bonn Econ Discussion Papers*, 2003/7.

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

Orbell, John M., Van de Kragt, Alphons J.C., and Dawes, Robyn M.. "Explaining Discussion Induced Cooperation." *Journal of Personality and Social Psychology*, 1988, *54*(5), pp. 811-819.

Orbell, John M. and Dawes, Robyn M.. "Social Welfare, Cooperators' Advantage, and the Option of Not Playing the Game." *American Sociological Review*, 1993, *58*(6), pp. 787-800.

Ostrom, Elinor. "Governing the Commons: The Evolution of Institutions for Collective Action." *New York: Cambridge University Press*, 1990.

Ostrom, Elinor. "A Behavioral Approach to the Rational Choice Theory of Collective Action." *American Political Science Review*, 1998, *92*(1), pp. 1-23.

Ostrom, Elinor; Walker, James, and Gardner, Roy. "Covenants with and without a Sword: Self-Governance is Possible." *American Political Science Review*, 1992, *86*, pp. 404-417.

Ostrom, Elinor; Gardner, Roy, and Walker, James. "Rules, Games and Common-Pool Resources." *Ann Arbor: University of Michigan Press*, 1994.

Page, T.; Putterman, L., and Unel, B.. "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency." *Brown University Working Paper*, 2003

Riedl, Arno and Ule, Aljaz. "Exclusion and Cooperation in Social Network Experiments." mimeo, 2003, University of Amsterdam.

Sefton, Martin; Shupp, Robert, and Walker, James. "The Effect of Rewards and Sanctions in the Provision of Public Goods." *CEDEX Research Paper*, 2002, University of Nottingham.

Turnbull, Colin M.. "The Forest People." *Touchstone Books*, 1987.

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

Appendix: Instruction Example FPN

Instructions for the experiment

General Information: At the beginning of the experiment you will be randomly assigned to one of **2 subpopulations each consisting of 12 participants**. During the whole experiment you will interact only with the members of your subpopulation. At the beginning of the experiment, **1000 experimental tokens** will be assigned to the experimental account of each participant.

Course of Action: The experiment consists of **30 rounds**. Each round consists of 2 stages. In Stage 1, the group choice and the decision regarding the contribution to the project take place. In Stage 2, participants may influence the earnings of the other group members.

Stage 1

(i) The Group Choice: In Stage 1, each participant decides which group she wants to join. There are two different groups that can be joined:

	Influence on the earnings of other group members
Group	A: No B: Yes, by assigning negative points

(ii) Contributing to the Project: In stage 1 of each round, each group member is endowed with 20 tokens. You have to decide how many of the 20 tokens you are going to contribute to the project. The remaining tokens will be kept by yourself.

Calculation of your payoff in stage 1: Your payoff in stage 1 consists of two components:

- **tokens you have kept** = endowment – your contribution to the project
- **earnings from the project** = 1.6 x sum of the contributions of all group members / number of group members

Thus, **your payoff in Stage 1** amounts to:
20 – your contribution to the project
+ 1.6x sum of the contributions of all group members / number of group members

The earnings from the project are calculated according to this formula for each group member. **Please note:** Each group member receives the same earnings from the project, i.e. each group member benefits from **all** contributions to the project.

Stage 2

Assignment of Tokens: In stage 2 it will be displayed how much each group member contributed to the project. (**Please note: Before each round a display order will randomly be determined.** Thus, it is not possible to identify any group member by her position on the displayed list throughout different rounds.) By the assignment of tokens you can reduce the payoff of a group member or keep it unchanged.

In each round each participant receives additional 20 tokens in stage 2. You have to decide how many from the 20 tokens you are going to assign to other group members. The remaining tokens are kept by yourself. You can check the costs of your token assignment by pressing the button *Calculation of Tokens*.

- Each **negative token** that you assign to a group member **reduces her payoff by 3 tokens**.
- If you assign **0 tokens** to a group member her **payoff won't change**.

Calculation of your payoff in stage 2: Your payoff in stage 2 consists of two components:

- **tokens you kept =** 20 – sum of the tokens that you have assigned to the other group members
- **less the threefold number of negative tokens** you have received from other group members

Thus, **your payoff in Stage 2** amounts to:
20 – sum of the tokens that you assigned to other group members
– 3x (the number of tokens you received from other group members)

Calculation of your round payoff: Your round payoff is composed of

	Your payoff from Stage 1	20 – your contribution to the project + 1.6 x sum of the contributions of all group members / number of group members
+	Your payoff from Stage 2	20 – sum of the tokens that you have assigned to other group members – 3 x (the number of tokens you have received from other group members
=	Your round payoff	

Special case: a single group member: If it happens that you are the only member in your group you receive 20 tokens in Stage 1 and 20 tokens in Stage 2, i.e. your round payoff amounts to 40. You neither have to take any action on Stage 1 nor on Stage 2.

Information at the end of the round: At the end of the round you receive a detailed overview of the results obtained in all groups. For every group member you are informed about her: Contribution to the project, payoff from the Stage 1, assigned tokens (if possible), received tokens (if possible), payoff from Stage 2, round payoff.

History: Starting from the 2nd round, in the beginning of a new round you receive an overview of the average results (as above) of all previous rounds.

Total Payoff: The total payoff from the experiment is composed of the starting capital of 1000 tokens plus the sum of round payoffs from all 30 rounds. At the end of the experiment your total payoff will be converted into Euro with an exchange rate of 1 € per 100 tokens.

Please notice: Communication is not allowed during the whole experiment. If you have a question please raise your hand out of the cabin. All decisions are made anonymously, i.e. no other participant is informed about the identity of someone who made a certain decision. The payment is anonymous too, i.e. no participant learns what the payoff of another participant is.

We wish you success!