A Comparison of Punishment Rules in Repeated Public Good Games

AN EXPERIMENTAL STUDY

TORSTEN DECKER
Institute for Operations Research
Humboldt University Berlin
ANDREAS STIEHLER
Strategic Interaction Group
Max Planck Institute for Research into Economic Systems
MARTIN STROBEL
Department of Economics and International Institute of Infonomics
Maastricht University, The Netherlands

One individual and three collective punishment rules in a public good setting are analyzed. Evidence and explanations for differences between the rules concerning punishment intensity, contribution, profit levels, and justice are presented. Influences crucial to participants' support for a collective rule when the individual rule is the status quo are also investigated. Results show that besides profit differences, the degree of consent required by the collective rule is essential for the degree of support by the participants.

Keywords: *public good; free riding; cooperation; punishment; experiment*

A distinctive characteristic of public goods is free access to the common good irrespective of each person's contribution to the provision of the good. This characteristic creates the incentive to free ride (i.e., enjoy the benefits while staying away from the provision). Although free riding is rational at the individual level, it produces socially undesirable outcomes at the aggregate level. Economists and scholars in other social sciences directed a lot of research to solve this dilemma. One way to do so is by implementing additional institutions that create incentives to participate in the provision (e.g., a sanctioning system). Unfortunately, in a framework of rationality and selfishness, a new dilemma arises: no individual is willing to bear the costs of implementing

AUTHORS' NOTE: We thank Jörg Breitung, Simon Gächter, Werner Güth, and Michael Kvasnicka for helpful comments as well as Urs Fischbacher for z-Tree. Financial support by Deutsche Handelsbank AG Berlin and Deutsche Forschungsgemeinschaft (through SFB 373) is gratefully acknowledged. Additional material for this article (i.e., complete data set and experiment instructions) can be found on the Web site of the journal (http://www.yale.edu/unsy/jcr/jcrdata.htm).

JOURNAL OF CONFLICT RESOLUTION, Vol. 47 No. 6, December 2003 751-772
DOI: 10.1177/0022002703258795

or supporting the institution. This famous problem is known as the second-order dilemma, or second-order public good (Oliver 1980; Bates 1988).

From experiments and casual observations in reality, however, we have evidence that punishment systems are indeed effective tools to promote cooperation. As a result of field studies, Elinor Ostrom (1990) worked out seven design principles, which are crucial to the success or failure of a group or society facing a social dilemma. One of them is sanctioning.[1] Fehr and Gächter (2000) showed experimentally that costly punishment opportunities, despite their dilemma characteristics, are used by the participants and are able to raise and stabilize cooperation in a public good environment. In their experiment, they introduced a sanctioning system in which each participant had to decide individually whether and to what extent he or she wants to punish another person. The leverage, however, was rather high. The cost for imposing a fine on somebody else was only one-third of the fine.

If the costs for punishment are high, however, members of a group probably wish to decide together on the potential punishment of free riders sharing the associated costs. Ostrom (1990) reports cases in which the decision on punishment was found in a joint and organized way (e.g., in a vote). Another example, the so-called Growth and Stability Pact, contains a sanctioning system in which the members of the European Economic and Monetary Union (EMU) decide together on the punishment of countries endangering the stability of the Euro. A collective sanctioning system may be useful if the very structure of individual punishment bears strong incentives to abstain from punishment (Weesie and Franzen 1998) or is likely to escalate and cause heavy damages.

Two crucial questions may arise when designs of collective rules are considered:

1. Is it feasible that people accept the collective decision with all its implications, although they favor a lower or higher punishment?
2. Given a positive answer to question 1, why is it not feasible to directly enforce a full contribution at the public good stage?

With respect to the first question, we argue that an institution, such as a collective punishment rule, is set up to continually gain benefits from it. Opposing it in one case could put the whole institution at risk. So, once people have erected an institution, they will probably continue to obey it for their long-term benefit. Moreover, it is conceivable that the enforcement of the institution is backed by an authority, which was erected by the people in advance and is out of control in the current situation. Ostrom (1990) reports on successful cases where people designed sophisticated systems to ensure that punishment costs are shared.[2] Whether people are indeed willing to submit themselves to an institution, such as a collective punishment rule, is one of the questions we seek to answer in this study.

1. Although Ostrom's main field of study is common pool resources (CPR), she argues that "given the similarity between many CPR problems and the problems of providing small scale collective goods, the findings of this volume should contribute an understanding of . . . the capabilities of individuals to organize collective action related to providing local public goods" (Ostrom 1990, 27).

2. She reports, for example, on irrigation systems in Spain where guards and so-called "ditch riders" were paid for monitoring, reporting violations, and bringing charges against farmers (Ostrom 1990, 69-78).

The answer to the second question makes it important to distinguish between the initial public good in the first stage and the public good "punishment" in the second stage. A certain feature of collective punishment regulation is that punishment becomes a central issue and is, therefore, easy to enforce as discussed above. In contrast, the contribution decision remains in private responsibility and cannot be transferred to a central or an outside institution. In some cases, the contribution is even not or only ex post observable. Thus, a participant deciding by himself or herself on his or her contribution still faces incentives to free ride. Even a contract among the participants that fixes contributions is no credible commitment and, therefore, is likely to be violated.

For the reasons given above, we assume in this study that contributions to the public good in the first stage are not enforceable, whereas the decided punishment in the second stage is. Our main goal is to investigate collective punishment rules and confront them with an individual rule, which we consider an initial position. The following questions are central to our analysis:

- Are collective punishment rules able to bring about stronger cooperation and/or higher profits in a public good setting than an individual rule?
- To what extent do different collective rules perform differently from each other with respect to contribution, efficiency, and justice?
- Will participants agree to submit themselves to a collective rule, even if this means to give up some individual freedom? Which rule is preferred?

To find answers to the above questions, we designed an experiment in which participants repeatedly play a public good game. Between the rounds, they had the possibility to punish each other according to different rules. Every participant experienced a collective punishment rule and an individual punishment rule. After that, participants could bid for the right to choose the institution for the last five rounds.

The results of our experiment suggest basically that the more severe an institution, the higher the contribution to the public good but the lower the willingness of participants to accept this institution. In the second section of the study, we will first describe the experimental design. Then, in the third section, we establish some links to the literature that guided our expectations and report detailed results in the fourth section. We conclude in the final section with some general remarks.

THE EXPERIMENT

In our experiment, participants played a two-stage game for 20 periods in groups of four. The first stage consisted of a standard linear public good game (see Ledyard 1995). Each group member i received an initial endowment of 20 ECU (experimental currency units) and had to decide on which amount x_i to contribute to a public good ("pot"). The pot was multiplied by 1.6 and then equally distributed among the four group members regardless of their contribution. Thus, ECU 1 contributed to the pot had a marginal return of ECU 0.40 to each participant, whereas ECU 1 kept for oneself

had a marginal return of ECU 1 to oneself only. The preliminary payoff of participant i in group G therefore was the following:

$$\prod_i^{pre} = 20 - x_i + 0.4 \sum_{j \in G} x_j.$$

In stage 2, the punishment stage, every participant got to know the other group members' actions and could propose punishment amounts for each other. They had to be chosen from all even amounts between ECU 0 (meaning no fine) and ECU 20. Thus, every participant faced three punishment proposals (henceforth called *triplet*) from the other group members against himself or herself. The fine that was actually imposed was determined according to specific rules described below. All rules have in common that they imposed punishment costs of the same amount. Hence, the rules basically determine the size of the fine and who is going to bear the costs.

Individual rule (indi). The highest proposal out of the triplet was chosen and put into action. The person who proposed that highest amount had to bear the costs of the punishment. If there happened to be two or three equally high maximal proposals, the person who had to bear the costs for punishing was chosen randomly among them.

We consider this rule to be the fallback possibility, which is always feasible without the presence of any institution. Many situations have the characteristic that the action of only one person is sufficient to reach a certain goal, and no additional gain can be made when another person joins the action. Imagine that people who are waiting in a queue observe a person jumping the queue. Although many may feel upset, it is probably sufficient if one person rises up and scolds the "free rider." The others with similar intentions may be satisfied then and abstain from further action.

Collective rules (mini, medi, maxi). We distinguish between three collective rules—mini, medi, and maxi—depending on which of the three proposals in a triplet was put into action. With all three collective rules, we assume that a person is monotone in his or her wish to punish (i.e., a person who proposes a certain fine will also approve any lower fine).

- In the mini-rule, the lowest proposal was chosen. This rule can be considered a unanimity vote.
- In the medi-rule, the medium proposal was chosen. This corresponds to a majority vote.
- In the maxi-rule, the highest proposal was chosen. This rule can be interpreted as a sort of minority voting.

As with the individual rule, incurred costs were equal to the actual punishment amount. In contrast to the former, costs for punishing were now shared by the three remaining participants. Each of them had to pay one-third regardless of his or her own proposal.

The final profit $\prod_i^{final}$ for a participant i was calculated from his or her preliminary profit $\prod_i^{pre}$, the sanction s_i he or she received, and his or her costs for sanctioning others $c(s_j)$ as follows:

$$\prod_i^{final} = \prod_i^{pre} - s_i - \sum_{j \in (i, j \neq i} c(s_j).$$

At the end of each period, participants were informed about the actual punishment they received themselves, the costs they incurred by punishing others, and their final profit. In addition, they got information on whether and to what extent other persons were punished. They did not get to know the individual punishment proposals. Their final profits were credited to their respective accounts, and a new period began.

Having described the design of one period, we now turn to the structure of the treatments. To compare different punishment rules as well as to measure people's willingness to accept them, we let participants experience two different rules (in periods 1-15) and then auctioned off the right to determine the rule for a third phase (periods 16-20). An overview of the treatments is provided in Table 1.

Any collective rule was followed by the indi-rule, whereas the indi-rule was followed by the medi-rule (majority voting) as a prominent example of the collective rules. Moreover, complementing the indi-rule with a collective rule allows us to check whether the sequence of rules matters for the rule selection.

The selection procedure was designed not only to find out which rule a person favors but also to receive an indication of how strongly he or she prefers this rule. Each participant was asked to announce the rule he or she favors and his or her willingness to pay to be dictator in the selection procedure. Therefore, an extra amount of ECU 60 was credited to each participant's account. The proposal supported by the highest amount was then applied, and the amount was removed from the dictator's account. All other proposals were ignored, and the assigned bids were not subtracted. If there were two or more equally high maximal bids, the dictator was chosen randomly among them. Periods 16 to 20, in which the selected rule was actually applied, served as an incentive for thoughtful participation in the selection process.

We converted the experimental design into a computer program using the software z-Tree (Fischbacher 1999). To run eight observations of each treatment, we invited 128 students of economics and business administration to the experimental laboratory at Humboldt University. Eight people were invited at a time and randomly assigned to two groups. The participants did not know who of the remaining 7 people belonged to their group. Although they were informed that the experiment consisted of three phases, the instructions were not handed out until the beginning of the corresponding phase and the selection procedure, respectively. During the experiment, all interaction took place through computers. Each group member was given a number (ID), which was used as an identifier for the other group members. Within a phase, the IDs were constant, but after each phase, they were randomly permutated to diminish carryover effects from one to another phase (participants were informed about it). The decisions in the experiment were materially motivated. Each participant's payoff was converted into German Marks after the experiment and paid in cash. The conversion rate was ECU 30 = DM 1 (roughly Euro 0.50). The average earning of a participant was Euro 9.50.

TABLE 1
Overview of the Four Different Treatments

Phase	*Periods*	*Treatment 1*	*Treatment 2*	*Treatment 3*	*Treatment 4*
1	1 to 10	Indi-rule	Maxi-rule	Medi-rule	Mini-rule
2	11 to 15	Medi-rule	Indi-rule	Indi-rule	Indi-rule
	Before 16	Rule selection	Rule selection	Rule selection	Rule selection
3	16 to 20	Selected rule	Selected rule	Selected rule	Selected rule

LINKS TO THE LITERATURE

Before formulating concrete hypotheses about the likely outcomes of the experiment, we briefly summarize some theoretical approaches suitable for our purpose.

Assuming rationality and common knowledge of rationality, punishment will not occur in the last period because it is costly and no additional profit can be gained from it. Therefore, the punishment threat is not credible. This, in turn, leads to zero contribution in the last period. Using backward induction, we conclude that neither punishment nor cooperation will occur in any period. For the finite game, this is the only Nash equilibrium, which is both subgame perfect and trembling hand perfect.[3]

From experimental studies and real life, we know that the assumption of rational individuals is often violated. Many people are willing to engage in cooperation despite adverse incentives. Dawes and Thaler (1988), for example, found nonnegligible cooperation levels in one-shot public good decisions. Moreover, participants may become angry about unfair behavior and are ready to punish even if it is costly. Experimental evidence therefore was found not only in public good situations but also in a variety of other settings. Responders in ultimatum games reject unfair offers (Güth and Tietz 1990). In gift exchange games, Fehr, Gächter, and Kirchsteiger (1996) found behavior that displayed patterns of cooperation and reciprocity. Fehr and Gächter (2000) presented evidence for people's willingness to punish unfair behavior, as well as report high cooperation levels in public good settings.

A variety of approaches exist, in which economists try to explain these contradictory results by introducing elements of altruism, fairness, and/or reciprocity into people's considerations, mostly by incorporating additional terms into their preferences.

Rabin (1993) emphasizes reciprocity in a person's behavior. The core of his approach is that "people like to help those who are helping them, and to hurt those who are hurting them" (p. 1281). Levine (1998) presents a theory of altruism and spitefulness in which people's utilities depend on their own and their fellow players' payoffs. The degree to which a person takes other people's payoffs into account is specific to that person and varies among the population. The former two approaches model both

3. For indi, no other subgame-perfect equilibrium exists. However, for mini, medi, and maxi, the voting procedure causes numerous subgame-perfect equilibria in weakly dominated strategies. This is because individuals do not have a decisive vote in many of these situations.

altruism and reciprocity in participants' utility functions, whereas Andreoni and Miller (2002) focus on altruism alone.[4]

Although cooperation in a public good game is in line with all three approaches, punishment can be explained only if reciprocity or spitefulness is part of the model, as is the case in Rabin's (1993) and Levine's (1998) approaches.

Recently, two approaches pioneered by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) have received attention in the public good context and beyond. They have in common the introduction of an inclination to equity of payoffs into people's motivation. This means that punishment could be used to reduce inequity after the public good stage. The application of inequality aversion is, however, questionable in the case of the indi-rule. It would only work out if the free riders were to engage in punishment, which is not the case (as we will see later).

Pruitt and Kimmel (1977) address the topic from a psychological perspective and emphasize the role of trust. Their goal/expectation theory suggests that most people recognize the need for and share the goal of establishing mutual cooperation. To achieve cooperation, the common goal "must be accompanied by an expectation that the other will cooperate" (Pruitt and Kimmel 1977, 375).

Yamagishi (1986) has extended the goal/expectation theory to the structural goal/expectation theory. He argues that people are conditionally willing to cooperate in the sense of Pruitt and Kimmel (1977). The main obstacle to cooperation is a lack of mutual trust. However, the opportunity to cooperate in a second-order public good (e.g., a sanctioning system) will be used to establish trust necessary for durable cooperation. Yamagishi provides experimental evidence that people showing a lack of mutual trust display uncooperative behavior (relative to groups of rather trusting people)[5] in the absence of a sanctioning system. The same people make relatively heavy use of punishment opportunities and achieve higher cooperation levels than their trusting counterparts when a sanctioning system is provided. Further evidence for conditional cooperation is given by Fischbacher, Gächter, and Fehr (2001).

We consider as the essence of all approaches that there are people who are conditionally willing to contribute to a common goal, despite adverse incentives, and that there are people ready to punish if they feel unfairly treated.

HYPOTHESES AND RESULTS

In most of our analysis, we concentrate on phase 1 (periods 1-10) and the selection procedure (choice of the rule). The data from phase 2 are difficult to interpret because it is not clear to which extent there are carryovers from one rule to another. Phase 3 was primarily introduced to give the participants incentives to take the auction seriously.

Figure 1 gives a general idea of the behavior in the first phase. Obviously, punishment occurred, but we have to be careful with conclusions about the effect of punishment because we did not run a treatment without punishment. The general contribu-

4. In addition, they allow for nonlinear preferences.

5. On the basis of a preexperimental questionnaire, Yamagishi (1986) subdivided participants into groups of so-called high and low trustees.

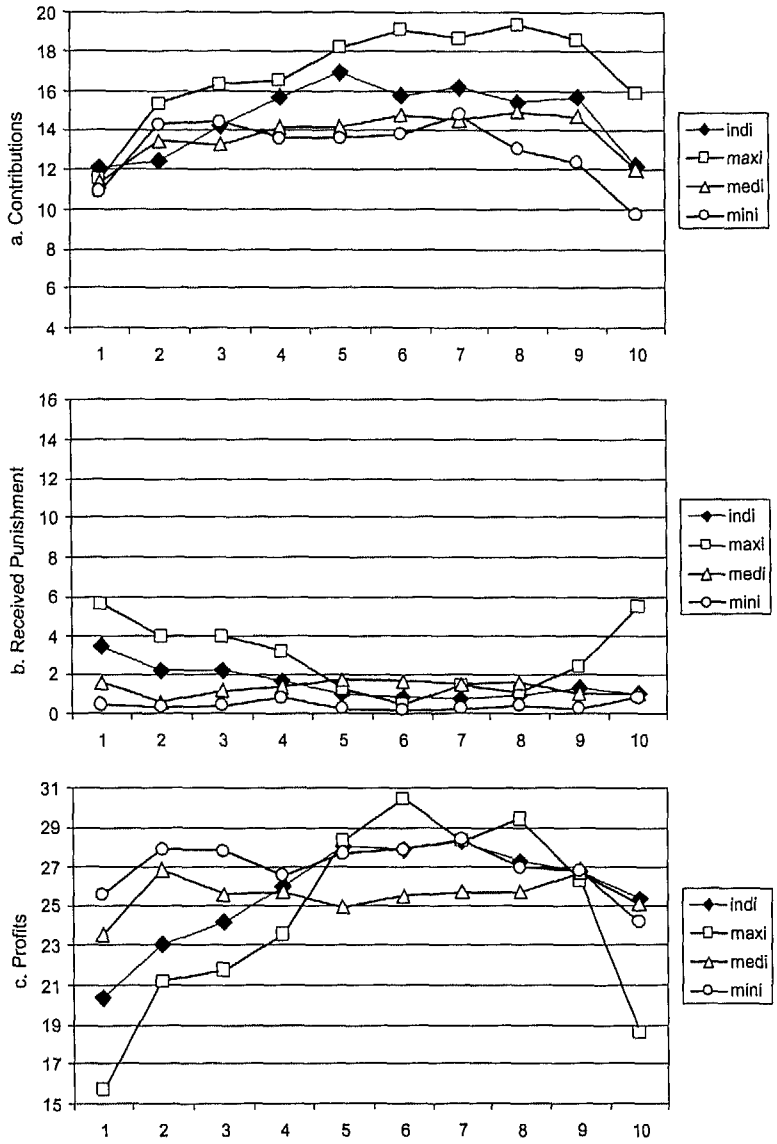

Figure 1: Average Contribution (a), Average Received Punishment (b), and Average Profit (c) over the First 10 Periods (Abscissa)

tion behavior, however, is hump shaped—quite similar to that found by Masclet et al. (forthcoming) and Noussair and Tucker (2002) in treatments with punishment. We do not find the typical strong deterioration of cooperation over time as described in the lit-

erature (e.g., Fehr and Gächter 2000; Ledyard 1995). We therefore assume that punishment helped to establish and stabilize cooperation.

There are, however, differences between the rules. Although in terms of contributions, they display visibly different courses with a clearly superior maxi-rule, the evolution of profits paints a rather vague picture. Interestingly, the initial average contribution is similar for all different rules. Apparently, people did not anticipate the rules' different peculiarities but rather react to them in the course of the repeated game.

In the following subsections, we analyze the data in more detail. Thereby, we use the relation ">" in a generic sense denoting higher and better, respectively. Most regressions rely on averages of single groups over the first 10 periods rather than on individual data because this allows us to base our tests on independent observations.

PUNISHMENT

Number and Extent of Punishment Proposals

That punishment is able to substantially change people's behavior in public good settings necessitates closer investigation of the motives behind punishment. On one hand, it can be the anger about another person's contribution. In this case, we speak of emotional punishment. On the other hand, one might wish to punish to induce higher future contributions by the punished person and, as a result, higher cooperation within the group. We refer to this as strategic punishment.

Hypothesis 1: Both strategic and emotional punishment exist and are nonnegligible.

If the strategic motive existed, punishment activity should weaken toward later periods, as the benefit from establishing future cooperation decreases. To check this, we calculated the total amounts of punishment proposals (i.e., the sum over all proposals for a given group and period) and compared the first five periods with the last five periods. Apart from the influence described above, we recognize the contribution's mean and standard deviation as possible determinants for punishment activity. Furthermore, it is reasonable that a particular group is more or less active in punishment throughout all periods. Thus, we performed a fixed-effect regression that explains the total amount of punishment proposals in a period by the mean and standard deviation of contributions of that period, as well as a dummy variable with values 0 or 1 for the first or last five periods, respectively. Moreover, the regression was done separately for the indi-rule, the collective rule, and all rules together. Eventually, we chose to report (White's) robust standard deviations to account for possible dependencies within the 10 observations from one group. The results are reported in Table 2.

As can be seen, the dummy is only significant for the indi-rule. There is no sign of weakening punishment behavior under the collective rules. One possible explanation is that other (emotional) motives grow as the strategic motive loses strength. We conjecture that in the collective rules, increasing emotional punishment neutralizes lower strategic punishment in the last periods, whereas emotional punishment is dampened

TABLE 2
Fixed-Effect Regression of the Total Amount of Punishment Proposals within a Group and a Period

	All Rules		*Indi Only*		*Collective Rules*	
Constant	21.05***	(6.88)	5.84	(7.60)	27.27***	(8.29)
Contribution mean	−0.71*	(0.38)	0.01	(0.49)	−1.06**	(0.44)
Contribution standard deviation	4.00***	(0.48)	2.76***	(0.95)	4.41***	(0.54)
Dummy (6 to 10)	−1.78	(1.97)	−8.45***	(3.94)	0.90	(2.01)
Adjusted R^2	0.69		0.56		0.70	

NOTE: Numbers in parentheses are (White's) robust standard errors.
*Significant at the 10% level. **Significant at the 5% level. ***Significant at the 1% level.

by the second-order dilemma in the indi-rule. If one person repeatedly bore the costs for punishment alone, he or she may feel exploited and understandably abstains from future engagement in punishment. However, this explanation cannot be verified in this study.

The existence of emotional punishment can be checked easily. If we observe punishment activity in the very last period of a phase, we have evidence for emotional punishment because strategic punishment is futile. The data strongly support emotional punishment. In 24 out of 32 groups (indi: 5/8, maxi: 5/8, medi: 7/8, mini: 7/8), punishment proposals occurred in period 10. One may object that there is still a reason for strategic punishment because people may speculate that they will encounter each other in a similar game in the second or third phase of the experiment. Thus, we also looked at the last period of the third phase, from which the participants knew that it was definitely the last period of the entire experiment. Again we found punishment activity across all rules (indi: 9/12, maxi: 1/2, medi: 9/12, mini: 5/6).

Participants anticipate, however, punishment to be strategic rather than emotional and decrease their contribution at the end of the phase (see Figure 1a). To the contrary, at least some individuals seem to be more emotional than anticipated, which can be seen by the punishment curve of the maxi-treatment (see Figure 1b).

The punishment curve for the indi-rule resembles best the characteristics of a standard public good dilemma. Carried out punishment starts rather high and decreases toward the end. It carries features of a so-called volunteer's dilemma.[6] The positive effects of punishment on cooperation can be enjoyed by everyone, whereas the costs are borne by that participant who proposed the highest amount. This dilemma in its various forms has been described and analyzed by Diekmann (1993), Weesie (1993, 1994), and Weesie and Franzen (1998). Most relevant is the work by Weesie and Franzen, in which they show both theoretically and experimentally that the probability of the public good (here: punishment) being provided increases under the condition of

6. Strictly speaking, we assume the existence of an asymmetric volunteer's timing dilemma for individual punishment. Following Weesie (1993), the person with the strongest need for the provision of the good (punishment) is most likely to go ahead and provide and pay for it, whereas the others who are able to observe this will then abstain from provision. Assuming that the person with the highest need for punishment will propose the highest punishment amount, our design of the indi-rule reflects both the selection mechanism and the consequences of a volunteer's timing dilemma.

cost sharing. In our setting, this translates into more and heavier punishment proposals under the collective rules. One may also take on a simpler view. Punishment is, for an individual, less costly under a collective rule than under the indi-rule; strictly speaking, the costs reduce to one-third. Hence, we state the following:

Hypothesis 2: Punishment activity (measured by proposals) is weaker under the indi-rule than under the collective rules.

To analyze this hypothesis, we confront the total extent of punishment proposals in a period averaged over 10 periods in indi-rule with the same measure from the collective rules. To control for the influence of the contributions' mean and standard deviation, also averaged over 10 periods, we applied a linear regression including a dummy variable with value 0 for observations from indi-treatments and value 1 for those from collective treatments. Table 3 (column: punishment proposed) shows that our hypothesis is confirmed, as the significant dummy indicates.

Enforcement of Punishment Proposals

Concerning the rules, the punishment, which was enforced depending on the proposals, is straightforward. In both the maxi- and the indi-rules, the most severe proposal is enforced; in the medi-rule, the median proposal is enforced; and in the mini-rule, the least severe proposal is put into action. Given all other things equal, punishments carried out in the four rules should be related in the following way: indi/maxi > medi > mini. However, carried-out punishment differs only if there are different proposals against a person. Moreover, the sizes of the punishment relations crucially depend on the characteristics (spread, skewness) of the proposal distribution within a triplet.

From our data, we calculated descriptive measures for this distribution. Taking into account all triplets in which at least one proposal was greater than zero, we found that the average distance between the maximum and the median proposal was 5.80, whereas the average distance between the median and the minimum proposal was 1.77. Furthermore, we calculated the average ratios of the min-med span to the (whole) max-min span over all triplets of a four-person group. The ratios have a mean of .251 and a standard deviation of .125. There was no group with an average ratio greater than .5. Thus, apart from the variation within the triplets, we find that the distribution is highly skewed. The medium proposal is much closer to the minimum than to the maximum proposal.

Carried-Out Punishment

The number, extent, and the enforcement of punishment proposals essentially influence the occurrence of punishment. Following hypothesis 2, the indi-rule generates less punishment activity than the collective rules. Because, under the aspect of enforcement, the indi-rule is equivalent to the maxi-rule, we expect the following:

TABLE 3
Ordinary Least Squares Regressions of the Averaged Total Amount of Punishment Proposals and of Carried-Out Punishment

	Punishment Proposed		*Punishment Carried Out*	
Constant	−3.05	(17.11)	1.20	(1.62)
Averaged contribution	−0.23	(0.87)	0.02	(0.08)
Averaged standard deviation of contribution	5.66**	(1.91)	0.56**	(0.17)
Dummy for collective rules	14.82**	(5.56)		
Dummy for indi			−1.56**	(0.60)
Dummy for medi			−2.00***	(0.62)
Dummy for mini			−3.24***	(0.64)
Adjusted R^2	0.42		0.49	

NOTE: Numbers in parentheses are standard errors.
**Significant at the 5% level. ***Significant at the 1% level.

Hypothesis 3: In the maxi-rule, punishment occurs more extensively than in the indi-rule.

Following our reasoning on enforcement, and because we have no reason to expect different punishment behavior across the collective rules, we expect the following:

Hypothesis 4: Concerning the extent of punishment carried out, the following relation holds: maxi > medi > mini.

In a linear regression, we explained the carried-out punishment averaged over 10 periods. We introduced dummy variables for three of the four rules (indi, medi, and mini) and controlled for the contribution's mean and standard deviation (also averaged over 10 periods). The intercept reflects the influence of the maxi-rule, whereas the dummies reflect the additional influences of either rule relative to the maxi-rule.

The results are reported in Table 3 (column: punishment carried out). The values for the dummies show that carried-out punishment is significantly higher in the maxi-rule than in the other rules. Applying an F test, based on the restriction that the mini- and medi-dummies are equal, we also find that the medi-rule generates significantly higher punishment than the mini-rule ($p = .044$, F test). Thus, all relations proposed in hypotheses 3 and 4 turn out to be statistically significant.

For further considerations, we will refer to a rule that is superior to another rule in terms of the carried-out punishment as the more severe rule.

CONTRIBUTION AND PUNISHMENT

In a public good game without punishment opportunities, an important determinant of a participant's contribution decision are the past contributions of the other group members and his or her expectation about their future contributions. This is in line not only with Yamagishi's (1986) approach but has recently been supported also experimentally by Fischbacher, Gächter, and Fehr (2001). With our setup, explicit punish-

ment and expectations about it come into play and also influence the contribution decision.

It is intuitively appealing that relatively small contributions are perceived as unfair, whereas higher contributions are seen as fair. In this context, Fehr and Gächter (2000) found that mainly negative deviations from average contribution levels were punished.[7] Consequently, we expect to find similar behavior in our experiment.

Hypothesis 5: Lower contributions (relative to the group average) receive heavier punishment proposals than higher contributions.

The correlation between a person's deviation from the average group contribution and the amount of punishment proposals he or she received in the same period was calculated as −.580 from all available data, supporting the assumption. The relation is stable across all rules. However, this number is calculated from the total of 1,280 observations that are partly dependent. To check for statistical significance, we calculated a correlation coefficient for each group and applied a t test on the 31 resulting independent measures.[8] The correlation coefficients have a mean of −.660, with an estimated standard deviation of .037. Based on a sample of size 31, one can reject the hypothesis of a nonnegative correlation at any reasonable significance level.

In a similar way, we consider one's own contribution as a major determinant of how fair or unfair one regards other persons' contributions (i.e., a person with a very high contribution may feel a stronger need to educate a free rider than a person with an average or even lower contribution).

Hypothesis 6: Participants with above-average contributions propose heavier punishment than others.

As with hypothesis 5, we calculated the correlation between a person's deviation from the group average and the total amount of punishment he or she proposed. The Pearson correlation coefficient is .257 if all observations are included. If one separates indi from the collective treatments, the coefficients are .256 and .271, respectively, indicating a stable relationship. The 31 correlation coefficients for each single group have a mean of .278, with an estimated standard deviation of .050. Again, we can reject the hypothesis of a nonpositive correlation by a t test at any reasonable significance level.

CONTRIBUTION

Having an idea of who punishes whom, we turn to the effects that punishment causes. A person's natural reaction to a received punishment should be to change his or her behavior. Following hypothesis 5, this means, above all, that this individual will

7. They also found that some low contributors engaged in punishment against high contributors, but the extent was small.

8. In 1 out of 32 groups, no punishment proposal occurred in the first phase, making it impossible to calculate a correlation coefficient.

raise his or her contribution.[9] All other contributions unchanged, this would lead to a higher cooperation level in the group. Keeping one's own contribution constant in an environment of rising cooperation, however, also means facing a higher risk of being punished. Therefore, observed punishment may induce other group members to raise their contributions as well.

The fear of being punished is not the only motive to raise one's contribution. As explained above, people are willing to contribute more in expectation of higher cooperation by other group members, even in the absence of punishment opportunities. The credible threat of punishment should be sufficient to trigger the same dynamics. In contrast, we expect lower cooperation levels if punishment fails to be credible.

From hypotheses 3 and 4, we saw that the occurrence (and thus the credibility) of punishment varies within the rules. Therefore, contribution should relate accordingly.

Hypothesis 7: Concerning the performance of contributions, the following relation holds: maxi > indi and maxi > medi > mini.

Table 4 shows that the contribution levels, averaged over all participants and periods under a certain rule, relate as proposed, although there are large dispersions within each group. The relations become clearer when only contributions of the last five periods of a treatment are averaged. The results of pairwise comparisons using a Mann-Whitney *U* test are reported in Table 5. The relations between maxi and medi as well as maxi and mini are significant at the 5% level. With a 10% level of significance, one might also reject that maxi and indi generate equal contribution behavior in the last five periods. The nonsignificance of the medi-mini comparison may be explained by the high skewness of the distribution of punishment proposals. The punishment amount resulting from a typical triplet would rise only modestly if the medi-rule were applied instead of the mini-rule. It would rise substantially if the maxi rule were applied.

Finally, it is worth mentioning that applying the same test to initial contribution levels under the different rules did not lead to any significant difference. As noted before, this shows that people's anticipation about the rule was limited.

EFFICIENCY

On the one hand, punishment seemingly raises contribution; on the other hand, it reduces income. Obviously, the question arises about which of the rules performs best in terms of efficiency.[10] Does the benefit from more cooperation outweigh the costs of the punishment needed to establish that additional cooperation? We abstain from generating a hypothesis to this question because it requires quantitative assumptions on the effects of punishment, which are difficult to make.

9. Hypothesis 5 contains a link between a person's contribution and the punishment proposals he or she receives, whereas, for a change in a person's behavior, actual punishment was needed. However, more or higher punishment proposals generally led to higher (at least not smaller) carried-out punishment within any rule.

10. By efficiency, we understand the group's total profit after all sanction costs have been covered (e.g., the final profit as defined in the second section summed over all four members).

TABLE 4
Descriptive Measures on Contribution and Profit Levels under Different Rules

Periods	*Mean 1-10*	*Mean 6-10*	*Minimum Group Mean 1-10*	*Maximum Group Mean 1-10*	*Range of Group Mean 1-10*
Contributions					
Indi	14.65	15.04	10.98	20.00	9.02
Maxi	16.94	18.31	14.13	19.13	5.00
Medi	13.73	14.19	10.43	17.75	7.32
Mini	13.06	12.74	6.58	20.00	13.42
Profits					
Indi	25.65	27.05	20.75	32.00	11.25
Maxi	24.35	26.61	16.68	31.02	14.34
Medi	25.54	25.74	21.78	28.25	6.47
Mini	26.96	26.84	23.45	31.28	7.83

NOTE: All measures are based on group averages under the specific rule.

Interestingly, the relations between the rules in terms of efficiency are opposite to the relations predicted and found for punishment and contributions: maxi < indi and maxi < medi < mini (see Table 4). However, these relations do not withstand a statistical test ($p > .1$, Kruskall-Wallis test), partly due to the high dispersion in profit levels among the observations.

The efficiency relations (see Figure 1c) are not stable throughout the course because of the striking pattern exhibited by the maxi-rule. Although it performs best in some intermediate periods, it lies far behind in the beginning and in the last period. Obviously, the heavy use of punishment to establish cooperation was both successful and expensive. In the intermediate periods, cooperation stabilized on a high level accompanied by few punishments, allowing the participants to earn generous profits. However, when the end of the treatment approached, contribution levels decayed generally. This effect could not be stopped, even though heavy punishment had to be expected. The immense extent of carried-out emotional punishment in the last period left the participants with relatively poor profits.

The medi- and mini-rule, in contrast, exhibit a rather stable pattern in terms of efficiency, whereas the indi-rule's evolution resembles the maxi-rule in the beginning but not in the last periods.

JUSTICE

Besides efficiency, people turn their attention to justice if they judge allocations or have to choose between different alternatives.[11] We, therefore, compare the four pun-

11. One may reason that justice is of importance for a long-term stabilization of cooperation. Given repeated free-rider attacks, repeated punishment would be needed to stabilize cooperation. If, however, a rule generates permanently and strongly unjust results after punishment, the willingness to punish and cooperate may decay.

TABLE 5
Test Statistics and Significance Levels (Mann-Whitney *U* Tests) for Contributions on All Relevant Pairs of Rules

Variable: Contributions	*Mann-Whitney* U		*Significance (One-Tailed)*	
Maxi/indi	20.0	(17.5)	.117	(.065)
Maxi/medi	11.0	(8.0)	.014	(.005)
Medi/mini	28.0	(23.5)	.361	(.191)
Maxi/mini	12.5	(11.0)	.019	(.014)

NOTE: Numbers in parentheses are the respective values using the averages of the last five periods. All tests are based on independent observations (group averages).

ishment rules with respect to justice, too. In our view, the highest degree of justice would be reached if the most generous contributors earned the highest profits. In addition, we would also regard as just a situation without any dispersion in both contributions and profits.[12]

In a linear public good game without punishment opportunities, different profits can only be attributed to different privately kept amounts. Therefore, the correlation between profits and contributions is perfectly negative. Moreover, the variance in contributions translates completely into the variance of profits. According to our definition, this is a very unfair situation.

We recognize three channels through which punishment is able to improve justice:

1. through weakening or even reversing the negative correlation between contributions and profits, given a negative correlation rate;
2. by reducing the variance in profits; and
3. by reducing variation in contributions themselves.

Let us first turn to channels 1 and 2, which can be considered the direct influences of punishment. Because, following hypotheses 5 and 6, mainly below-average contributions are punished from above-average contributors, the relative benefit from free riding may be reduced or even reversed. This is for the better of justice. The costs of punishment, however, lessen this improvement. Thus, it is of particular importance who bears the costs for punishment in the indi-rule. In most cases, participants who already have relatively small preliminary profits after the public good stage suffer a further deterioration in payoffs due to punishment costs.

For the collective rules, the results of the hypotheses 3 and 4 suggest that the more severe a rule is, the more likely free riders were punished and thus the greater the improvements in justice will be.

We come to the same conclusion when turning to channel 3, the indirect effect of punishment. The more severe a rule is, the fewer deviations from a social norm (e.g., the average contribution) are tolerated. And the less tolerant a rule is, the more it is able to generate homogeneity in contributions.

12. Alternatively, one could simply regard the variance in profits within a group as a measure of justice (i.e., low variance means higher justice than high variance). This definition rests on the fact that in each period, every group member starts with the same endowment.

Taking all three channels into account, we arrive at the following:

Hypothesis 8: In terms of justice, the following relations hold: maxi > indi and maxi > medi > mini.

To test our hypothesis, we calculated a measure for each of the justice channels: correlation coefficients between profits and contributions, variances of profits, and variances of contributions. The measures were calculated for each group and each period from 1 to 10 and then averaged over periods.[13] Eventually, we tested whether the resulting measures differ.

Regarding the first relation between the maxi- and indi-rules, we find only a weakly significant difference for profit variances (see Table 6). The other two measures do not significantly differ. When comparing the three collective rules, the variances of profits as well as correlations between profits and contributions turn out to be significantly different ($p = .013$ and $p = .011$, Kruskall-Wallis test), whereas the variances in contribution do not differ significantly ($p = .105$, Kruskall-Wallis test). In pairwise comparisons (see Table 6), we find (weakly) significant differences with respect to the correlation of profits and contributions as well as to profit variances. The difference in contribution variances is only significant between maxi and mini. After all, we consider the second part of the hypothesis (i.e., the relations among the collective rules) as supported.

INSTITUTIONAL CHOICE

In previous subsections, we compared collective rules in terms of normative criteria (e.g., efficiency and justice). Now we turn to the questions about whether and to what extent people are willing to support a collective rule when their alternative is the individual rule.

Before we turn to identify the determinants of the decisions made by the participants, we present some aggregate results. The approval of the collective rules follows the pattern mini > medi > maxi. When the mini-rule was the alternative to indi, 22 out of 32 (68.8%) chose it. For medi and maxi, the rates of approval were 42/64 (65.6%)[14] and 16/32 (50%).

The bids submitted by the participants to support their votes averaged ECU 14.18, which is nearly 25% of the amount they received for this purpose. The most frequent bids were ECU 0, 1, 10, and 20; the maximum amount of ECU 60 occurred four times.

With respect to the decision determinants, we expect individuals to weigh their costs and benefits. One variable that is easily observable and comparable is personal profit. We expect the following:

13. The variances in profits and contributions are disputable, particularly if the correlation between profits and contributions is positive. However, because only 1 out of 31 groups shows a resulting positive measure, we neglect this problem.

14. Because the order in which the rules are applied does not significantly influence the rate of approval (two-sample test for equality of proportions with continuity correction, $p = .792$), we pool the data of indi/medi (22/32) and medi/indi (20/32).

TABLE 6
Significance Levels (Mann-Whitney *U* Tests) of the Differences between Different Rules Regarding Correlations between Profits and Contributions and the Variances of Profits and Contributions

Rules	*Correlations between Profits and Contributions: One-Tailed Significance*		*Variance of Profits: One-Tailed Significance*		*Variance of Contributions: One-Tailed Significance*	
Maxi/indi	.232	(21.5)	.081	(18.0)	.287	(26.0)
Maxi/medi	.065	(17.5)	.065	(17.0)	.221	(24.0)
Medi/mini	.053	(16.5)	.053	(16.0)	.164	(22.0)
Maxi/mini	.002	(5.5)	.003	(6.0)	.018	(12.5)

NOTE: Numbers in parentheses are Mann-Whitney *U* statistics. All tests are based on independent observations (group averages).

Hypothesis 9: Differences in personal profits are of importance; a person tends to choose the rule under which he or she can expect higher profits.

Besides profits, we consider personal freedom essential to a person's evaluation of punishment rules. Any collective rule constitutes a restriction to an individual's freedom of action that can be regarded as a cost in a nonpecuniary sense. The restriction occurs in two possible ways:

1. A participant has to bear the costs for a carried-out punishment (one-third of the amount) even if he or she proposed a lower punishment or no punishment at all.
2. A participant has to be satisfied with the lower carried-out punishment even though he or she proposed a higher amount.

Crucial to the restriction of a person's freedom by a rule is the degree of consent required by the rule. The mini-rule requires the agreement of all three proposers to fix the punishment amount. Consequently, in the mini-rule, the first kind of restriction can be ignored, whereas the second kind may occur frequently. On the other hand, in the maxi-rule, the most severe person decides on the punishment amount. Thus, the second kind of restriction is negligible, whereas the first kind may play an important role. The medi-rule lies between the two extremes.

Hypothesis 10: Both kinds of restriction weaken the approval of collective rules.

If the choice is between indi and maxi, a person who is very keen to punish others will prefer the maxi-rule because this rule would enable this individual to put through a high punishment while he or she can share the costs with the other group members. On the other hand, a person who wishes little or no punishment and is not willing to share costs will prefer the indi-rule. A similar reasoning applies when people are choosing between the mini-rule and the indi-rule. A person very willing to punish may feel too restrained under this collective rule and vote for the indi-rule.

There are some other possible influences on a person's attitude toward a rule—for example, the extent to which a person was the target of past punishments, the extent of undue punishments to one person, and the degree of justice of a rule. However, we consider these influences to play a minor role because the first is at least partially included in the profits, and the second rarely happens. Justice was difficult to observe because the participants had to keep track of all the payments of the others.[15]

According to the distinction concerning the restriction to a person's freedom, we calculated two measures:[16]

1. drag-in: absolute value of the sum of negative differences between a person's proposed punishment and the carried-out punishment, averaged over 10/5 periods;
2. curb: sum of positive differences between a person's proposed punishment and the carried-out punishment, averaged over 10/5 periods.

To measure the degree of approval for a particular collective rule, we put a sign to each bid, depending on whether the amount was dedicated to support a vote for the indi-rule (–) or a collective rule (+). On these measures, we used a linear regression to estimate the influence of drag-in, curb, and profit differences on the approval of a rule. We use aggregated measures to base the regression on independent observations. For the dependent variable, we calculated the total approval for the collective rule in a group as the sum of the signed bids over the four members of the group (i.e., bids for the collective rule were added, and bids for the indi-rule were subtracted). For independent variables, we used the averaged profit differences and summed the variables' "curb" and "drag-in" over the four group members. The results confirm our hypotheses (see Table 7). Profit differences as well as drag-in contribute most strongly to the explanation of participants' approval for the collective rule, but curb also has an influence. Given our prior assumptions on the sign of these influences, a one-tailed test is appropriate. Based on this, the variables are significant at levels of 5% and 10%, respectively.

CONCLUSION

Like Elinor Ostrom, we are aware that

> whether or not an equilibrium would be an improvement for the individuals involved (or for others who are in turn affected by these individuals) will depend on the particular structures of the institutions. . . . Further, the particular structure of the physical environment involved also will have a major impact on the structure of the game and its results. Thus, a set of rules used in one physical environment may have vastly different consequences if used in a different physical environment. (Ostrom 1990, 22)

15. Participants received information only about their own accumulated payoff.

16. The measures were, of course, calculated from the first phase. For groups starting with the indi-rule, the second phase was accounted.

TABLE 7
Ordinary Least Squares Regression of the Total Approval of the Collective Rule in a Group

Variable	*Coefficient*	*Significance (t Test, One-Tailed)*
Constant	30.89 (13.04)	.013
Profit difference	2.98 (1.55)	.033
Curb	−0.72 (0.52)	.087
Drag-in	−1.50 (0.86)	.046
Adjusted R^2	0.18	

NOTE: Numbers in parentheses are standard errors.

We consider our experimental design as one of many possible structures to be found in practice. Changing certain characteristics of the design (e.g., the number of players, the number of periods, the cost or information structure, or the degree of anonymity in the voting process) may lead to changes of particular results. Therefore, we cannot generalize each particular finding, but we want to make some general remarks about the insights we gained from this study.

Our experiment was designed to learn about people's behavior under different punishment rules and to answer several questions. First we asked, "Are collective punishment rules able to bring about stronger cooperation and/or higher profits in a public good setting than an individual rule, and to what extent do different collective rules perform differently from each other?"

We find that punishment rules differ from each other due to differences in the occurrence of punishment proposals and their enforcement. Collective rules generally induce heavier punishment activity. If we suppose that a higher probability of being punished leads to a higher contribution, it is no surprise that the maxi-rule performs best in terms of contribution. But when comparing punishment rules, it is not enough to consider contribution behavior only. One has to take punishment costs into account as well. Although efficiency does not differ significantly in our setting, the relations are reversed. We conclude the following:

Remark 1: More punishment, although it is able to enforce more cooperation, does not necessarily generate higher profits.

Our data verify that emotional punishment exists and that it has a major impact on efficiency. In the maxi-rule, although it performed best in terms of contribution, emotional punishment was not sufficiently suppressed and consequently damaged the gains from cooperation. We conclude the following:

Remark 2: When designing punishment rules, the ability to suppress exaggerated emotional punishment is at least as important as providing a sufficient punishment threat.

Turning to justice, we argued that the extent to which a rule is able to generate and enforce punishment is responsible for the improvement of justice, provided that

contributions below average are punished more heavily than those above average. This argument was confirmed by our empirical findings, in which the maxi-rule turned out to be superior to all other rules.

Altogether, considering efficiency and justice, collective rules do not perform worse than the individual rule. Moreover, depending on their design, they are able to surpass individual punishment in different aspects. The following question remains: will participants agree to submit themselves to a collective rule, even if this means giving up some individual freedom?

We investigated influences on participants' approval or rejection of collective rules and conclude the following:

> *Remark 3:* As long as collective rules allow for higher profits, people support them. On the other hand, restrictions to the personal freedom weaken the support for collective rules or even cause support for the individual rule.

Mainly, participants who have to bear the costs of a carried-out punishment higher than the one they have proposed are likely to oppose the collective rule. This finding suggests that the higher the required degree of consent by the collective rule, the more likely it is to be supported by the participants. This may explain the high popularity of unanimity voting and veto rights in institutions of the United Nations and the European Union. The other kind of restriction—that participants have to accept a certain punishment even if they proposed a higher amount—turned out to be less influential.

Finally, we turn to the last question: is there an optimal punishment rule?

To design an optimal punishment rule, one has to solve several trade-offs. The punishment rule has to provide sufficiently high probability of being punished, on one hand, but it should be able to suppress exaggerated emotional punishment, on the other hand. To what extent punishment is sufficient or exaggerated can only be answered case by case, taking into account the environment and characteristics of the individuals involved. A particular result of our experiment is that a rule that is severe enough to establish cooperation in one group may fail to provide sufficient punishment in another group. Moreover, a rule that provides sufficient but not exaggerated punishment in one group may allow for too excessive punishment in another group. Because both people and groups differ in their initial inclination to cooperate and punish, and because environments are special and unique, we arrive at the final remark—that a generally optimal punishment rule may not exist.

REFERENCES

Andreoni, J., and J. Miller. 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* 70:737-53.

Bates, R. H. 1988. Contra contractarianism: Some reflections on the new institutionalism. *Politics and Society* 6:169-217.

Bolton, G. E., and A. Ockenfels. 2000. ERC—A theory of equity, reciprocity, and competition. *American Economic Review* 90:166-93.

Dawes, R. M., and R. Thaler. 1988. Cooperation. *Journal of Economic Perspectives* 2:187-97.

Diekmann, A. 1993. Cooperation in an asymmetric volunteer's dilemma game: Theory and experimental evidence. *International Journal of Game Theory* 22:75-85.

Fehr, E., and S. Gächter. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90:980-94.

Fehr, E., S. Gächter, and G. Kirchsteiger. 1996. Reciprocity as a contract enforcement device. *Econometrica* 65:833-60.

Fehr, E., and K. M. Schmidt. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114:817-68.

Fischbacher, U. 1999. z-Tree: Zurich toolbox for readymade economic experiments. Working Paper No. 21, University of Zürich.

Fischbacher, U., S. Gächter, and E. Fehr. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71 (3): 397-404.

Güth, W., and R. Tietz. 1990. Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology* 11:417-49.

Ledyard, J. O. 1995. Public goods: A survey of experimental research. In *Handbook of experimental economics*, edited by J. Kagel and A. Roth. Princeton, NJ: Princeton University Press.

Levine, D. K. 1998. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1:593-622.

Masclet, D., C. Noussair, S. Tucker, and M. C. Villeval. Forthcoming. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review.*

Noussair, C., and S. Tucker. 2002. Combining monetary and social sanctions to promote cooperation. Working paper, Emory University, Atlanta.

Oliver, P. 1980. Rewards and punishments as selective incentives for collective action: Theoretical investigations. *American Journal of Sociology* 85:356-75.

Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action.* Cambridge, UK: Cambridge University Press.

Pruitt, D. G., and M. J. Kimmel. 1977. Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *American Review of Psychology* 28:362-96.

Rabin, M. 1993. Incorporating fairness into game theory and economics. *American Economic Review* 83:1281-302.

Weesie, J. 1993. Asymmetry and timing in the volunteer's dilemma. *Journal of Conflict Resolution* 37:569-90.

———. 1994. Incomplete information and timing in the volunteer's dilemma. *Journal of Conflict Resolution* 38:557-85.

Weesie, J., and A. Franzen. 1998. Cost sharing in a volunteer's dilemma. *Journal of Conflict Resolution* 42:600-18.

Yamagishi, T. 1986. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51:110-16.