

Who to Punish? Individual Decisions and Majority Rule in the Solution of Free Rider Problems*

Abstract

Experiments on public goods dilemmas have found that the opportunity to punish leads to higher contributions and reduces the free rider problem. But a substantial amount of punishment is targeted on high contributors. Furthermore, whether subjects would allow punishment if given the choice has been questioned. We gave subjects opportunities to vote on who, if anyone, can be punished. We found that, from their first opportunity to vote, no group ever allowed punishment of high contributors, most groups eventually voted to allow punishment of low contributors, and the result was both high contributions and high efficiency levels. Thus, we found evolution in the lab of societies that solve their free rider problems by allowing low contributors alone to be punished.

JEL Classification: C91, C92, D71, H41

Keywords: Public goods, collective action, punishment, voting, institutions.

* The research reported here was supported by N.S.F. grant SES-0001769.

Who to Punish? Individual Decisions and Majority Rule in the Solution of Free Rider Problems

0. Introduction

In organizations such as teams, firms and military units, which depend on cooperative effort to succeed, sanctions are very common, and research on the role of sanctions (or punishment) in increasing cooperation has rapidly grown in recent years. There are of course downsides of punishment. Punishment is costly to the punished, it is often costly to the punisher, and it is occasionally targeted at cooperative individuals. A rival may harm such an individual to gain personal advantage; a free rider who is punished, presumably by a cooperator, may punish back in revenge. Even with these downsides, researchers have found in experiments on voluntary contributions to a public good that punishment has dramatic effects on increasing contributions. But they also found smaller and sometimes negative effects on earnings and efficiency.¹ Because of the disappointing results on earnings, some researchers have questioned whether, if given a choice of allowing or prohibiting punishment, the opportunity to punish might not be chosen.

In this paper we asked, for an experiment on voluntary contributions to a public good: (1) if subjects could choose by majority vote, not just between the two extreme alternatives of allowing or not allowing punishment, but also among possibilities of allowing punishment with various restrictions, would they choose one of the alternatives of restricted punishment? And (2) if the subjects choose a restricted alternative that allowed punishment of low contributors but not high contributors, would this rule of punishment substantially increase contributions and efficiency compared with the alternatives of no punishment and unrestricted punishment?

It is not obvious that subjects would vote for this restricted rule of punishment. In other recent experiments, subjects have shown a preference for no punishment over unrestricted punishment, and the intermediate rule allowing punishment of low-but-not-high contributors was not an option that was studied. As a second-order public good,

¹ We follow the usual definition of efficiency in experiments, as the observed sum of earnings divided by the maximum possible sum of earnings, for a specified group or groups and periods.

punishing low contributors invites free-riding of others' punishment of low contributors, and it was not clear how the incentives to free ride would affect the voting outcomes. In addition, while there is now much evidence that many subjects engage in costly punishment of free riders, there is no clear prediction that if subjects voted for the intermediate rule, it would raise contributions by enough to offset punishment's costs to punishers and punished.

Predicting how subjects would vote is difficult. From our own experiments and reviews of others we estimated that about 15% of punishment was directed at the highest contributor in a group, and about 25% at those who contributed more than their group's average. We called this targeting of high contributors "perverse punishment" because it would seem to have negative effects on cooperation. If subjects voted as they punished, this would suggest that perverse punishment would be controlled by majority rule, at least in most votes. On the other hand, in a given period most subjects did not punish, and we could only conjecture as to the frequency with which allowing any punishment at all would be chosen, and its implications if chosen.

But the results of the experiment are unequivocal. We found that out of 160 group votes, even when groups had no prior experience with unrestricted punishment, no group ever voted to allow unrestricted punishment and no group ever allowed punishment of high contributors. Over a series of votes and periods of learning we found a distinct reluctance to allow any punishment at the beginning, with a gradual but clear evolution toward allowing punishment of low-but-not-high contributors. And groups adopting this rule of controlling perverse punishment achieved high levels of contributions and efficiency, among the highest in the literature on social dilemmas.

In sum, our main contributions are (1) to show how rules of punishment can evolve endogenously to address free rider problems; (2) to find strong evidence that perverse punishment is an important phenomenon and its control has important effects on cooperation, in terms of both contributions and efficiency; and (3) to show that a balance between rules collectively chosen by voting and decentralized, individually imposed punishment within the chosen rules can improve upon results reached by decentralized choice alone.

Our paper is related to the literature as follows. In a much-cited and by now much-replicated public goods experiment, Fehr and Gächter found (2000a) that subjects were willing to punish free riders even though it was costly to the punisher and even in the last period when no strategic benefit was possible.² The punishment led to less free-riding and higher contributions. However, Cinyabuguma, Page and Putterman (forthcoming) and Botelho, Harrison, Costa Pinto and Rutström (2006) analyzed Fehr and Gächter's data, finding lower earnings when punishment was allowed than when it was not allowed.³ Cinyabuguma *et al.* argued that a major reason why punishment reduces efficiency is the punishment of high contributors. They demonstrated that punishing high contributors reduces efficiency because it leads those targeted to reduce their contributions.⁴

Our paper is one of several contemporaneous studies that ask whether institutions of punishment will emerge and be sustained. A variety of results has been found. When subjects were allowed to choose between an institution with punishment and another without it, Botelho *et al.* (2005) found that the subjects voted overwhelmingly for the institutions without punishment. In a related experiment, Sutter *et al.* (2005) found that subjects overwhelmingly voted to allow rewards rather than punishment even though the latter raised contributions more (these experiments allowed only one vote for each group). Neither study allowed subjects to choose rules that restrict punishment to a specific domain.

The studies most closely related to ours are those of Gürerk, Irlenbusch and Rockenbach (2005, 2006). Their experimental designs allowed subjects to “vote with their feet” in choosing between two groups, one allowing unrestricted punishment and the other no punishment. Subjects initially avoided the group with punishment, but with

² Ostrom, Walker and Gardner (1992) conducted a similar common pool resource experiment using an infinitely repeated game protocol.

³ Botelho *et al.* found similar results for Fehr and Gächter (2002) as did Cinyabuguma *et al.* for public goods and sanctions experiments by Carpenter and Matthews (2002), Sefton, Shupp and Walker (2002), Page, Putterman and Unel (2005), and Bochet, Page and Putterman (2006).

⁴ The authors used regression to study the impact of punishment upon changes in the punished subject's contribution, and found that each dollar of punishment of a group's highest contributor substantially decreased his or her next period contribution. Their calculations also show that in the related public goods and sanctions experiments by Bochet *et al.* and Page *et al.*, earnings would have been higher with punishment than without it but for the presence of perverse punishment. However, earnings are not always lower in a VCM condition with unrestricted punishment than in one without punishment opportunities, as evidenced by Masclet *et al.* (2003) and by Gürerk, Irlenbusch and Rockenbach (2005, 2006).

repeated opportunities to choose, almost all eventually chose the group with punishment, in result achieving high contributions and efficiency. Gürerik *et al.* don't mention perverse punishment. They explain their results in term of an evolution to a "norm to cooperate and punish free-riders" with the result that "punishment becomes ever more unnecessary" (2006, p. 110). Our results resemble theirs in the endogenous evolution of effective punishment, but our experiment differs in that we investigate the problem of perverse punishment, our subjects vote for rules instead of selecting groups, and we include rules that partially restrict punishment.

Although not involving endogenous institutional choice, an experiment by Casari and Luini (2005), which contrasts outcomes under differing punishment rules, includes a rule requiring a subject to be targeted for punishment by at least two group members (in a group of five) before the punishment takes effect. Their finding that the restriction decreased punishment of high contributors resembles our result that punishing high contributors is never favored when subjects vote on who can be punished. Even with the restriction, however, the average contribution doesn't exceed half of the endowment, in their experiment.

The paper is organized as follows. Section 1 explains the experimental design, Section 2 presents the analysis and results, and Section 3 provides a concluding discussion.

1. Design

1.1 Basic Design

Our design extends the basic voluntary contributions mechanism (VCM) in which subjects are randomly assigned to groups that remain fixed (a "partners" design) for a finite and known number of periods. Each subject in a group is provided with an initial endowment that he or she is asked to divide between a private account and a group account. Any funds placed in the group account are scaled up by the experimenter and divided equally among the subjects in the group without regard to individual contribution. This design generates the familiar result (for self-interested subjects) that it is socially optimal to contribute everything to the group account, but privately optimal to contribute nothing.

To this basic VCM we added punishment and voting opportunities, using two designs to study how rules restricting or allowing punishment might evolve over a series of votes. Our “3-Vote” design allowed for a gradual process of learning and familiarization, see Figure 1A. At the beginning of the experiment, subjects were read and shown on the computer screen instructions describing the basic VCM mechanism (without mention of punishment or voting to come later). Then they participated in the VCM for three periods without punishment or voting.

At the beginning of the 4th period, the subjects received their second instructions. These explained the opportunity of voluntary punishment, which was unrestricted except for some budgetary constraints (these instructions did not mention the voting to come later). Then the subjects participated in three periods of the VCM with unrestricted punishment.

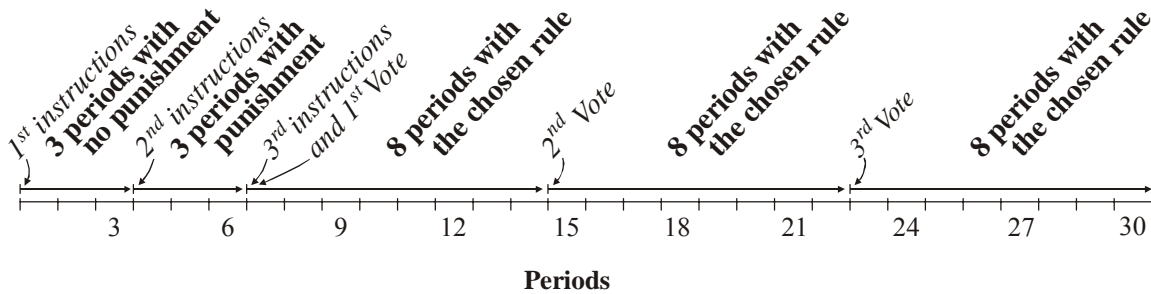


FIGURE 1A. THE 3-VOTE DESIGN

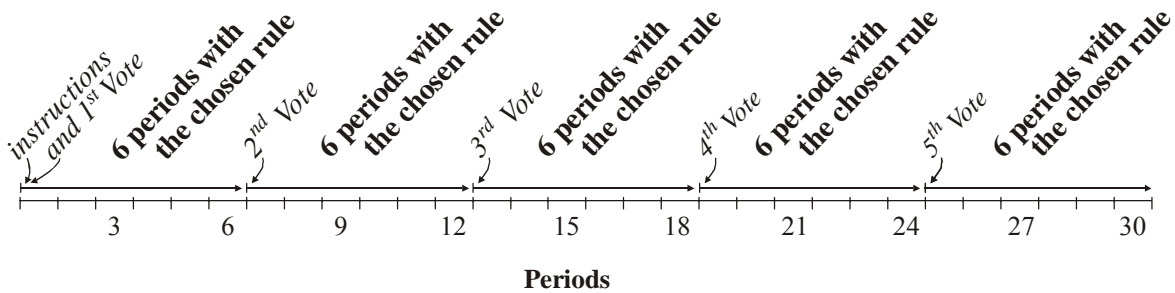


FIGURE 1B. THE 5-VOTE DESIGN

At the beginning of the 7th period, the subjects received their third instructions, which explained the voting process, and took their first vote on the rules governing who, if anyone, could be punished for the next eight periods. At the beginning of the 15th

period a second vote was taken and new rules regulating punishment were chosen. Then the subjects participated in eight periods of the VCM with punishment (if any) governed by the second chosen rules. At the beginning of the 23rd period the third and final vote was taken, and the remaining eight periods were conducted with possible punishment governed by this last vote. We included practice exercises in each of the three sets of instructions.

Surprised to find that out of 60 group votes none allowed punishment of high contributors, we asked if this uniform pattern of voting would hold up if experimental subjects had no prior experience with unrestricted punishment and more opportunities to vote. The question led to an additional, “5-Vote” design, Figure 1B. In this design the task of learning and familiarization was harder. The first and only instructions, given at the beginning of the experiment, explained the basic VCM mechanism without punishment, possible rules governing punishment, and the opportunity to vote on them. Again, the instructions included practice exercises.

After the instructions at the beginning of the 5-Vote design, the subjects voted to allow or restrict punishment. Then they participated for six periods in the VCM, governed by the chosen rules of punishment. At the beginning of the 7th period, the subjects voted again, and then participated in 6 periods of the VCM, governed by the chosen rules of punishment. The same process repeated for three more times, as shown in Figure 1B. The 5-Vote design had the same number of periods (30) as the 3-Vote design and took about the same amount of time.

In both designs, sessions had sixteen subjects assigned randomly to four groups of four subjects who remained together throughout the session. Each subject knew there were 16 subjects in the experiment room but could not tell which among the others in the session belonged to her group. Contribution and punishment choices (if any) were announced to other group members under randomly changing labels B, C, and D, for one’s fellow members, so that the behaviors of individuals could not be tracked from period to period, except by conjecture. A subject learned the total amount of punishment she had received, but not which group members punished her or by how much.

1.2 Payoffs

All periods shared the same underlying structure. In each period, each subject had to decide on a division of 10 experimental dollars, in integer amounts, between a private account and a group account, before observing the choices of fellow group members. In a period, subject i earned

$$(1) \quad y_i = (10 - C_i) + (0.4) \sum_{j=1}^4 C_j$$

where C_i is i 's contribution to the public account and the summation is taken over all members of i 's group, including i . After all four made their decisions, each was informed of the choices of the others. When punishment was permitted, it cost a subject 0.25 experimental dollars to reduce the earnings of another person by 1.00 experimental dollar. Subject i 's earnings after punishment were thus

$$(2) \quad y_i = (10 - C_i) + (0.4) \sum_{j=1}^4 C_j - (0.25) \sum_{j \neq i} R_{ij} - \sum_{j \neq i} R_{ji}$$

where R_{ij} is the number of dollars by which i reduced j 's earnings, and conversely for R_{ji} . General constraints on punishment in all treatments were: (i) a subject could not spend more than her/his pre-punishment earnings for the period on reducing the earnings of other subjects, (ii) a subject's post-punishment earnings for a period would be set to zero if earnings y_i in equation (2) were negative, and (iii) a subject i could not spend more on reducing the earnings of a subject j in any period than would single-handedly reduce j 's earnings according in (2) to less than zero.⁵ The Appendix shows the screen design for entering an individual's contribution and punishment decisions.

1.3 Voting

In a voting stage, each subject checked off one of three boxes beside each of three ballot items, on a screen set up as follows:

⁵ The purpose of (i) and (ii) was to keep all decisions financially independent of each other while maintaining a guaranteed minimum payment for recruiting reasons. The purpose of (iii) was to help subjects to avoid pointless spending on punishment in view of constraint (ii). Note, however, that it remained possible for subjects to "overspend" on punishing in the sense that both subject i and, say, subject k might each spend enough to reduce j 's earnings for the period to zero, although only one subject's punishment would actually be effective in that case, given (ii). This could happen because subjects did not learn of punishment not carried out by or aimed at them, and the design (as in Fehr and Gächter, 2000a) keeps such information private so as not to encourage free riding on punishment.

I vote to allow a person's earnings to be reduced if

- | | | | |
|--|---------------------------------|--------------------------------|---|
| (a) that person assigns less than the average amount ⁶ to the group account | Yes <input type="checkbox"/> | No <input type="checkbox"/> | No preference <input type="checkbox"/> |
| (b) that person assigns the average amount to the group account | Yes <input type="checkbox"/> | No <input type="checkbox"/> | No preference <input type="checkbox"/> |
| (c) that person assigns more than the average amount to the group account | Yes <input type="checkbox"/> | No <input type="checkbox"/> | No preference <input type="checkbox"/> |

In each group of four subjects, of those expressing a preference in ballot item (a), if there was a majority or tie of “No” votes against punishment of low contributors, then punishment of low contributors would be prohibited for the next 8 periods in the 3-Vote design and 6 periods in the 5-Vote design; and if a majority voted “Yes,” punishment of low contributors would be allowed; and correspondingly for ballot items (b) and (c).⁷ After the vote, each group's members received a message reporting the voting outcome, which was one of $2^3 = 8$ possible punishment rules (*i.e.*, combinations of the three ballot item choices).⁸

When a group voted to restrict punishment, a fixed zero appeared in the punishment box⁹ for all individuals to which the restriction applied during the punishment stages that followed each contribution stage. For example, members of a group that had voted to prohibit all punishment saw the standard punishment stage screen with fixed 0's in all the punishment boxes, indicating that no punishment was allowed in this case. Just before the second and later votes, each subject was informed of the punishment rule chosen in the preceding votes of each of the four groups in a session along with each group's average contributions and earnings during the periods the rule

⁶ As explained in the instructions, “average amount” meant the average over the four members of the group in the contribution stage of the period in question. It could vary among groups and within a given group from one period to the next. Note that a vote to allow punishment of those contributing less than the group average of 4 players is the same event as a vote to allow punishment of those contributing less than the average of the 3 others.

⁷ We expected few cases where someone was exactly an average contributor, but for symmetry we treated the average contributor on a separate ballot item.

⁸ Only “Yes” and “No” votes were counted in determining majorities; for example, if 2 voted “Yes” and 2 voted “No,” the proposal did not pass, but if 2 voted “Yes,” 1 “No” and 1 “No preference,” the proposal passed. Subjects were informed of whether a ballot item passed or not, but not by how many votes or who voted which way.

⁹ See the boxes labeled b', c' and d' on the lower left portion of the diagram in the Appendix showing the screen design.

governed (the information was new for the most recently taken vote, and was repeated for the earlier votes). While revealing this information forces us to treat the session, rather than the group, as the statistically independent observation from the second vote onwards, it has the virtue allowing learning from the examples of others, as occurs in many real-world settings.

We conducted five sessions of each design using a total of 160 subjects (see Table 1).¹⁰ All of the sessions were conducted by computer in a computer lab at Brown University. At the end of each session, cumulative earnings for the thirty periods were totaled and converted to real money at the rate of 25 experimental dollars to one real dollar, and \$5 was added as a participation fee. Sessions typically lasted a little less than two hours including instructions, and subjects' overall earnings averaged approximately \$25. Instructions for both designs are similar and available in our Working Paper.¹¹

TABLE 1. NUMBERS OF GROUPS, SUBJECTS, AND VOTES

| Session design | Number of sessions | Number of groups in each session | Number of subjects in each group | Total number of subjects | Total number of group votes on rules |
|----------------|--------------------|----------------------------------|----------------------------------|--------------------------|--------------------------------------|
| 3-Vote | 5 | 4 | 4 | 80 | 60 |
| 5-Vote | 5 | 4 | 4 | 80 | 100 |

2. Results

2.1 The Voting Pattern

¹⁰ Subjects were Brown undergraduates, recruited by (a) distribution of flyers in the mailboxes of all undergraduates, (b) distribution of flyers in a large introductory economics course, (c) distribution of table slips at a student dining hall, and (d) advertising under the heading of employment in an on-line campus magazine, the *Brown Daily Jolt*. Analysis of information provided in the post-experiment debriefing shows that the subjects majored in a large range of concentrations, with the economics concentration being that of only 15%, about 5% more than the proportion in the overall student body. A little less than half the subjects had taken no economics courses at the college level. 67% of the subjects were female, somewhat higher than the 53% share in the general student body. Brown's undergraduate population numbers about 5500, so students participating in a given session tended not to know one another.

¹¹ In the instructions and experiment we used neutral language and did not use words like "free riding," "punishment," and "perverse punishment." See also Cinyabuguma *et al.* forthcoming, where we point out that punishment is most clearly perverse when aimed at a group's highest contributor. Here as in that experiment we distinguish between punishment of above average, average, and below average contributors, rather than between punishment of highest and of other contributors, because this seems more symmetrical and less likely to convey a biased framing of the problem to subjects.

In the 3-Vote design there were 720 individual votes (80 subjects each voting 3 times on 3 ballot items), and in the 5-Vote design 1200 individual votes. Table 2 shows the number of individual votes on each ballot item. The table shows a substantial number of individuals voted to allow punishment of higher-than-average contributors, but many more voting to allow punishment of less-than-average contributors.

TABLE 2. NUMBERS OF INDIVIDUAL VOTES TO ALLOW PUNISHMENT OF HIGH, AVERAGE, AND LOW CONTRIBUTORS, BOTH DESIGNS

| | Yes | No | No Preference |
|--|-----|-----|---------------|
| Allow punishment of less than average contributors | 410 | 211 | 19 |
| Allow punishment of average contributors | 46 | 577 | 17 |
| Allow punishment of above average contributors | 111 | 493 | 36 |

In the 3-Vote design there were 60 group votes (see Table 1), and in the 5-Vote design there were 100 group votes. In the 160 group votes altogether, only 4 of the 8 possible combinations of rules were ever chosen by majority rule. These were to allow: (i) no punishment, 56 group votes; (ii) punishment of lower-than-average contributors and no other punishment, 98 votes; (iii) punishment of low-or-equal-to-average contributors and no other punishment, 4 votes; and (iv) punishment of equal-to-average contributors and no other punishment, 2 votes. Conspicuously absent from this list is that no group ever voted to allow punishment of higher-than-average contributors.

RESULT 1. No group ever voted to allow punishment of higher-than-average contributors, so perverse punishment was ruled out from the first opportunity to vote.

In ruling out perverse punishment, every group also ruled out unrestricted punishment from the beginning. The two rules “punishment of lower-than-average contributors and no other punishment” and “punishment of low-or-equal-to-average contributors and no other punishment” are similar and we grouped them together under the heading of allowing punishment of “low-but-not-high” contributors. Figure 2 shows

how the group voting evolved, over the sequence of votes for the 3-Vote design and 5-Vote designs. Result 2 summarizes the voting pattern over time.

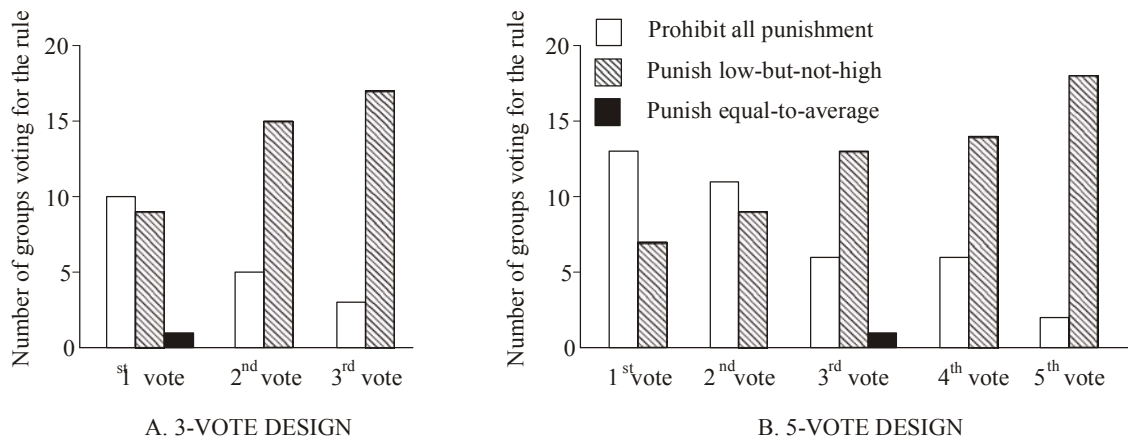


FIGURE 2. EVOLUTION OF THE VOTING RULES

RESULT 2. *In both designs, a plurality of groups voted in their first vote to prohibit all punishment, with a substantial minority of groups voting to allow punishment of low-but-not-high contributors. Over the sequence of votes, this ordering reversed, so that in the final vote, nearly all groups voted to allow punishment of low-but-not-high contributors, with only a few remaining groups voting to prohibit all punishment.*

2.2 Contributions and Efficiency

Figure 3 shows period-by-period contributions of groups for the two composite rules most frequently chosen. In both the 3- and 5-Vote designs, groups that allowed punishment of low-but-not-high contributors achieved substantially higher levels of contributions than did groups that prohibited punishment altogether, significant ($p < 0.001$) in Mann-Whitney tests.¹² In both designs, contributions in groups that permitted punishment of low-but-not-high contributors tended to increase over time until the end-

¹² The test compared all groups over the periods governed by voted rules. The observations used are averages of contributions in groups during the 8 periods that a given voted rule governs. For example, in the 3-Vote design if a particular group of 4 subjects voted to allow no punishment during periods 7-14, voted to allow punishment of low contributors in periods 15-22, and again voted to allow punishment of low contributors in periods 23-30, its average contribution of the first 8 periods would constitute one observation in the no-punishment category, those of the second 8 periods one observation in the punish-low-but-not-high category, and those of the last 8 periods another observation in that category.

game fall off. In contrast, 3-Vote design groups that prohibited all punishment had falling levels of contributions over time, replicating earlier results on basic VCMs without punishment, and in the 5-Vote design contributions had a slightly increasing trend in the middle periods.¹³

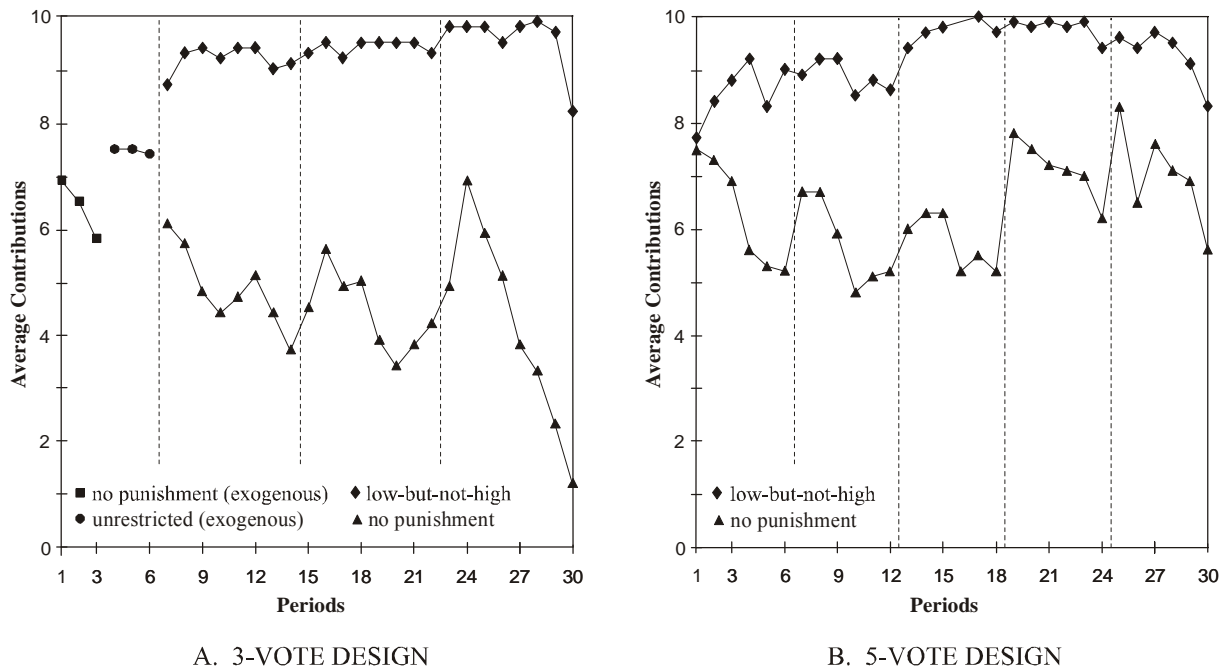
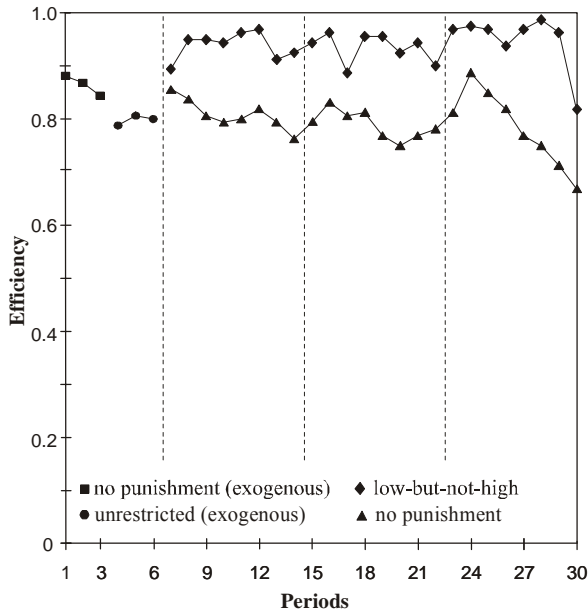


FIGURE 3. AVERAGE CONTRIBUTIONS FOR THE TWO DESIGNS, BY PERIOD AND PUNISHMENT RULE

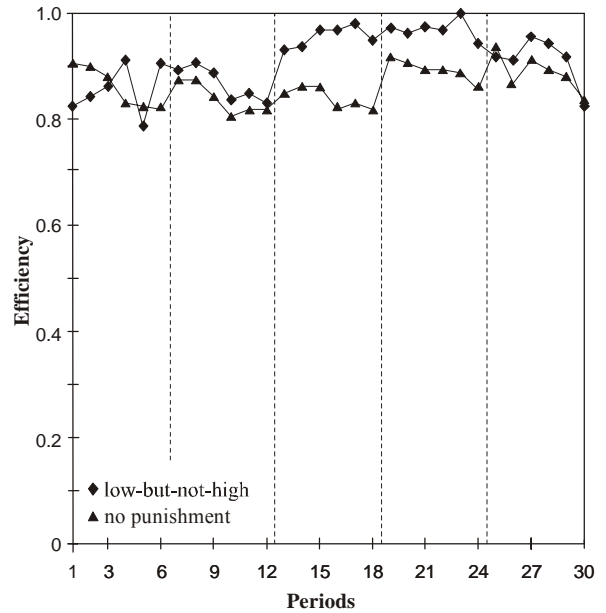
Figure 4 shows period-by-period efficiency¹⁴ of groups that voted to prohibit all punishment and groups that voted to prohibit perverse punishment while allowing punishment of low contributors. In Figure 4A, average period efficiency was always higher under the rules allowing punishment of low contributors, and similarly in Figure 4B, except in six periods. In Mann-Whitney tests comparing all groups over the periods governed by voted rules, the groups which voted to allow punishment of low-but-not-high contributors had significantly higher efficiency than groups which voted for no punishment, at the 0.1% level for the 3-Vote design and 1% for the 5-Vote design.

¹³ In Figure 3 contributions under the endogenously chosen rule of no punishment are more sustained and decline more slowly than is typical in a VCM without punishment. But endogenous choice includes its process, including repeated voting and the ability to change rules, possibly leading to commitment effects (see Sutter *et al.*), restart effects (see the dashed vertical lines in Figure 3), and selection effects as groups change rules in response to free-riding behavior.

¹⁴ We follow the usual definition of efficiency in experiments, as the observed sum of earnings divided by the maximum possible sum of earnings, for a specified group or groups and periods.



A. 3-VOTE DESIGN



B. 5-VOTE DESIGN

FIGURE 4. EFFICIENCY FOR THE TWO DESIGNS, BY PERIOD AND PUNISHMENT RULE

Table 3 compares contributions and efficiency under the two most voted rules, and the exogenously imposed conditions of unrestricted punishment (periods 4-6 of the 3-Vote design) and no punishment (periods 1-3).¹⁵ The results of the five tests of Table 3 are summarized in Result 3:

RESULT 3. For each of the Wilcoxon matched pair tests on contributions, contributions are higher under the rule of punish low-but-not high than under the rule of unrestricted punishment, and contributions are higher under the rule of unrestricted punishment than under the rule of no punishment, and this ordering is transitive. Correspondingly, efficiency is higher under punish low-but-not high than under unrestricted punishment, and

¹⁵ For example, in comparing contributions under the rule of punish low-but-not-high with contributions under the (exogenous) rule of unrestricted punishment in Test 1, we considered the 17 groups of the 3-Vote design that eventually chose the rule allowing punishment of low-but-not-high contributors (see Figure 2A). For each of these groups we calculated the average group contribution over the first three periods that the group was governed by this rule. We matched this average with the same group's average contribution over the three periods of unrestricted punishment (periods 4-6 of the 3-Vote design). In the 17 matched pairs, 14 groups had higher contributions under the rule of punish low-but-not-high, 2 groups had higher contributions under unrestricted punishment, and 1 group was tied. The difference is significant ($p = 0.001$) in a two-tailed Wilcoxon matched pair test.

efficiency is higher under no punishment than under unrestricted punishment, and this ordering is transitive.

TABLE 3. EFFECTS OF THE PUNISHMENT RULE ON CONTRIBUTION AND EFFICIENCY

| Test | Ranks of Contributions by the Punishment Rule | Test | Ranks of Efficiency by the Punishment Rule |
|------|---|------|--|
| 1 | punish low > unrestricted**** | 4 | punish low > no punishment*** |
| 2 | unrestricted > no punishment** | 5 | punish low > no punishment** |
| 3 | unrestricted > no punishment** | 2 | no punishment > unrestricted* |
| 4 | punish low > no punishment*** | 3 | no punishment ~ unrestricted |
| 5 | punish low > no punishment**** | 1 | punish low > unrestricted*** |

Notes: Wilcoxon Matched Pair Tests. Tests 1–4 are for groups in the 3-Vote design. **Test 1** compares the average contributions of the first three periods, if any, in which a group chose “punish low” matched with the average contributions of the same group in periods 4–6 of “unrestricted punishment” (the number of distinct groups matched and compared is $n = 17$); and correspondingly for efficiency. **Test 2** compares contributions (efficiency) for groups in periods 1–3 with contributions for the same groups in periods 4–6, $n = 20$. **Test 3** compares periods 4 – 6 with 7-9, for the groups that chose no punishment in periods 7–9, $n = 10$. **Test 4** compares the same groups before and after a switch from “no punishment” to “punish low,” comparing the 8-period averages before and after the switch, $n = 9$. **Test 5** is the same as Test 4, except it is for the 5-Vote design and 6-period averages are compared before and after the switch, $n = 17$. **** indicates significance at the 0.1% level, *** at the 1% level, ** at the 5% level, * at the 10% level, and ~ insignificant, in two-tailed tests. “Punish low” indicates “punish low-but-not-high.”

Because of the difference in the orderings for contributions and efficiency, the sequence or tests in Table 3 for efficiency are rearranged to show the transitivity. The difference in the orderings of contributions and efficiency is likely due to the cost of punishment.

2.3 Mitigating the Free-Rider Problem

In the literature on public goods games, it is common practice to use the term “free rider” loosely to denote any individual who contributes less than the socially optimal amount. It’s worth noting, however, that whereas in the absence of punishment anyone who contributes less succeeds in earning more and thus in obtaining a “free ride” on others’ contributions, when punishment is possible a low contributor may fail to earn more, and therefore fail to free ride in actuality. To compare how successfully different sets of rules address free riding, we adopt in this section a definition that considers the full outcome, not simply the contribution decision.

Specifically, the symmetric design of this and other VCM experiments suggests a simple definition of free riding: a subject A experiences free riding when someone else in his group, B, contributes less to the public good but earns more than A does.¹⁶ For a specified punishment rule, sequence of periods, and collection of groups, we define the *frequency* of free riding as the number of cases of free riding divided by the number of observations, and an *observation* as a pairing in a group, where one subject in a group has a higher contribution than the other subject of the pair. By the design of a basic VCM without punishment and its payoff equation (1), every time someone contributes more than someone else, there is a case of free riding because the higher contributor always has lower earnings. Thus, in this definition of free riding, the frequency of free riding for the basic VCM is 100% (as shown in the first bar of Figure 5). But the frequency of free riding may decrease when sufficient punishment is directed at low contributors.

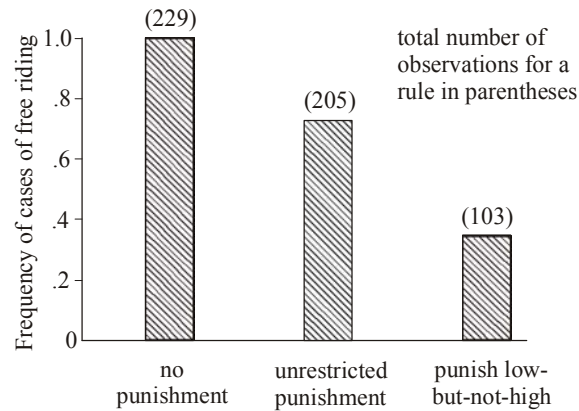


FIGURE 5. FREQUENCY OF CASES OF FREE RIDING, BY PUNISHMENT RULE

For the rule of unrestricted punishment, over all 20 groups in periods 4-6 of the 3-Vote Design, there were 205 observations of pairs of unequal contributions by subjects in a group, and 148 cases of free-riding, for a frequency of 72% (see the middle bar). In comparison, the frequency of free riding in the first three periods after a group voted for the rule of punishing low-but-not-high contributors was 35% of the 103 observed unequal

¹⁶ Under this definition, if everyone in a group contributed the same low amount, there would be no free riding (it is only defined for unequal contributors).

pairs. This is a striking reduction, considering that the rule of punish low-but-not-high does not prevent a higher-than-average contributor from free riding on a still higher contributor. The difference in free riding between unrestricted punishment and punish low-but-not-high contributors is significant ($p < 0.0001$) in a Fisher exact test.¹⁷

RESULT 4. *In comparing VCMs with rules governing punishment, we find the highest frequency of free-riding in groups operating with no punishment, less free-riding in groups with unrestricted punishment, and least free riding in groups allowing punishment of low-but-not-high contributors.*

A regression analysis of incentives to free ride finds the same ordering as in Result 4. In the regressions below, we follow Fehr and Gächter in defining subject i 's absolute negative and positive deviations from the average of others' contributions as:

$$\begin{array}{l} \text{Absolute} \\ \text{Negative} \\ \text{Deviation} \end{array} = \begin{cases} |C_i - \bar{C}_{-i}| & \text{if } C_i < \bar{C}_{-i} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \begin{array}{l} \text{Positive} \\ \text{Deviation} \end{array} = \begin{cases} |C_i - \bar{C}_{-i}| & \text{if } C_i > \bar{C}_{-i} \\ 0 & \text{otherwise} \end{cases}$$

where $\bar{C}_{-i} = \frac{\sum_{j \neq i} C_j}{3}$ is the average of others' contributions.

Using Fehr and Gächter's specification (see their Table 5, p. 991), we first consider behavior in the three periods of the exogenously imposed rule of unrestricted punishment (periods 4–6 of the 3-Vote design, see column (1) of Table 3), and compare this with the first three periods of the endogenously chosen rule allowing punishment of low-but-not-high in both the 3 and 5-Vote designs (columns (2) and (3)).¹⁸ Then, secondly, we consider behavior for the punish low-but-not-high rule over all the periods which it governs in the 3- and 5-Vote designs (columns (4) and (5)).

In each regression of Table 4 the dependent variable is each subject i 's punishment received in each period (three periods for regressions (1), (2), and (3), and up

¹⁷ We also did a Wilcoxon matched pair test, which is also significant; see the Working Paper for details.

¹⁸ We include observations for only the first three periods under a rule in columns (2) and (3) to achieve comparability with the regression for periods 4-6 (column (1)), in view of the possibility that learning or other factors might change behaviors with more repetitions.

to 24 and 30 periods in regressions (4) and (5) respectively). The independent variables are the Average Contribution of Others, i 's Absolute Negative Deviation, i 's Positive Deviation, and period and group dummies (not shown).^{19,20}

TABLE 4. DETERMINANTS OF PUNISHMENT RECEIVED

| <u>Dependent Variable: experimental dollars of punishment</u> | | | | | |
|--|--|--|--|--|--|
| Independent variables | First three periods of the rule | | | All periods of the rule | |
| | Unrestricted Punishment 3-Vote Design (1) | Punish Low-But-Not-High 3-Vote Design (2) | Punish Low-But-Not-High 5-Vote Design (3) | Punish Low-But-Not-High 3-Vote Design (4) | Punish Low-But-Not-High 5-Vote Design (5) |
| | -0.74 (1.067) $p = 0.490$ | 4.086* (2.353) $p = 0.088$ | 19.754*** (4.587) $p < 0.001$ | 0.587 (2.222) $p = 0.792$ | 11.483*** (4.367) $p = 0.010$ |
| <i>Constant</i> | | | | | |
| Average Contribution by Others | 0.388** (0.175) $p = 0.028$ | -0.230 (0.244) $p = 0.350$ | -1.090*** (0.405) $p = 0.009$ | 0.228 (0.206) $p = 0.269$ | -0.654** (0.269) $p = 0.016$ |
| Positive Deviation | 0.377** (0.152) $p = 0.014$ | n.a. | n.a. | n.a. | n.a. |
| Absolute Negative Deviation | 0.888*** (0.221) $p < 0.001$ | 1.217*** (0.148) $p < 0.001$ | 1.039*** (0.122) $p < 0.001$ | 1.054*** (0.138) $p < 0.001$ | 0.967*** (0.095) $p < 0.001$ |
| R-squared | 0.54 | 0.91 | 0.86 | 0.75 | 0.78 |
| Observations | 240 | 82 | 92 | 241 | 176 |

Notes: Punishment received as a function of deviation from group average in unrestricted and restricted punishment conditions. OLS regressions with period and group fixed effects, not shown. Unrestricted punishment, in Column 1, is observed in periods 4-6, where each observation is for one subject and one period. Columns 2–5 include one observation per subject under the rule allowing punishment of low-but-not-high contributors. In Columns 2 and 3, only the first three periods in which a group adopted the rule for the first time are included, while Columns 4 and 5 include all periods of restricted punishment. Numbers in parentheses are White heteroskedasticity-consistent standard errors; *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level.

To compare how the rules of punishment affect the estimated incentives toward contributing \$1 less, Table 5 focuses on coefficients of Absolute Negative Deviation and Positive Deviation from Table 4. For example, in Column (1) of Table 5 the coefficient

¹⁹ In both the unrestricted (Column 1) and restricted (Columns 2-5) punishment regressions, only the observations of individuals who could potentially be punished are included. The difference is that under unrestricted punishment, anyone can be punished.

²⁰ The regressions were also estimated by the Tobit method, treating 0 punishment observations as potentially left-censored. Resulting coefficients are similar and similarly significant except in the case corresponding to Column (1), where they are not significant at conventional levels.

for Absolute Negative Deviation is \$0.89, the estimated punishment for a \$1 reduction in contribution for a less-than-average contributor, under the rule of unrestricted punishment, in the first three periods of the 3-Vote design, and shown as a negative gain of \$-0.89 in Column (1) of Table 5. In Column (2) of Table 5 the coefficient for Absolute Negative Deviation is \$1.22, the estimated punishment for a \$1 reduction in contribution for a less than average contributor, under the rule of punish low-but-not-high contributors, in the first three periods of the 3-Vote design, and shown as a negative gain of \$-1.22 in Column (2) of Table 5, etc.

The \$+0.60 throughout Table 5 is the \$1 gain in the private account from reducing one's contribution by \$1, minus the \$0.40 loss in the individual's earnings from the group account. In Column (1) of Table 5 the coefficient for Positive Deviation is \$0.38, the estimated punishment for each \$1 of additional contribution for a higher-than-average contributor, under the rule of unrestricted punishment, in periods 4-6 of the 3-Vote design. The \$+0.38 in Column (1) of Table 5 is the positive gain from contributing \$1 less and avoiding \$0.38 in perverse punishment, for a higher-than-average contributor. The cases labeled n.a. in Table 5 are for the rule of punish low-but-not-high in Columns (2)-(5), in which case punishment of higher-than-average contributors is not allowed.

TABLE 5. INCENTIVES TO CONTRIBUTE \$1 LESS

| | Less-than-average contributors, subject to punishment | | | | |
|---------------------|---|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| | (1) | (2) | (3) | (4) | (5) |
| | unrestricted punishment | punish low- but-not-high | punish low- but-not-high | punish low- but-not-high | punish low- but-not-high |
| Abs. neg. deviation | \$-0.89 | -1.22 | -1.04 | -1.05 | -0.97 |
| \$1 account shift | + 0.60 | + 0.60 | + 0.60 | + 0.60 | + 0.60 |
| Net gain | \$-0.29 | -0.62 | -0.44 | -0.45 | -0.37 |
| | Higher-than-average contributors, subject to punishment only in Column (1) | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| Positive deviation | +0.38 | n.a. | n.a. | n.a. | n.a. |
| \$1 account shift | + 0.60 | + 0.60 | + 0.60 | + 0.60 | + 0.60 |
| Net gain | +0.98 | +0.60 | +0.60 | +0.60 | +0.60 |

Note: Net gain is the change in earnings from contributing \$1 less.

Table 5 shows that for less than average contributors the net gain from contributing \$1 less is negative for each of the cases in Columns (1)-(5). The \$-0.29 in Column (1) suggests that unrestricted punishment can reverse a subject's incentive to free ride, for a subject contributing less than average, replicating Fehr and Gächter's earlier finding for the case of less-than-average contributors. But the negative gains for less than average contributors is even more negative in Columns (2)-(5), suggesting that the incentive against free-riding is strengthened for less than average contributors under the rule of punish low-but-not-high.

Table 5 suggests that the incentives to contribute \$1 less for higher-than-average contributors is *not* reversed under unrestricted punishment or the rule of punish low-but-not-high. In Column (1) under unrestricted punishment, a subject with a higher-than-average contribution makes an estimated net gain of \$0.98 from contributing \$1 less (a gain of \$0.38 from reduced perverse punishment added to the \$0.60 gain from shifting away from the group account). In Columns (2)-(5), under the rule of punish low-but-not-high, a higher-than-average contributor bears no punishment, but still gains the \$0.60 from a \$1 shift from the public account. While neither rule reverses the incentive for a higher-than-average contributor to contribute less, the incentive toward free riding is less under the rule of punish low-but-not-high than under unrestricted punishment.²¹

In other words, when it is allowed, perverse punishment appears to exacerbate the incentive problem for high contributors. In fact there was considerable perverse punishment in periods 4–6 of the 3-Vote design. Of the 129 events of punishment, 28% were punishments aimed at higher-than-average contributors for the period and group in question, 19% at the highest contributor for the period and group in question and 11% at individuals who contributed their full endowment.²²

²¹ When subjects make their contribution decision, they don't know what the other subjects' contributions will be, and are uncertain of what will be the average and its boundary line of punishment risk. This uncertainty creates an incentive toward higher contributions to be on the safe side of the unknown boundary.

²² These percentages are calculated by counting each event (rather than dollar amount) of someone punishing someone else. They may be atypically high due to the short duration of the unrestricted punishment portion of our experiment. Yet similarly large amounts of perverse punishment are found in some other studies; see for example Anderson and Putterman (forthcoming), Gächter and Herrmann (2005), and for a regression result similar to column (1), in which the absolute positive deviation term also has a positive significant coefficient, Ones and Putterman (forthcoming), Table 2.

2.4 Do Subjects Vote “According to their Type”?

We conjectured that even though some subjects use opportunities to perversely punish (when punishment is unrestricted) and would likely vote to allow perverse punishment in our experiment, punishment of high contributors might nonetheless be ruled out since few groups would have a majority of members of this type. Results at group level are consistent with this conjecture. Is there also evidence at the level of individuals, however, that subjects tended to vote “according to type”? Logit regressions provide some affirmative evidence.

We estimated regressions in which the dependent variable is 1 if a subject voted to permit punishment specified by a particular rule and 0 otherwise. Explanatory variables included the subjects’ contributions relative to their group averages during the periods preceding each vote, measures of how much punishment they had given and received, and vote and group dummies. The coefficients on the subjects’ relative contribution were positive in the regressions on voting to allow punishment of low contributors, significant at the 5% level or better for both the 3- and the 5-Vote design, and negative in the regressions on voting to allow punishment of high contributors, significant at the 10% level in the regression for the 3-Vote but not in that for the 5-Vote design.

RESULT 5: Subjects were more (less) likely to vote to allow punishment of less- (greater-) than-average contributors the higher on average was their contribution above their group’s average contribution in the eight (six) previous periods.

Details are in the Working Paper.

3. Discussion and Conclusion

The main conclusion is that the endogenous choice of institutional rules of punishment can mitigate the free rider problem and reduce perverse punishment. We found that, given the opportunity to vote, no group voted to allow unrestricted punishment and no group allowed punishment of high contributors. The favored choice of a punishment rule, allowing punishment of low-but-not-high contributors, increased

both contributions and efficiency, compared with unrestricted punishment and no punishment. Whereas permitting subjects to punish whomever they please raised contributions but not efficiency, in our experiment, the combination of majority determination of rules with voluntary individual choice of contributing and punishing allowed substantially high efficiencies to be attained.

In a parallel way with Fehr and Gächter's and others' VCM experiments, our results are inconsistent with the predictions of the iterated dominance equilibria for purely self-interested players, but consistent with predictions for players with heterogeneous preference types, including both conditional cooperators and active resisters of cooperation (e.g. perverse punishers). Many subjects punish low contributors even in the last period, suggesting a value placed on "negative reciprocity" (Fehr and Gächter 2000b). Yet conditional cooperators are not the only non-payoff-maximizing type present. Analysis of earlier experiments and of the periods of unrestricted punishment in our 3-Vote design suggests about 25% of punishment is targeted at higher-than-average contributors.

With a "demography" of preference types in which roughly one in five favor such a rule and with considerable punishment of high contributors occurring when it is permitted, fully decentralized decisions about punishment leads to "scattershot" behavior reducing punishment's efficiency. Majority votes on when punishment is allowed suppressed this behavior, because the proposal to allow punishment of cooperators was voted down every time (out of 640 individual votes on the ballot item to allow punishment on high contributors 111 votes, or 17%, were "Yes" votes). Casari and Luini's (2005) finding that most punishment of those contributing 75% or more of the endowment was suppressed for lack of a "seconded," in their treatment in which at least two must punish, is also suggestive of such a demography.

We suspect that perverse punishment – or more generally, resistance to those promoting group cooperation – may be a widespread phenomenon, not just limited to public goods experiments with unrestricted punishment. One student, who had severely punished someone who had contributed her whole endowment to the group account, offered in his debriefing a quite general reason for his action, explaining that when someone does something morally superior to you, you feel uncomfortable and want to get

back at that person. The inclination to retaliate against those punishing one's free riding is also manifest in recent experiments by Nikiforakis (2004) and Cinyabuguma *et al.* (forthcoming). In ordinary life, such retaliation might be done informally by poor-mouthing, ridicule, or ostracism.

In our experimental design, inefficient free riding is a dominant strategy for purely self-interested actors, leading to low efficiency. But under majority rule voting self-interested actors combined with cooperators to outvote a minority of perverse punishers, thus freeing cooperative types to punish free-riding without fear of retaliation, thereby ameliorating the free-rider problem and achieving higher efficiency. In the larger context of practical politics, Ordeshook (pp. 213-215) analyzed a model of majority rule voting for political pork (legislators voting for projects benefiting one's own district with the costs borne by the whole country). In his model of purely self-interested legislators it is a dominant strategy for every legislator to externalize the costs of his district's project, for cases when the costs of each project exceed the benefits, leading to low efficiency.

But when there is a combination of self-interested types combined with some conditional cooperators, the cooperators may decline to add a pet project to a pork bill and vote against it. Would ever there be enough cooperators to form a majority and vote down a pork barrel bill? Elsewhere, we found (Page, Putterman, and Unel, 2005) that cooperators could be a majority, under favorable conditions when they were allowed to group themselves together, but this majority did not manifest itself in the baseline condition when the experimental subjects were grouped randomly. In actual legislatures of course, legislators form coalitions and it is common for some legislators to vote against pork bills, occasionally voting down such bills.

As another example, Meltzer and Richard's (1981) model of the level of redistributive taxation uses a median voter solution assuming strictly self-regarding preferences. More accurate explanations of the level of redistribution and its variation over time and place would consider the strength of preferences for greater equality, on the parts of some citizens, and resentment of the "undeserving poor," on the parts of others (see, for instance, Benabou and Tirole, 2005). Such an addition of two almost opposite social preference types alongside self-interested types resembles the specific case studied in this paper, where self-interested subjects co-exist with both cooperation-

preferring and cooperation-resisting types, with the associated demographic leading to predictable voting outcomes. Applying the emerging study of other-regarding preferences (Camerer and Fehr, 2004) to the study of public choice is a promising area for future research. Our study of endogenous institutional choice illustrates how the experimental method can be applied to study majority rule voting when self-interested and other preference types interact.

References

- Anderson, Christopher and Louis Putterman, forthcoming, "Do Non-strategic Sanctions Obey the Law of Demand? The Demand for Punishment in the Voluntary Contribution Mechanism," *Games and Economic Behavior* (in press).
- Benabou, Roland and Jean Tirole, 2005 "Belief in a Just World and Redistributive Politics," C.E.P.R. Discussion Paper 4952.
- Bochet, Olivier, Talbot Page and Louis Putterman, 2006, "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior and Organization* 60: 11-26.
- Botelho, Anabela, Glenn Harrison, Ligia M. Costa Pinto and Elisabet E. Rutström, 2005, "Social Norms and Social Choice," unpublished paper, Dept. of Economics, University of Central Florida.
- Camerer, Colin and Ernst Fehr, 2004, "Measuring social norms and preferences using experimental games: a guide for social scientists," Chapter 3, pp. 55-95 in *Foundations of human sociality: economic experiments and ethnographic evidence from fifteen small scale societies*, (ed. Joseph Henrich, et al.); Oxford; New York: Oxford University press.
- Carpenter, Jeffrey and Peter Matthews, 2002, "Social reciprocity," Middlebury College Department of Economics Working Paper #29.
- Casari, Marco and Luigi Luini, 2005, "Group Cooperation Under Alternative Punishment Institutions: An Experiment" Working Paper, Department of Economics, University of Siena.
- Cinyabugama, Matthias, Talbot Page and Louis Putterman, forthcoming, "Can Second-Order Punishment Deter Perverse Punishment?" *Experimental Economics* (in press).
- Ertan, Arhan, Talbot Page and Louis Putterman, 2005, "Can Endogenously Chosen Institutions Mitigate the Free-Rider Problem and Reduce Perverse Punishment?" Brown University Department of Economics Working Paper 2005-13.
- Fehr, Ernst and Simon Gächter, 2000a, "Cooperation and Punishment," *American Economic Review* 90: 980-94.
- Fehr, Ernst and Simon Gächter, 2000b, "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives* 14 (3): 159-81.
- Fehr, Ernst and Simon Gächter, 2002, "Altruistic Punishment in Humans," *Nature* 415: 137-40.

Gächter, Simon and Benedikt Herrmann, 2005, "Norms of Cooperation among Urban and Rural Dwellers: Experimental Evidence from Russia," unpublished paper, University of Nottingham.

Gürerk, Ö., B. Irlenbusch and B. Rockenbach, 2006, "The Competitive Advantage of Sanctioning Institutions," *Science* 312 pp. 108-110, April 7 2006.

Gürerk, Ö., B. Irlenbusch and B. Rockenbach, 2005, "On the evolution of institution choice in social dilemmas," University of Erfurt, Working Paper.

Kreps, David, Paul Milgrom, John Roberts and Robert Wilson, 1982, "Rational Cooperation in Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory* 27: 245-52.

Masclot, David, Charles Noussair, Steven Tucker and Marie-Claire Villeval, 2003, "Monetary and Non-Monetary Punishment in the VCM," *American Economic Review* 93(1): 366-80.

Meltzer, Allan and Scott Richard, 1981, "A Rational Theory of the Size of Government," *Journal of Political Economy* 89 (5): 914-27.

Nikiforakis, Nikos, 2004, "Punishment and Counter-punishment in Public Goods Games: Can we Still Govern Ourselves?" unpublished paper, Royal Holloway University of London.

Ones, Umut and Louis Putterman, forthcoming, "The Ecology of Collective Action: A Public Goods and Sanctions Experiment with Controlled Group Formation," *Journal of Economic Behavior and Organization* (in press).

Ordeshook, Peter, 1986, *Game Theory and Political Theory: An Introduction*, Cambridge; New York: Cambridge University Press.

Ostrom, Elinor, James Walker and Roy Gardner. 1992, "Covenants with and without a Sword: Self Governance is Possible." *American Political Science Review*. 86 (2): 404-416.

Page, Talbot, Louis Putterman and Bulent Unel, 2005, "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency," *Economic Journal* 115: 1032-53.

Sefton, Martin, Robert Shupp and James Walker, 2002, "The Effect of Rewards and Sanctions in Provision of Public Goods," Working Paper, University of Nottingham and Indiana University.

Sutter, Matthias, Stefan Haigner, and Martin Kocher, 2005, "Choosing the stick or the carrot? – Endogenous institutional choice in social dilemma situations" unpublished paper, University of Cologne, University of Innsbruck and University of Amsterdam.

Appendix

Figure A is the screen design for an individual to enter her contribution to the group account (box a), to learning of others' contributions (boxes b, c, and d), to enter her punishment decisions (boxes b', c', and d'), and to observe the computer's calculation of net earnings for a round.

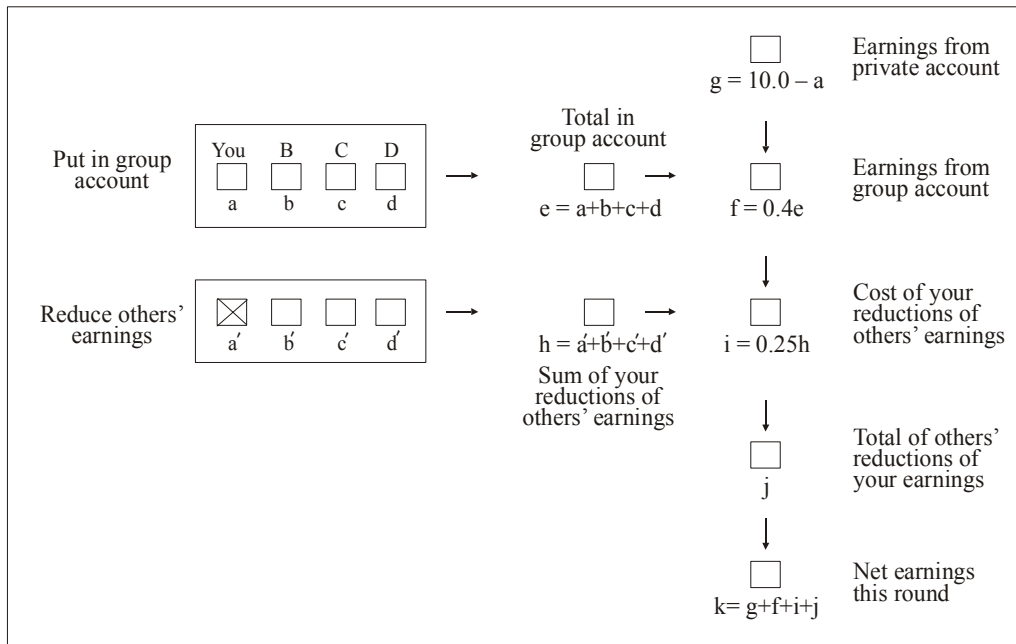


FIGURE A. SCREEN DESIGN FOR ENTERING CONTRIBUTION AND PUNISHMENT DECISIONS, RECEIVING INFORMATION, AND CALCULATING NET EARNINGS