On Perverse and Second-Order Punishment in Public Goods Experiments with Decentralized Sanctioning

by Matthias Cinyabuguma*, Talbot Page** and Louis Putterman**

Abstract

The fact that many people take it upon themselves to impose costly punishment on free riders helps to explain why collective action sometimes succeeds despite the prediction of received theory. But while individually imposed sanctions lead to higher contributions in public goods experiments, there is usually little or no net efficiency gain from them, because punishment is costly and at times misdirected. We document the frequency and probable causes of punishment of high contributors in several recent studies, and we report a new experiment which shows that introducing higher-order punishment opportunities offer a partial solution to the problem, but also reveal the deep-seatedness of retaliatory tendencies.

JEL numbers: C91, H41, D71

Keywords: Public goods, collective action, experiment, punishment, demand.

* Candidate for the Ph.D. in Economics, Brown University
** Professor of Economics, Brown University

* The research reported here was supported by National Science Foundation grant SES-0001769 and by university funds administered by the Brown University Department of Economics. We thank Simon Gächter for providing the data from Fehr and Gächter (2000), and Xiaotong Wang and Ioannis Garos for their role in early stages of the data analysis.

On Perverse and Second-Order Punishment in Public Goods Experiments with Decentralized Sanctioning

by Talbot Page, Louis Putterman, Theodore Marr and Matthias Cinyabuguma

> "In the state of nature there often wants power to back and support the sentence when right ... resistance many times makes the punishment dangerous, and frequently destructive to those who attempt it." John Locke, Second Treatise on Civil Government.[1]

0. Introduction

A dozen undergraduates take up residence in an old house a quarter mile from campus and try to organize themselves for joint housekeeping. Each of them is to pitch in to clean the kitchen and bathrooms, sweep the hallways, and take the trash out on time. It soon becomes clear that some are more civic-minded than others. Among the conscientious members of the group, a few are particularly inclined to take it upon themselves to criticize those who fail to do their parts in a proper and timely manner. These critics are quietly supported by most of the others and successfully prod some laggards into action. But they provoke so much hostility from one or two die-hard procrastinators that they find it better to back off somewhat. In the end, an equilibrium of moderate effort mixed with tolerated laxness emerges.

This story, which echoes experiences that most of us have had as members of a family, a committee, a work team, or an organization, points to an aspect of otherwise much-studied collective action problems that has captured little formal attention from economists. In this paper, we argue that the interplay between enforcement and resistance prevents many group efforts from achieving maximum efficiency. And we demonstrate, with evidence from voluntary contribution mechanism (VCM) experiments, that in accounting for the variation of behavior among individuals which marks most groups, it is important to take into account the existence of individuals inclined to act not just self-interestedly but in a decidedly perverse fashion.

In particular, we show that perverse punishment of high contributors is a major reason why the recent VCM experiments in which subjects have the opportunity to

[1] P. 185 in Somerville and Santoni, eds.

sanction one another fail to increase efficiency, though they raise contribution levels. And we introduce a new experiment in which subjects can punish punishment itself. In this experiment, we show that direct retaliation can be curbed, and efficiency be increased, by allowing higher order punishment. However, we also find evidence that the tendency underlying perverse behavior is quite deep-seated: retaliatory action migrates from one level of the experiment to another.

The paper is organized as follows. In Section 1, we introduce the recent experimental literature on cooperation and punishment, discuss the importance of heterogeneity of agent types for understanding observed behaviors, and set out the goals for this paper's analysis. In Section 2, we review contribution trends in ordinary VCM experiments and in VCM experiments with targeted punishment opportunities, and we focus new attention on earnings and efficiency differences between the two types of experiments. In Section 3, we re-analyze data from extant experiments to explore the frequency and consequences, and in Section 4 the causes, of what we will call "perverse punishment." In Section 5, we present the design of a new experiment in which subjects have recourse to a higher order of punishment, to control retaliation, and in Section 6, we present its results. Section 6 concludes the paper.

1. Cooperation, punishment, and agent heterogeneity

Standard economics and game theory, which assume that individuals are rational and that each seeks to maximize her material pay-offs, predicts that no one will contribute to the provision of a public good when the private cost of doing so exceeds the private benefit. Numerous experimental tests of this hypothesis use the VCM, a linear *n*-person dilemma game where subjects are assigned to groups and each is asked to divide her endowment between group and personal accounts. Although the socially efficient outcome is attained when all contribute their full endowment to the group account, individuals maximize their payoff by retaining their endowments, regardless of what others do. In trials, subjects typically contribute an average of over 50% of their endowments in one-shot play or in the initial round of repeated play. Contributions then tend to decay toward zero if play is repeated (Ledyard, 1995; Davis and Holt, 1993), although levels of 10 to 15% in the pre-announced final period remain common. While

the initial contributions contradict the standard theory interpreted strictly, the decaying trend is consistent with the theory if learning takes time.

Fehr and Gächter (2000a, hereafter FG) demonstrated that the frequently replicated result of decaying contributions can be reversed by allowing subjects to direct costly monetary punishments at other group members after learning of their contributions. Fehr and Gächter's qualitative findings that many subjects engage in costly punishment, that such punishment is aimed mainly at low contributors, and that contributions accordingly stabilize or rise, have been confirmed by a number of subsequent studies including Bochet, Page and Putterman (forthcoming, hereafter BPP), Carpenter and Matthews (2002), Falk, Fehr and Fischbacher (2001), Fehr and Gächter (2002), Masclet, Noussair, Tucker and Villeval (2003), Page, Putterman and Unel (2003, hereafter PPU), and Sefton, Shupp and Walker (2002). Fehr and Gächter (2000b) interpret the evidence as suggesting that substantial numbers of individuals have a propensity to punish free riding, perhaps because it violates a norm of, or a predisposition towards, reciprocity, or because it triggers the negative reciprocation of behaviors perceived as being exploitative.

In insightful discussions, Fehr and Gächter (2000a, 2000b) suggest that the decay in contributions that is typical in the VCM without punishment does not result from learning by the "typical" member of a uni-modal population, but instead is the consequence of the clash between two types: conditional cooperators, who are willing to contribute provided they see others do so, and free riders or strictly self-interested individuals. Without punishment opportunities, they point out, conditional cooperators who initially make large contributions, testing the waters to see if others follow suit, have no way to defend themselves from being exploited by free riders. Upon encountering free riders, who are likely to be found in most groups, they gradually reduce their contributions both to reduce their losses and to signal their disapproval. When a new degree of freedom is added by introducing a punishment option, conditional cooperators can signal disapproval and punish free riding without reducing their contributions. This leads self-interested free riders to adjust by raising their contributions.

We agree with Fehr and Gächter that reciprocity is a common element of human behavior that alters the predictions of game theory in important ways.[2] However, we propose an addition to the heterogeneity of behaviors to which they allude. In particular, our re-analysis of the original FG data and of the data from several of our own experiments indicate that in addition to pay-off maximizers and pro-social reciprocators willing to punish free riders, most subject pools also contain a few individuals so aggressively anti-social as to significantly alter outcomes for randomly formed groups. When the enforcement of social rules is put in individual hands, John Locke pointed out over three centuries ago, there is a likelihood that some will seek revenge, and this makes it more dangerous and costly to undertake enforcement of social norms on an individual basis. We demonstrate that introducing punishment opportunities in VCMs usually raises contributions, but not efficiency. And we show that a major reason for this is that a significant amount of punishment is aimed at high contributors, and that this tends to prevent contributions rising to their fully efficient level.

2. Contributions and Earnings in Public Goods Experiments with and without Punishment

The baseline and punishment treatments of the experiments of FG, BPP, and PPU, on which we focus in this section and the next, are largely alike. In each case, subjects were anonymously and randomly placed in groups of four in sessions having enough participants to support several such groups. In baseline conditions, each subject's task was to select an integer number of units of an endowment to contribute to a group account. Subject *i*'s earnings for the period were given by

$$y_i = (E - x_i) + 0.4 \, \Sigma_j x_j \qquad (1)$$

Here E is the endowment received each period ($E = 20$ in FG, $E = 10$ in BPP and PPU), x_i is the amount that *i* put in the group account, and the summation is taken over all members of *i*'s group, *i* included. The number of periods is finite and known in advance: 10 or 6 in FG and BPP, 20 in PPU. In punishment conditions, there was a second stage of each period in which a subject could choose at some cost to himself to reduce the

[2] There are many other useful discussions of reciprocity in the recent literature; see, for example, Hoffman, McCabe and Smith (1998).

earnings of one or more of the other group members, having information only about the amount each contributed in the current period. Instructions to subjects did not use the term punish or suggest any reason for reducing earnings, or whose earnings to reduce. The two stage process was repeated 10 times in FG and BPP, 20 times in PPU.

A small difference between the FG experiments, on the one hand, and those of BPP and PPU, on the other, lies in the details of punishment cost. In FG, *i*'s earnings in the punishment condition are given by

$$y_i = (E - x_i) + 0.4\ \Sigma_{j=1}^{3} x_j\ \}[1 - (0.1)\Sigma_{j\neq i} r_{ji}] - \Sigma_{j\neq i} C(r_{ij}) \qquad (2)$$

where r_{ij} are the number of punishment points that *i* gave to *j*, and conversely for r_{ji}. This means that for every punishment point received, *i* loses 10% of his pre-punishment earnings for the period. The cost of punishment to *i*, $C(r_{ij})$, increased at an increasing rate with the number of points given to the individual in question, i.e. r_{ij}.[3] In BPP and PPU, by contrast, there were constant costs of 0.25 to the punisher and 1.0 to the recipient of punishment, for each unit of punishment given. Hence, income for a period with punishment option is given by

$$y_i = (E - x_i) + 0.4\ \Sigma_j x_j - .25\Sigma_{j\neq i} r_{ij} - \Sigma_{j\neq i} r_{ji}\ [4] \qquad (2')$$

The structure of the experiments also differs slightly. In FG, subjects in a given session played both ten (or six) periods in the baseline condition and ten (or six) periods in the punishment condition, with roughly equal numbers of sessions being characterized by each of the possible orders (baseline followed by punishment, or punishment followed by baseline). In addition, different matching protocols were used in different sets of sessions. In the partner treatment, the same group of four played together the entire session, although with identifying letters randomly reassigned each period to prevent continuing identification of individuals. In the stranger treatment, subjects were randomly reassigned to new groups each period. Both treatments involved ten periods of play in each condition. In the perfect stranger treatment, subjects were reassigned each

[3] In particular, the relationship between C and r is given by:

r	0	1	2	3	4	5	6	7	8	9	10
C	0	1	2	4	6	9	12	16	20	25	30

[4] FG's subjects earned money in imaginary units called "francs" and BPP and PPU's in units called "experimental dollars." After conversion to real money and addition of show-up fees, FG's subjects earned an average of $34 and BPP and PPU's an average of $25 for two hour and 100 minute sessions, respectively.

period in such a way that no two subjects would be in the same group twice. In this treatment, subjects played only six periods of each condition due to limitations imposed by the size of the subject pool.

In BPP and PPU, subjects in a given session played a baseline game or a game with punishment option but not both, so that baseline and punishment outcomes must be compared across subjects. The treatments from those experiments that we analyze here involved partner groups, only. Because we consider only the partner experiments with (first order) punishment stage in BPP and PPU, because these treatments differ only with respect to the number of rounds (10 versus 20), and to distinguish them from experiments with second order punishment, to be discussed later, we will henceforth refer to the BPP and PPU punishment treatments (without communication or endogenous group formation) as 1-Ord-10 and 1-Ord-20, respectively.[5]

FG found that contributions stayed roughly level, averaging about 60% of endowment, in the ten periods of the punishment condition in their stranger and perfect stranger treatments, and rose with repetition, from an average of about 60% to an average of about 90% of endowment, in the punishment condition of their partner treatment. In the 1-Ord-10 and 1-Ord-20 treatments of BPP and PPU, contributions remained roughly level, at about 70% of endowment, with a drop off to about 60% in the last period. In all three of FG's treatments and when comparing the no punishment baseline sessions in BPP and PPU to 1-Ord-10 and 1-Ord-20, respectively, contributions began at lower levels in baseline than in punishment conditions, and in baseline conditions they declined with repetition, as in other VCM experiments.

That average contributions were higher even in the first period of a punishment condition than in the first period of baseline play indicates that at least some subjects anticipated, before any punishment occurred, that they might be punished if they did not contribute enough. But it was not only the *anticipation* of punishment that raised and sustained higher contributions: most subjects *actually* punished at least once in 1-Ord-10, 1-Ord-20, and all three treatments of FG. If the *threat* of punishment alone had been enough to deter free riding, subjects would have enjoyed higher earnings when there was

[5] Other treatments in BPP, not discussed here, permit various types of communication among the subjects, while other treatments discussed in PPU permit self-sorting of subjects and new group formation.

a punishment option, since there were significantly higher contributions, which implies higher earnings by (1). However, earnings were not on average higher with punishment in any of the experiments.

FG report that subjects earned less in punishment than in baseline conditions in early periods and more in punishment than in baseline conditions in later periods. They do not report an overall result. Accordingly, we had to calculate average earnings over the ten (six) periods of baseline play and average earnings over the ten (six) periods of punishment play for each of their treatments. We found that overall average earnings were higher in the baseline than in the punishment condition in all three: 24.48 francs versus 23.74 francs in the partner treatment, 22.24 francs versus 18.98 francs in the stranger treatment, and 21.96 francs versus 16.69 francs in the perfect stranger treatment.

In our 1-Ord-10 treatment, average earnings were also higher in the baseline than in the punishment treatment in early periods, with this ordering reversing in periods 6 and 8-10. However, earnings for the ten periods as a whole were lower, at 12.52 experimental dollars (78.3% of maximum efficiency) in 1-Ord-10 than in the 10 period baseline treatment of BPP, where they were 12.85 (80.3% of maximum). Looking only at the rank order of earnings in the 12 baseline groups and 12 punishment groups, a Mann-Whitney test finds no statistically significant difference in earnings.

In the 20 period experiments of PPU, average earnings per period were slightly higher in 1-Ord-20, at 12.9 experimental dollars (81% of maximum), than in the baseline treatment, at 12.3 experimental dollars (77% of maximum). However, a parallel Mann-Whitney test comparing earnings in the 16 baseline groups with those in the 16 punishment groups also finds no statistically significant difference.

All treatments discussed here share payoff structures such that total and per subject earnings are at a maximum when all subjects contribute their full endowments to the group account every period and there are no expenditures on punishment. In BPP and PPU (FG), this would yield incomes of 16 (32) experimental dollars (francs) per person per period, whereas subjects would earn only 10 (20) per period if no one contributed. Focusing on BPP, we find that in its baseline treatment, subjects contributed an average of 4.75 experimental dollars per period to the group account, and therefore earned an average of $(4.75 \times 1.6) + (10 - 4.75) = 7.6 + 5.25 = 12.85$. In BPP's punishment treatment

(1-Ord-10), subjects contributed an average of 6.93 per period, so they would have earned an average of (6.93x1.6) + (10 – 6.93) = 11.088 + 3.07 ≅ 14.16 per period but for the deduction of punishment costs. That average earnings were in fact 12.52 per period in 1-Ord-10 thus results from average punishment costs of 1.64 per period, of which 33 cents are costs of punishing and 1.31 are costs of being punished. Subjects chose a punishment greater than zero in 14.5% of their interactions with other subjects, punishing at least one group member in 45.5% of rounds, over the course of this experiment,[6] with the average punishment amount being $3.01 experimental dollars. Had it proven possible to boost contributions from the average of 4.75 characterizing the baseline treatment to the 6.93 of the punishment treatment using 20% less punishment than in fact took place, then earnings would have matched those of the baseline treatment. Had *more* than 20% of the observed punishment been avoided, punishment treatment earnings would have been the higher of the two. We demonstrate in the next section that punishment of high contributors is of roughly this magnitude, and that if its perverse effects on contributions are considered even using a conservative estimate, it suffices to explain why earnings with punishment did not exceed those without.

3. Perverse punishment

FG, BPP, and PPU show that most punishment in their experiments was directed at group members contributing less than their group's average to the group account. To demonstrate this directedness of punishment, FG estimate a regression in which the number of punishment points received by subject *j* is the dependent variable, and the independent variables are the average contribution of the others in *j*'s group, the positive deviation of *j*'s contribution from this average (set to zero if *j* contributed less than the average), and the negative deviation of *j*'s contribution from the average (set to zero if *j* contributed more than the average). We estimated parallel regressions for 1-Ord-10 and 1-Ord-20, and obtained qualitatively similar results, as seen in Table 1.

[6] Each subject has three interactions per period, so if a subject punished one out of the three other group members each period, she would be punishing in 1/3 of her interactions. Thus the 14.5% figure would result if each subject punished one other subject during an average of 45.5% of all rounds played, a fairly high frequency of punishing.

Table 1: OLS Regressions, First Order Punishment Received as a Function of Deviation of Recipient's Contribution from Others' and Average Contribution Level

	FG[a]	1-ord-10	1-ord-20
Constant	0.988	1.599 ***	0.934 ***
(standard error)	0.680	0.360	0.213
Positive Deviation	-0.036	-0.111	0.038
(standard error)	0.036	0.065	0.038
Negative (absolute) Deviation	0.417 ***	0.602 ***	0.593 ***
(standard error)	0.051	0.050	0.028
Average Contribution	-0.011	-0.115 **	-0.073 **
(standard error)	0.046	0.043	0.024
Number of observations	400	480	1280
Adjusted R-squared	.68	.284	.286

* indicates significance at the .05 level
** indicates significance at the .01 level
*** indicates significance at the .001 level
a Source: Fehr and Gachter, (2000a), Table 5, partner treatment

When costly punishment is directed at free riders in a repeated VCM, there is at least the possibility that it will contribute to efficiency by inducing increases in contributions large enough to offset its cost. The same cannot be said of punishment of high contributors. A careful analysis of the data from FG, 1-Ord-10 and 1-Ord-20 indicates that the latter occurs frequently enough to account for the failure of the punishment option to generate efficiency gains. We will call targeted reductions of others' earnings "perverse" if they tend to reduce overall efficiency by inducing declines rather than increases in contributions to the public good. Punishment of a subject who has just contributed the maximum amount possible seems unlikely to encourage continued high contributions. To test this idea, we estimated regression equations in which the dependent variable is the change in subject *i*'s contribution from one period, *t*, to the next, $t + 1$, and the independent variables are the total amount of punishment received by *i* in period *t* interacted with a set of dummy variables for four possible conditions: condition 1, *i*'s contribution to her group's account was the smallest in the group in period *t*; condition 2, i's contribution was the second smallest; condition 3, *i*'s contribution was the second largest; and condition 4, *i*'s contribution was the largest in

the group.[7] Although only one dummy variable can be positive for a given subject in a given period, observations on many subjects in many periods permit us to estimate the sign and magnitude of the impact of punishment on contributions under each condition.[8] Table 2 shows the estimates for the 1-Ord-10, 1-Ord-20, and FG experiments. These results are remarkably consistent, with negative significant coefficients for amount of punishment if *i* was top contributor, and positive significant coefficients for amount of punishment if *i* was a lowest or next-to-lowest contributor.[9] The insignificant coefficients for next-to-highest contributors suggest that there was no consistent pattern of responses to punishment in that case, presumably because some took punishment as a warning to contribute more and others as a sign that their higher-than-average contributions were not appreciated.

Table 2: Change in Contribution in Response to Punishment

Variable	1-Ord-10	1-Ord-20	FG (partner)	FG (stranger)
r_{ji} of Top Contributor	-0.544 ***	-0.536 ***	-1.163 ***	-1.053***
	0.151	0.106	0.229	0.259
r_{ji} of Next to Top Contributor	0.252	0.022	0.182	0.492**
	0.491	0.195	0.242	0.223
r_{ji} of Next to Bottom Contributor†	0.315 *	0.427 ***	1.053 ***	0.852***
	0.124	0.101	0.194	0.131
r_{ji} of Bottom Contributor	0.543 ***	0.643 ***	1.049 ***	1.093***
	0.058	0.050	0.088	0.098
(Constant)	-0.486 ***	-0.479 ***	-1.111 ***	-1.053***
	0.139	0.133	0.259	0.259
Number of observations	431	824	648	648
Adjusted R-squared	.205	.208	.237	.234

†Where only three levels of contribution were observed, the medium level was coded as a medium-low contribution
* significant at the .05 level

[7] If two subjects were tied for highest contributor in a certain period, both will be treated as highest contributor; the same applies for lowest contributor. If a group exhibits only three contribution levels in a period, the middle contributor is treated as second lowest. If there is only one level, as happened on a few occasions, in the BPP regression we treat each group member as a highest contributor, if that level is 9 or 10, and as lowest contributor if it is 0 (the only cases observed). In the regression for FG's data, we discarded the few cases of ties.

[8] Because the dummy variables are interacted with the amount of punishment, which varies across observations, it is not necessary to have an omitted category.

[9] That the magnitudes of the significant coefficients are about twice as high for the FG as for the BPP and PPU data actually underscores the consistency of the results since per period endowments were twice as large in experimental currently units (20) in FG's as in the other (10) two experiments.

** significant at the .01 level
*** significant at the .001 level

Based on the evidence in Table 2, we conclude that when punishment is aimed at a group's highest contributor, it has the perverse effect of discouraging contributions to the public good, thereby reducing efficiency. We define perverse punishment operationally, then, as punishment of a highest contributor. (When a description seems helpful, we refer to punishment that has the opposite effect, that of encouraging contributions to the public good, as "pro-social.")

Table3: Perverse punishment incidence in five BPP, PPU and FG treatments.

	1-Ord-10 (BPP)	1-Ord-20 (PPU)	FG-partner	FG-stranger	FG-perfect stranger
Perverse punishment E$ or points per period	$0.24	$0.21	1.01	0.32	0.46
Share of punishment E$ or points that are perverse	18%	18%	32%	8.5%	12.5%
Share of punishment Events that are perverse	28%	25%	35.3%	8.0%	13.6%

How pervasive was perverse punishment in these experiments? Table 3 shows the result of our calculations, the first row indicating the average number of punishment points (FG) or dollars of punishment (1-Ord-10 and 1-Ord-20) directed against a highest contributor per period, the second row the average proportion of punishment that was directed against a highest contributor. Punishment against highest contributors accounts for more than 18% of punishment points or dollars, and for more than 25% of punishment events, in the three partner treatments (1-Ord-10, 1-Ord-20, and FG-partner) and for smaller but still significant shares of punishment in the two FG stranger treatments.[10]

[10] It strikes us as noteworthy that about 1/3 of punishment in FG's partner treatment was directed not just at above-average contributors, but at those contributing the highest amount in their group in the period in question. Clearly perverse punishment is not only an artifact of our American subject population; it was at least as prominent in the comparable treatment of FG. The substantially lower shares of perverse punishment in FG's stranger treatments is consistent with most punishment being an attempt to retaliate for

Notice that the average of 18.1% of punishment that was perverse in 1-Ord-10, for example, is only a little shy of being enough (the 20% mentioned above) to directly explain the lower earnings in the punishment compared to the baseline condition of that experiment. However, looking at the direct cost of perverse punishment understates its negative impact on efficiency, because it fails to include the negative effect on contributions demonstrated in Table 2. A conservative estimate of that effect can be made using the assumption that when perverse punishment reduces a targeted subject's next-period contribution, the effect lasts for that one period only. With this assumption, we can calculate the per subject cost of perverse punishment in a given period in 1-Ord-10 and 1-Ord-20 as

$$\text{Cost of Perverse Punishment} = (0.25)*(\$PP) + \$PP + (0.6)(\$PP)(MEP) \qquad (3)$$

where \$PP is the average number of dollars of punishment aimed at highest contributors in the period divided by number of subjects, MEP is the marginal effect of punishment on contributions, 0.25 is the cost to the punisher, and 0.6 represents the loss of total earnings from a \$1 decline in contribution. In the last period of play, the term containing MEP disappears, since perverse punishment can have no further effect on contributions. Substituting for MEP the estimated coefficient on punishment aimed at highest contributors from Table 2, we get

$$(.25)*(\$114) + \$114 + (0.6)(\$106)(0.544) = \$177.10 \text{ in 1-Ord-10} \qquad (3a)$$

and

$$(.25)*(\$261) + \$261 + (0.6)(\$251)(.536) = \$406.97 \text{ in 1-Ord-20} \qquad (3b)$$

or in terms of cost per period per subject, \$.45 in 1-Ord-10 and \$.32 in 1-Ord-20.

In 1-Ord-10, average earnings in the treatment with punishment (1-Ord-10) were lower than in the baseline treatment by 2.6%. If the costs of perverse punishment are negated using the conservative estimate in (3a), then 1-Ord-10 earnings are *higher* than in the baseline treatment, by 0.3%.[11] For 1-Ord-20, average earnings were already 5%

---

being punished—see the next section. However, the larger shares in the perfect stranger than in the stranger treatment are contrary to what would be expected, if perverse punishment were only for retaliation and subjects properly understood that they could never meet the same individual twice. Anderson and Putterman (forthcoming) get still higher proportions of perverse punishment in their perfect stranger VCM with varying punishment costs treatment.

[11] Neither the difference before nor the difference after this adjustment for perverse punishment is statistically significant, according to a Mann-Whitney test. Of course, the conclusion that earnings would

higher with punishment (1-Ord-20) than without punishment, but a Mann-Whitney test at group level showed no significant difference. With perverse punishment costs deducted, average punishment treatment earnings are 7.5% higher than those in the baseline treatment, and the difference is now statistically significant at the 10% level in a Mann-Whitney test.

4. What accounts for perverse punishment?

Why would subjects in a VCM punish group members who contribute to the public good when they themselves benefit from those contributions? Recall, first, that the instructions given to subjects in these experiments made no mention of punishment as such, and left it entirely up to subjects to decide whether to use the opportunity to reduce a group member's earnings, by how much, and when. Some instances of perverse punishment may have resulted from simple confusion, and others from subjects' desires to amuse themselves by undertaking secretive, malicious behavior at a modest monetary cost. Also, at least one subject[12] cast the problem—mistakenly, we would say—as one of cooperation *with the experimenter*, and he punished high contributors to express disdain for their "goody goody" behaviors. But the more common causes of perverse punishment fit into two categories. First, some subjects punished both high and low contributors in order to increase their own *relative* pay-off, thus acting out of "spite" in the sense of Saijo and Nakamura (1995) and Falk, Fehr, and Fischbacher, (2001). Second, some subjects who were themselves punished, usually for contributing little, punished high contributors in an attempt at retaliation or to dissuade them from punishing in the future. Subjects were not told who in particular had punished them, so these attempts at retaliation fit Ostrom *et al*.'s (1992) characterization as "blind revenge."

To get a sense of the importance of the various reasons for punishing top contributors, we count as spiteful those perverse punishment cases in which a subject

---

have been higher with than without punishment, but for the presence of perverse punishment, could be strengthened by adopting a less conservative assumption about the persistence of the downward contribution changes due to perverse punishment.

[12] According to his debriefing statement.

simultaneously punished both high and low contributors in his/her group,[13] and as retaliatory or dissuasion those perverse punishment cases in which the punisher of a group's highest contributor in a period, say *t*, was herself punished in period *t* – 1. Table 4 shows our results for 1-Ord-10 and 1-Ord-20, with row 1 reporting the proportion of perverse punishment satisfying the criterion for revenge, row 2 the proportion of perverse punishment characterized by spite, row 3 the proportion for which the two criteria overlap—that is, cases in which a subject was punished in the previous period and proceeded to punish high and low contributors alike—and row 4 the proportion of PP events that satisfy neither criterion. Both proportion of dollars of punishment and proportion of instances of punishment (events) are shown. The results indicate that a substantial majority of dollars of punishment aimed at top contributors can be explained by a revenge or dissuasion motive, with more than half of the revenge attempts being aimed at high and low contributors alike. Spiteful punishment not likely attributable to revenge accounts for only 9 to 14% of perverse punishment dollars, with 13 to 21% of perverse punishment dollars being accounted for by neither factor, hence possibly the results of confusion, resentment, and amusement.

Table 4: Breakdown of Perverse Punishment by Cause	1st Ord-10		1st Ord- 20	
	Dollars	Events	Dollars	Events
Revenge	73%	52%	70%	57%
Spite	53%	48%	37%	52%
Both Spite and Revenge	39%	41%	28%	35%
Neither Spite nor Revenge	13%	41%	21%	26%

To what extent was retaliatory punishment an emotional response carried out for the satisfaction of getting revenge, to what extent was it aimed at increasing the punisher's monetary payoff by dissuading a suspected punisher from punishing again? This question can be answered by comparing behaviors in earlier periods to those in the final period in partner treatments.[14] A Mann-Whitney test finds less perverse punishment

[13] By low contributor, we here mean a subject who contributed less than the group average amount in the period, averaging over all group members. To count as perverse, punishment must be directed at the group's highest contributor, as before.

[14] Table 3 shows that perverse punishment was less common in FG's stranger and perfect stranger treatments than in their partner treatment and in 1-Ord-10 and 1-Ord-20, also partner treatments.

in the last period than in other periods of 1-Ord-20, significant at the 5% level, but no such difference is found in 1-Ord-10. There is no difference in the amount of punishment given to *low* contributors in the last versus other periods in these experiments. The latter finding supports the idea that punishing low contributors is not mainly strategically motivated,[15] but the former finding carries a mixed message, since it suggests that in at least one of the treatments studied a significant part, but not all, of perverse punishment may have been strategically motivated.

5. Taming the Perverse Punisher with Second-Order Sanctions: Experimental Design

The tendency of some individuals to retaliate when punished for free-riding and of others to use punishment to raise their relative earnings makes giving each individual the unilateral power to sanction an inefficient if not an entirely ineffective instrument for promoting cooperation. Perverse punishment is reduced in FG when individuals interact once only, but constantly changing group membership is infeasible or unattractive in collective action environments like firms, villages, and civic associations. How might the willingness of so many to punish free riding be harnessed without unleashing the perverse actions of a small minority?

In real-world settings, informal discipline may emerge through the establishment of norms of cooperative behavior, including enforcement actions. If most group members see their goal as mobilizing collective effort and agree that it is legitimate to sanction non-cooperators but not others, then those who take it upon themselves to sanction non-cooperators may expect the support and approval of others, whereas those who engage in retaliatory and other perverse punishment may anticipate criticism and perhaps even punishment. Such a dynamic cannot emerge in FG and the experiments

---

Somewhat at odds with this is the finding by Anderson and Putterman (2003) that in their perfect stranger VCM with punishment experiments, in which the cost of punishing varied from one period to another, fully 23% of punishment was aimed at highest contributors--although in groups of three rather than four, which makes highest contributors a larger share of each group's population. In any case, perfect stranger (and stranger) treatments eliminate (or attenuate) not only the prospect of influencing to one's advantage a team members' future actions, but also the possibility of revenge, which requires striking back in the period *after* one received punishment oneself, at which point one's punisher is (most likely) no longer in one's group at all. So comparing perverse punishment in perfect stranger and partner treatments may help to separate retaliatory and dissuasive perverse punishment from other forms of perverse punishment, but it is not a promising way to distinguish between the retaliation and the dissuasion motives themselves.

[15] See also Falk, Fehr and Fischbacher (2001), who find that the proportion of punishment of free riders that can be explained by strategic considerations is negligible

patterned on it, because subjects are given limited information about who punishes whom and because even this history is lost due to identities being scrambled each period. In such an information environment, a group member can sanction a high contributor knowing that the act is observed by the person targeted, alone, so that even if others wanted to punish it, they can do so only by striking blindly.

To explore an environment in which cooperative norms, including punishment of anti-social punishers, might emerge, we conducted a new experiment in which subjects were periodically given information about one anothers' punishing rather than contributing behaviors, and were allowed to engage in a round of punishment while presented with that information alone. In particular, we ran partner treatment VCMs with a first stage of contributions and a second stage of punishment identical to 1-Ord-10 and 1-Ord-20 (hence, similar to FG as well). In the new treatments, after every third period (one period consisting of the two stages just mentioned) subjects saw a list of the amounts by which each had reduced the earnings of below-average contributors, of average contributors, and of above-average contributors,[16] and subjects could again engage in costly punishment—as before, at a cost of .25 to the punisher per 1.0 to the person targeted.[17] We wanted to see whether adding this second-order punishment stage to the experiment would curb perverse punishment of high contributors to the public good.

We conducted two versions of this second-order punishment experiment. In the first, which we'll call 2-Ord-OG (for "own group"), subjects saw, at the second-order punishment stage, only the punishment behaviors of members of their own group of four, and could impose costly punishment on any or all of their three fellow members. In the second variant, which we'll call 2-Ord-FS (for "full session"), subjects saw, at this stage, the punishment behaviors of all sixteen participants in their session, without identification of those in their own group. In this version, subjects could impose second- (but not first-)

[16] Here, contributing the average means contributing an amount equal to the average contributed by the others in one's group of four in the particular period in question.

[17] As can be seen in the appended instructions, two budget rules applied to second-order punishment: (a) an individual could spend no more on punishing others than he or she had earned net of first-order punishment during the previous three periods, and (b) a person targeted for punishment could not lose more than he had earned net of first-order punishment during the previous three periods. If the combined second-order punishment of several individuals violated constraint (b), all concerned had their chosen punishments adjusted downwards by the common proportion just sufficient to cause the constraint to be observed.

order punishment on any other subject. We did this to see whether the inability to identify one's own group members might alter punishment behavior. One conjecture of interest to us was that low contributors' fervor for revenge might be dampened by the need to expend money to punish three or four people to get at the individual one wanted to impose this cost upon, whereas "pro-social" subjects' desires to punish perverse punishers, being more "altruistic" in character, might persist even when much of the punishment would go to members of other groups.[18] In both variants, the experiment lasted a total of 18 periods, which allowed for six second-order punishment stages (following periods 3, 6, etc.).

A different motivation for studying second-order punishment comes from the theoretical discussions of Boyd *et al.* (2002), Boyd and Richerson (2002), Henrich and Boyd (2001), and Henrich (forthcoming). They ask how a propensity to punish free-riders could have emerged in the course of human genetic and cultural evolution, and they give a prominent place to higher-order punishment in their anwers. The crux of their theory is that the evolution of a propensity to punish free riders can be supported by a propensity to punish those who fail to join in punishing free-riders which can in turn be supported by a propensity to punish those who fail to punish those who fail to punish, and so on. What is important is that when a large enough number of individuals engage in first-order punishment of free riding, there is little such free riding, and it is therefore rarely necessary to punish it and even more rarely necessary to punish failure to punish free riding. The result is that being of the type with a tendency to engage in, say, 2nd order punishment, gives one only a small reproductive disadvantage. The authors show how this small individual disadvantage can be outweighed, under the right circumstance,

[18] This sort of behavior on the parts of pro-social players is exhibited with regard to first-order punishment in Carpenter and Matthews (2002). There, the opportunity to engage in first-order punishment of "out group" members was introduced in an otherwise standard VCM with punishment stage. In their experiment, two groups of 4 played a VCM in the same session, and subjects punish (after learning of contributions) both members of their own group and members of the other group. They found a substantial amount of punishment of free-riding aimed at members of the other group, enough so that earnings were significantly higher in the "all session" punishment treatment than in a control "own group only" treatment. Note that our 2-Ord-FS treatment differs from that of Carpenter and Matthews not only in that the opportunity for "out group" punishment arises only in the second-order punishment stage, but also in that our subjects could not tell which individuals belonged to their own group, whereas the two groups were clearly differentiated in the (first-order) punishment stage of their experiment.

by the advantage of being in a group with people inclined to punish, which is where punishers are more likely to find themselves.

Although second- and higher-order punishment are central to this argument, the concern of Boyd *et al*. is not with perverse punishment, which they never consider, but rather with "second order free riding"—that is, leaving it to others to carry out costly pro-social punishment. What would be observed in the second-order punishment stage of their models, then, is only the punishment, by those of the punishing type, of those who shirked on first-order punishment (but not on contributing, since almost everyone contributes, to avoid first-order punishment). It is of interest to see whether second-order punishment of non-punishers is common in our new experiment, as well as whether the introduction of a second-order punishment opportunity ameliorates the rather different problem that we have identified, that of perverse first-order punishment.

6. Second-order Punishment: Results

Four sessions of each variant of the second-order punishment experiment were conducted in a computer classroom at Brown University. 16 undergraduates drawn from the entire undergraduate population of the university participated in each session, for a total of 128 subjects.[19] In this section, we compare behaviors in experiments with second-order punishment (2-Ord-OG and 2-Ord-FS) to those in experiments with first-order punishment only (1-Ord-10 and 1-Ord-20), sometimes grouping each pair of treatments as a class, sometimes considering the results of each treatment separately.

Result 1: Subjects contributed and earned more in experiments with 2nd order punishment, considered as a class, with the earnings difference significant at the 5% level in a one-tailed test. Considering treatments separately, contributions and earnings in 2-Ord-OG exceeded those in 1-Ord-10 but not 1-Ord-20, and those in 2-Ord-FS are not significantly different from those of either first-order punishment treatment.

[19] Subjects were recruited by receiving flyers in their campus mailboxes or by reading an advertisement in an on-line magazine. They were promised a minimum payment of $5 and the possibility of more, with most likely outcomes falling in the $20 to $25 range. No subject participated twice in the same treatment, and most had not participated in an economics experiment before. At the end of the experiment, subjects received their accumulated earnings translated into real money at a rate of 8 cents to the experimental dollar, plus a show-up fee of $5. Earnings averaged about $25 for a 90 minute session.

Figure 1 shows the average contribution level in each period for the 2-Ord-OG and 2-Ord-FS treatments, and for comparison purposes, average contribution in the 1-Ord-10 and 1-Ord-20 treatments. The four treatments are similar, with initially high average contributions that are sustained with repetition, relative to VCM experiments without punishment. Despite noticeable late period declines, average contributions reamin well above the 10 to 20% of endowment common in the ordinary VCM. Contributions are somewhat higher in 2-Ord-OG, with those in 2-Ord-FG resembling those in 1-Ord-20, and those in 1-Ord-10 tending to be a little lower still. Mann-Whitney tests, using groups as units of analysis[20], find no statistically significant differences in average contribution levels between the four treatments except that contributions are higher in 2-Ord-OG than in 1-Ord-10, significant at the 5% level in a one-tailed test. When groups from both 1-Ord treatments are combined and compared with groups from both 2-Ord treatments, contributions are higher in the latter, but the difference is not significant ($p = 0.114$ in a one-tailed Mann-Whitney test).

[20] For each group, we compute the average contribution per period and per subject over the full duration of its session. The result is in per period terms, avoiding spurious differences due to having three different session lengths.

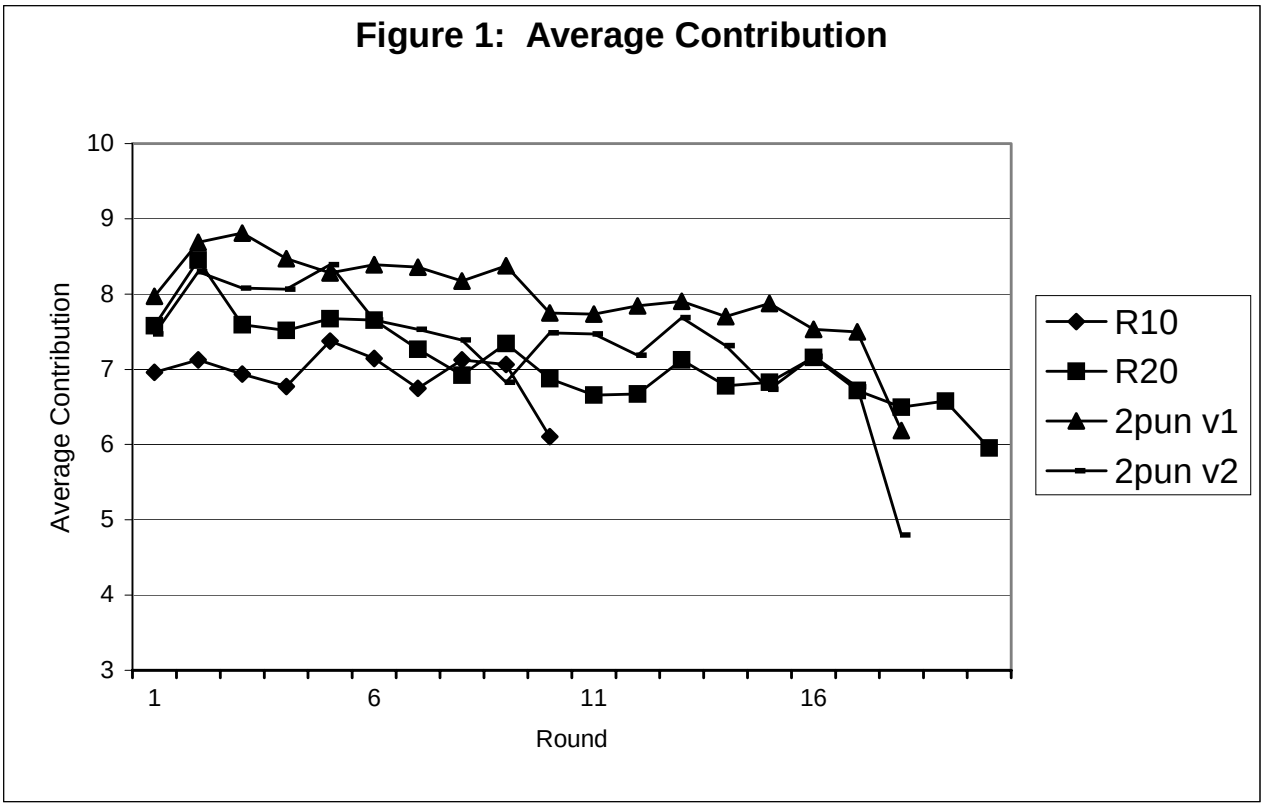

Figure 2 shows average earnings by period for the same four treatments. In 1-Ord-20, 2-Ord-OG and 2-Ord-FS, earnings appear to similarly fluctuate around a mean between 13 and 14 dollars, while 1-Ord-10 earnings initially decline, then rise gradually, without ever matching those in the other three treatments. Mann-Whitney tests comparing average earnings, again using groups as units of analysis, show no statistically significant differences among the treatments except that earnings in the 2-Ord-OG, 2-Ord-FS and 1-Ord-20 treatments are all higher than those in the 1-Ord-10 treatment, significant at the 1% level in a one-tailed test for 2-Ord-OG and at the 5% level in a one-tailed test for 2-Ord-FS and 1-Ord-20. When groups from both 1-Ord treatments are combined and compared with groups from both 2-Ord treatments, the Mann-Whitney test finds earnings significantly higher in the latter (p = .032 in a one-tailed test). Average contributions and earnings in the two 2nd order treatments and those in the two 1st order treatments can be seen at a glance in rows 1 and 8 of Table 5.

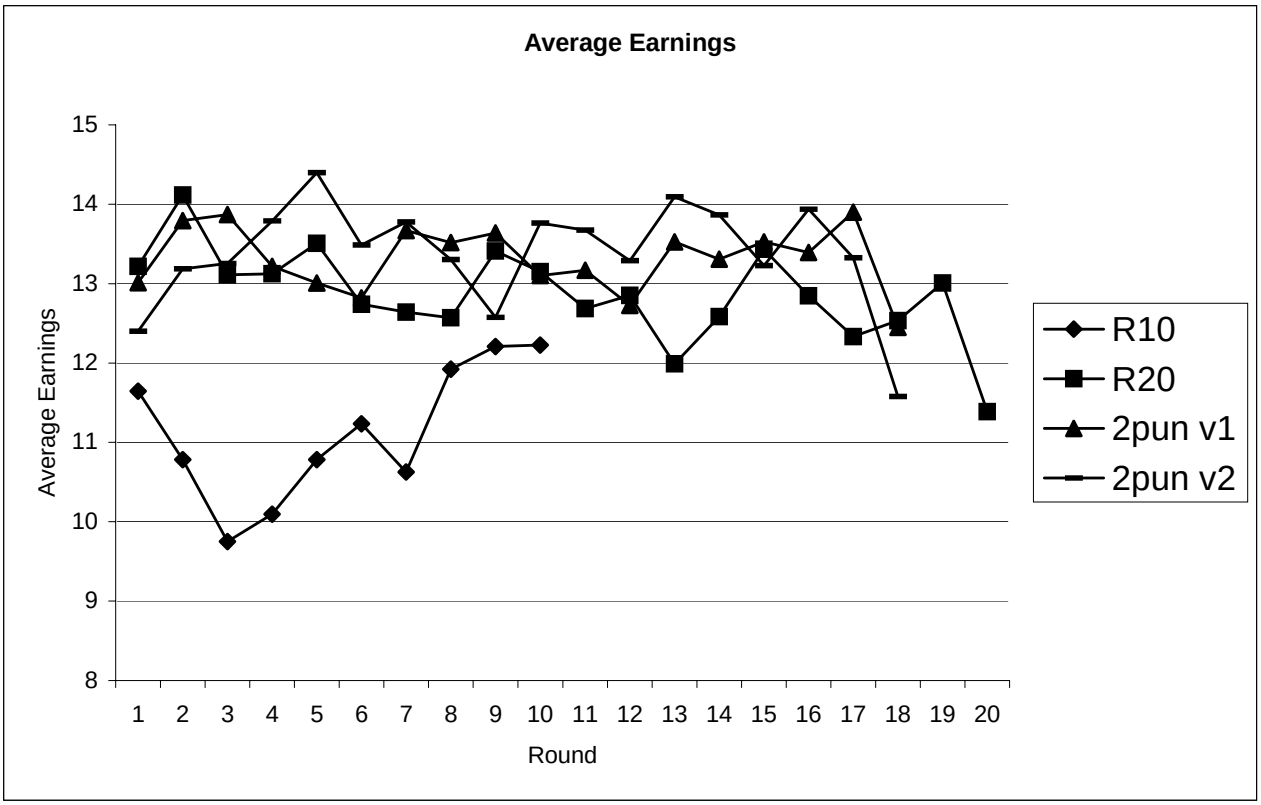

Figure 2

Table 5. Average Period Contribution, Punishment Costs, and Earnings

	1-Ord-10	1-Ord-20	2-Ord-OG	2-Ord-FS
1. Average Contribution	6.94	7.09	7.97	7.35
2. Amount Earned if no Punishment Cost	14.16	14.26	14.78	14.41
3. Average 1st Order Punishment Cost	1.64	1.39	0.96	0.74
4. % of Potential Earnings (2.) Lost	0.12	0.10	0.07	0.05
5. Average 2nd Order Punishment Cost	n.a.	n.a.	0.51	0.29
6. % of Potential Earnings (2.) Lost	0.00	0.00	0.03	0.02
7. Total % of (2.) Lost [(4) + (6)]	0.12	0.10	0.10	0.07
8. Average Earnings (Actual)	12.52	12.86	13.31	13.38

These comparisons of contributions and earnings suggest that informing subjects of one anothers' punishment behaviors and allowing them to further sanction each other based on that information may have been mildly successful at making sanctions more efficient. To learn more, we look at punishment behavior itself.

Result 2. A substantial number of subjects engaged in costly first- and second-order punishment in 2-Ord-OG and 2-Ord-FS. However, the amount of first- order punishment was less in these experiments than in 1-Ord-10 and 1-Ord-20. In fact, expenditure on first- and second-order punishment combined was less, on average, in the two 2-Ord treatments than in the two 1-Ord treatments .

75% of subjects in 2-Ord-OG and 81% of subjects in 2-Ord-FS used the opportunity to first-order punish at least one time in their sessions. 72% and 93% of subjects, respectively, were targeted for first-order punishment at least once in their sessions. In 2-Ord-OG, the typical subject first-order punished one or more others in 25% of the first-order punishment stages in her session. The corresponding ratio is 24% for 2-Ord-FS.

From Table 5, it can be seen that expenditure on first-order punishment is lower in both 2-Ord treatments than in 1-Ord-10 and 1-Ord-20, the average *total* punishment expenture per period being fully 44% lower in the 2-Ord than in the 1=Ord treatments. This lower punishment cost contributes, along with higher contributions, to the higher overall earnings in both 2-Ord than in both 1-Ord treatments. (Details on "pro-social" and perverse first-order punishment are discussed in Result 6.)

64% of subjects in 2-Ord-OG and 48% of subjects in 2-Ord-FS used the opportunity to second-order punish at least one time in their sessions. 72% and 59% of subjects, respectively, were targeted for second-order punishment at least once in their sessions. The typical subject punished at least one other subject in 27% of the second-order stages in her session of 2-Ord-OG and in 21% of those stages in 2-Ord-FS.

Result 3a. 1st order punishment in the 2-Ord treatments was (as usual) mainly directed at low contributors.

The direction of first-order punishment in the new treatments is qualitatively the same as in FG, 1-Ord-10, and 1-Ord-20. Table 6 shows estimates of regressions paralleling those in FG and in Table 1, above, for 2-Ord-OG and 2-Ord-FS. In these, the amount of (first-order) punishment received by subject *i* is predicted by the absolute positive deviation of *i*'s contribution from the average contribution of the remaining three group members in the period, the absolute negative deviation, and the average

contribution. In both 2-Ord treatments, the absolute negative deviation has a highly significant positive coefficient. Indeed, the estimated coefficients on absolute negative deviation in Tables 1 and 5 lie in a remarkably narrow band, indicating that for each dollar contributed below the group's average, a subject's earnings were reduced by other group members by about 60 cents (the latter being precisely enough to negate the private financial gain from not contributing it). The significant tendency for low contributors to be punished helps to explain why contributions to the group account in the new treatments remained well above those typical of VCM experiments without punishment.

Table 6: OLS Regressions, First Order Punishment Received as a Function of Deviation of Recipient's Contribution from Others' and Average Contribution Level

		2ord-OG		2ord-FS	
Constant		-0.057		-0.02	
(standard error)		0.203		0.12	
Positive Deviation		0.098	**	0.023	
(standard error)		0.037		0.025	
Negative (absolute) Deviation		0.615	***	0.581	***
(standard error)		0.024		0.018	
Average Contribution		0.027		0.009	
(standard error)		0.022		0.014	
Number of observations		1152		1152	
Adjusted R-squared		.373		.489	

* indicates significance at the .05 level
** indicates significance at the .01 level
*** indicates significance at the .001 level

Result 3b. However, perverse first-order punishment did occur.

Despite the findings just reported, first-order punishment of groups' highest contributors did occur in the new treatments. A Mann-Whitney test finds no difference that is significant at conventional levels between the *proportion* of 1st order punishment that is perverse in either 2-Ord treatment versus either 1-Ord treatment. The same type of test finds no difference between the *absolute amount* of perverse first-order punishment in the 2-Ord-OG treatment and either 1-Ord treatment, but finds the absolute amount of

perverse first-order punishment to be lower in 2-Ord-FS than in each 1-Ord treatment, significant at the 10% although not at the 5% level in one tailed tests.

Table 6 also shows that in 2-Ord-OG higher contributors were actually significantly more likely to be punished than were average contributors, receiving an average of about 10 cents of punishment per dollar contributed above the mean. This coefficient provides more evidence of perverse punishment in the 2-Ord-OG treatment.[21]

Result 4. 2nd-order punishers targeted firstt-order punishers of all types, with perverse first-order punishers receiving the most second-order punishment per dollar of first-order punishment they gave.

We begin to study who was targeted for second-order punishment by looking at Table 7, which presents simple regressions similar to those of Table 6, but for the second-order punishment stage. Here, the amount of punishment aimed at subject *i* in this stage is predicted by the three pieces of information on subjects' screens at the time of this decision, namely the amount by which *i* reduced below- , above-, and average contributors in the previous three periods.[22] The results clearly show that those who punished, in any fashion, were more likely to receive second-order punishment than those who did not punish at all. Thus, unlike the models of Boyd *et al*., the higher-order punishment stage in our experiment was not marked by the sanctioning of those who free rode on punishing.

Table 7: Determinants of second-order punishment received: preliminary analysis

Variable	2ord-OG		2ord-FS	
Punishment given to low contributors	0.294	***	0.384	***
(standard error)	0.030		0.026	
Punishment given to average contributors	0.351	***	0.977	
(standard error)	0.065		0.861	

[21] The fact that the other coefficients on absolute positive coefficient here and in FG's regression for their own data are not statistically significant, and that the insignificant coefficients in the 1-Ord-10 and 2-Ord-FS regressions are positive, are also reflections of the perverse punishment that we've seen to occur in those experiments. Were it to have been the case that highest contributors were never punished but contributors of more than the group's average and less than its maximum sometimes were punished, the coefficient would be significantly *negative*. not positive.

[22] Recall that subjects had no information about the contribution behaviors of each individual, thanks to the re-scrambling of subject identification letters.

Punishment given to high contributors	0.856	***	0.488	***
(standard error)	0.040		0.072	
(constant)	0.320	**	-0.020	
(standard error)	0.115		0.085	
Number of observations	384		384	
Adjusted R-squared	.623		.493	

* indicates significance at the .05 level
** indicates significance at the .01 level
*** indicates significance at the .001 level

Result 5. First- order punishment of both low and high contributors was significantly discouraged when those doing the punishing received second-order punishment for it.

Table 8 shows results of regressions for the 2-Ord-OG (columns 1 and 3) and 2-Ord-FS (columns 2 and 4) treatments, respectively. In columns 1 and 2, the change in subject *i*'s punishment of low contributors from the previous three periods to the three periods following a second-order punishment stage is the dependent variable, and the amount of second-order punishment which *i* received for punishing low contributors in that stage is the explanatory variable. Columns 3 and 4 are parallel regressions, except that the dependent variable is the change in punishment of high contributors, and the explanatory variable is the amount of second-order punishment received for punishing high contributors.[23] The regression results strongly support the conjecture that second-order punishment discouraged both pro-social and perverse punishers from persisting in those first-order punishment behaviors. For the "own group" treatment, the coefficients imply that for every dollar of second-order punishment given due to perverse punishment, the punishment given per dollar of positive deviation decreased by 3 cents, and for every dollar of second-order punishment given for normal punishment, the punishment given per dollar of negative deviation decreases by 1.7 cents. For the "full session" treatment, the corresponding numbers are 20.4 and 2.5 cents, respectively. To

[23] Because the information shown to subjects at the second-order punishment stage organized first-order punishment activity into that directed at contributors of less than, more than, and equal to the group average, we define punishing low contributors here as punishing those who contributed less than their group average for the period, and punishing high contributors as punishing those who contributed more than their group average for the period. Some judgment is required to identify how much second-order punishment was given "for punishing low contributors" or "for punishing high contributors." The approach we follow is to consider all second-order punishment received as being "for punishing low (high) contributors" if the person targeted punished only low (high) contributors, and otherwise to simply apportion the punishment received in proportion to the shares of the target's punishment that were directed towards low and high contributors, respectively.

set these numbers in context, in the "own group" treatment, the average level of punishment per dollar of positive deviation in the three periods preceding a second-order punishment event was 4.2 cents, and the average punishment per dollar of negative deviation was 10.4 cents. The corresponding numbers for the full session treatment are 6.8 and 10.3 cents, respectively.

Table 8: OLS Regression results: Impact of 2nd order punishment on 1st order punishment intensity

	change in punishment for positive deviation		change in punishment for negative deviation	
	2ord-OG	2ord-FS	2ord-OG	2ord-FS
2nd Order Punishment received for Perverse Punishment	-0.030 ***	-0.204 ***		
(standard error)	0.003	0.014		
2nd Order Punishment received for Normal Punishment			-0.017 ***	-0.025 ***
(standard error)			0.004	0.006
(constant)	0.008	-0.001	0.010	0.008
(standard error)	0.007	0.009	0.011	0.011
Number of Observations	320	320	320	320
Adjusted R squared	.243	.382	.043	.057

* indicates significance at the .05 level
** indicates significance at the .01 level
*** indicates significance at the .001 level

Result 6. There was significantly less first-order punishment of low contributors in 2-Ord treatments than in 1-Ord treatments. There was also less perverse first-order punishment in 2-Ord treatments, and the amount and share of perverse first-order punishment declined with repetition.

Table 9 shows the average dollars of total and of perverse first-order punishment per period in the first and second halves of sessions of each of the 1-Ord and 2-Ord treatments, and the averages in the last period of each treatment. Total first-order

Table 9. Average total and perverse punishment per period, by half or period

	1-Ord-10	1-Ord-20	1-Ord Avg	2-Ord-OG	2-Ord-FS	2-Ord Avg
Total punishment, 1st half	1.366666667	1.067188	1.216927	0.848958	0.743056	0.796007
Perverse punishment, 1st half	0.245833333	0.232813	0.239323	0.177083	0.055556	0.116319
Total punishment, 2nd half	1.258333333	1.1625	1.210417	0.690972	0.440972	0.565972
Perverse punishment, 2nd half	0.229166667	0.175	0.202083	0.097222	0.013889	0.055556
Total punishment, last period	1.416666667	1.75	1.583333	0.765625	0.890625	0.828125
Perverse punishment, last period	0.166666667	0.15625	0.161458	0	0	0

punishment shows no decline with repetition in the 1-Ord treatments, and its high level in the last period supports the idea that it was not performed for strategic purposes. For 2-Ord treatments, overall first order punishment shows signs of some decline between the first and second halves of sessions, although it rebounds in the last period, again indicating the moral or visceral as opposed to strategic nature of punishment. Perverse first-order punishment, however, shows a definite declining trend, actually falling to zero in the last period of every session of both 2-Ord-OG and 2-Ord-FS.

Mann-Whitney tests show that the amount of punishment aimed at low contributors was less in both the 2-Ord-OG and 2-Ord-FS treatments than in the 1-Ord-10 treatment, significant at the 5% level. Both the amount and the share of first-order punishment that are perverse appear to be lowest in the 2-Ord-FS treatment. Mann-

Whitney tests at group level confirm that there was less perverse first-order punishment per period in 2-Ord-FS than in either 1-Ord-10 or 1-Ord-20, although at a borderline level of significance.[24] The other treatments, including 2-Ord-OG, are not significantly different from one another, nor do any of the tests for the proportion of first-order punishment which was perverse show significant differences. OLS regressions in which a time trend is the only explanatory variable confirm that the amount and share of perverse first-order punishment out of overall first-order punishment declined with repetition in the second-order punishment treatments, significant at the 10% and 5% levels for 2-Ord-OG and 2-Ord-FS, respectively.[25]

Result 7. Highly significant and intuitively sensible relationships hold between punisher and punishee attributes and amount of second-order punishment given. For example high contributors, especially those perversely punished, were more likely than others to punish perverse punishers, while low contributors, especially those punished when contributing little, were more likely than others to punish punishers of low contributors. The proportion of second-order punishment that can be explained by these and similar factors is a remarkable 39% in 2-Ord-OG, but only 8% in 2-Ord-FS.

Analysis of who punished whom in second-order punishment stages shows subjects behaving "true to type." The regressions displayed in Table 10 allow us to test many intuitive conjectures, and the number of statistically significant coefficients that result is truly impressive—and reassuring that the experiment's complexity did not lead to chaos. For example, consider the conjecture that subjects who were punished for free riding in the first-order punishment stage used second-order punishment to try to retaliate against those who punished them. That conjecture is strongly supported by the highly significant positive coefficient on the interaction of punisher *i*'s contribution deviation and target *j*'s level of punishment to below-average contributors (variable 6), and especially by the highly significant positive coefficient on the interaction between the

[24] The *p*-value for one-tailed tests of the hypothesis that there is less perverse punishment in the "full session" treatment is .087 for the comparison with 1-Ord-10 and .080 for the comparison with 1-Ord-20.
[25] There is one observation per group and period. If period 18 is excluded, the coefficients on the time trend in both treatments are significant with *p*-values of about .06. Similar regressions for the percentage of all punishment that was perverse also show negative coefficients on the time trend , but they are not significant at conventional levels.

amount of punishment received by *i* when a low contributor and the amount of punishment that *j* gave to low contributors (variable 11). The coefficients on *j*'s first-order punishment (variables 3 – 5) continue to be significant and positive in all cases. The estimates suggest that a dollar of perverse first-order punishment, which cost *j* 25 cents to impose, attracted 17 to 19 cents of second-order punishment in the OG treatment and 2 to 6 cents of second-order punishment in the FS treatment, while a dollar of first-order punishment aimed at a low contributor attracted about 6 to 7 cents of second-order punishment in OG and 2 cents in FS. Hence, a dollar of first-order punishment resulted in up to three times as much second-order punishment when aimed at high contributors than at low ones, showing a pro-efficiency response to be predominant, though not the only type present. The same amount of first-order punishment elicited substantially less second-order punishment in the FS treatment, most likely because subjects had no way of knowing whether they were aiming it at a member of their own group or a member of another group in their session. All else being equal, the more *j* contributed relative to others' average, the more *j* engaged in second-order punishment, while larger negative deviations of contribution did not affect second-order punishment except through the interaction with the punishment behavior of the person targeted.

Table 10: Determinants of 2nd order Punishment Given (by *i*) and Received (by *j*)								
Variable	2ord-OG				2ord- FS			
1. Positive Deviation of *i*'s Contribution	0.005 0.011	0.003 0.011	0.003 0.011	0.00 0.01	0.003 0.002	0.007 *** 0.002	0.007 *** 0.002	0.007 *** 0.002
2. Negative Deviation of *i*'s Contribution	0.011 0.009	0.012 0.010	0.011 0.010	0.02 0.01	0.005 *** 0.002	0.002 0.002	0.002 0.002	0.004 0.002
3. 1st Order Punishment of Low contributors by *j*	0.068 *** 0.014	0.067 *** 0.016	0.065 *** 0.016	0.06 *** 0.02	0.020 *** 0.002	0.025 *** 0.002	0.018 *** 0.002	0.018 *** 0.002
4. 1st Order Punishment of Average contributors by *j*	0.118 *** 0.021	0.117 *** 0.021	0.117 *** 0.021	0.09 *** 0.02	0.062 0.067	0.064 0.067	0.039 0.066	0.039 0.066
5. 1st Order Punishment of High contributors by *j*	0.169 *** 0.030	0.174 *** 0.031	0.190 *** 0.037	0.17 *** 0.04	0.024 *** 0.006	0.024 *** 0.006	0.057 *** 0.007	0.057 *** 0.007
6. *i*'s Deviation X *j*'s Punishment of Low		0.000 0.002	0.000 0.002	0.00 0.00		0.004 *** 0.000	0.003 *** 0.000	0.002 *** 0.000
7. *i*'s Deviation X *j*'s Punishment of Average		-0.106 0.083	-0.104 0.083	-0.10 0.08		-0.025 0.013	-0.030 * 0.013	-0.031 * 0.013
8. *i*'s Deviation X *j*'s Punishment of High		-0.004 0.003	-0.003 0.004	0.00 0.00		-0.001 0.001	0.007 *** 0.001	0.007 *** 0.001
9. Normal Punishment Received by *I*				-0.02 0.02				-0.004 0.003
10. Perverse Punishment Received by *I*				0.08 *** 0.02				-0.018 0.014
11. Normal Punishment Received by *I* X Punishment of Low by *j*	0.004 *** 0.001	0.005 ** 0.002	0.005 ** 0.002	0.01 ** 0.00	0.002 *** 0.000	-0.001 0.001	0.003 *** 0.001	0.003 *** 0.001
12. Normal Punishment Received by *i* X Punishment of High by *j*			-0.003 0.004	0.00 0.00			-0.017 *** 0.002	-0.017 *** 0.002
13, Perverse Punishment Recieved by *i* X Punishment of High by *j*	0.011 *** 0.003	0.008 ** 0.004	0.008 * 0.004	0.01 0.00	0.041 *** 0.010	0.046 *** 0.010	0.055 *** 0.010	0.059 *** 0.011
(constant)	0.104 ** 0.045	0.105 ** 0.045	0.104 * 0.045	0.11 ** 0.05	-0.016 0.009	-0.020 * 0.009	-0.019 * 0.009	-0.015 0.009
Number of Observations	1152	1152	1152	1152	5784	5784	5784	5784
Adjusted R squared	0.38	0.381	0.381	0.39	0.0559	0.069	0.081	0.082

* indicates significance at the .05 level

** indicates significance at the .01 level
*** indicates significance at the .001 level

Notes: Positive and negative deviation of contribution are deviations from the average contributed by other's in *i*'s group, set to zero if *i* contributed less than (more than) the average, as in the regressions of Table 6. *i*'s deviation, which enters variables 6 – 8 multiplicitavely, is the average contributed by others in *i*'s group minus *i*'s contribution. In variables 9 – 13, punishment received by *i* is counted as normal if *i* received it when contributing less than the group average and as perverse if *i* received it when contributing more than the group average. This use of "perverse" is broader than elsewhere in the paper, where it means aimed at the group's highest contributor. The broader definition is adopted in this case because information seen by subjects in the second-order punishment stage was classified by this above and below average criterion.

Result 8. Second-order punishment of perverse first-order punishers does not appear to be attributable to strategic motives, since there is no less of it in the last period. By contrast, second-order punishment of normal first- order punishers is less in the last period, suggesting a substantial strategic element.

Like first-order punishment, second-order punishment in a repeated interaction with fixed groups might be explained as being intended to influence others' future choices for the benefit of one's own payoff—a low contributor, for example, might rationally prefer to increase his chances of getting away with future low contributions for a certain cost now rather than raise his contributions to avoid future first-order punishment. Since no such end can be accomplished in the last second-order punishment stage, after period 18, the hypothesis that second-order punishment is strategically motivated can be tested by comparing the amount in that period with amounts in earlier periods. Mann-Whitney tests of average second-order punishment of those who first-order punished high contributors—what can be called "normal" or "pro-social" first order punishment—show that it is not statistically different in amount in the last period versus the other five such stages. Corresponding Mann-Whitney tests for second-order punishment of those who first-order punished low contributors—"perverse" second-order punishment—show that it is lower in the last period than in the other five 2nd order punishment stages.

In summary, experiments adding second-order punishment stages showed that subject heterogeneity continued to play itself out in those stages, with those inclined towards high contributions and pro-social punishment now punishing perverse first-order punishers, while those inclined toward low contributions and perverse first-order punishment now punish pro-social first-order punishers. More neutral subjects largely refrain from punishing and thus avoid bearing its costs. Both pro-social and perverse first-order punishment decline relative to treatments without second-order punishment, but perverse punishment declines more decisively, so contributions remain at least as high as in experiments with only first-order punishment, and earnings are somewhat higher. The experiments thus provide further evidence of subject heterogeneity, and, while showing that perverse first-order punishment can be mitigated by introducing a

higher order punishment opportunity, they also underscore the strength of the motivation underlying it by showing it to migrate from the first- to the second-order stage so long as it continues to be permitted expression.

7. Discussion and Conclusion

Fehr and Gächter are almost certainly correct in arguing that collective interactions are unlikely to be understood without taking into account the presence of many individuals who are willing to reciprocate one another's contributions to the public good and to incur costs to punish those who free ride. But the characterization of typical populations as composed of strict payoff maximizers and reciprocators only is oversimplified. Most groups of moderate size will also include some individuals with a taste for punishing the reciprocator types, making for a more complex "moral ecology." When reciprocators cannot sanction free riders without fear of being punished in return, willingness to punish and to contribute are both negatively affected, and the efficiency of punishment is reduced by the admixture of counter-productive punishment with the efficiency-enhancing punishment of free riders. The frequency of perverse punishment is adequate to account for the failure of adding a punishment stage to a VCM to raise earnings—to raise them at all, in one of two treatments studied in detail, to raise them significantly, in the other—even though it raises contributions to the public good.

How can the pernicious effects of perverse, often retaliatory, punishment be countered? We conjectured that perverse punishment might be common in the experiments discussed because the designs allow the punisher to act in secret without possibility of sanctioning by the majority of group members who either actively promote or passively accept a cooperative pattern of behavior. To see whether group members would sanction perverse punishment if given the opportunity, we modified a VCM with punishment stage to include stages of second-order punishment in which group members could elect to impose costly sanctions on others identified only by their first-order punishment behaviors displayed as aimed at above-average, average, or below-average contributors to the group account. We found that perverse (first-order) punishers were in fact routinely sanctioned by other group members, but pro-social punishers were also sanctioned, though less intensely. We confirmed that punishment of pro-social punishers

was disproportionately done by low contributors who had themselves received punishment for free riding. While perverse first-order punishment declined significantly with the introduction of the second-order punishment stage, in effect it migrated to that stage, contributing to a decline in the incidence of pro-social punishment. Contributions remained high enough that earnings rose slightly, although significantly so in one treatment. Perverse second-order punishment persisted into the final period but at a lower level, indicating that it was partly but not entirely explained by the self-interest of payoff maximizing punishers, rather than a taste for "getting even."

Another way to deal with the strongly anti-social individuals who appear to constitute a small but bothersome minority in most groups is to exclude them from interacting with more cooperative peers and with those who quietly adopt cooperation if threatened with punishment. Cinyabuguma, Page and Putterman (2003) permitted subjects in a VCM experiment to expel to a secondary, low-reward group those whom a majority cast costly votes against, and this led to significantly higher contributions and earnings on average, even when the few expelled individuals were included in the calculations. Putterman and Ones (in process) report on a VCM experiment in which subjects were (without their knowledge) classified as high contributor/pro-social punisher or low contributor/perverse punisher types based on behaviors in early, "diagnostic" play, then assigned to more homogeneous groups for further interaction. All but the groups with the lowest contributors and most perverse punishers achieved higher levels of cooperation and earnings when the latter were excluded.

If the supposition that perverse punishers are usually in a minority is correct, then a mechanism of majority rule may be able to tame perverse punishment. Ertan, Page and Putterman (2003) report on a VCM experiment in which subjects periodically voted on whether to permit earnings reductions aimed at low, high, and/or average contributors to the public good. In 160 votes, not a single group of four subjects ever had a majority vote to allow punishment of high contributors. Groups which chose to allow punishment of low contributors tended to achieve both high contributions and higher earnings, and once a few groups in a session made this choice, others caught on and imitated, so that sessions endogenously evolved towards complete adoption of a regime of punishing free riders only without any suggestion by the experimenters. This experiment suggests that a

judicious mix of collective (voting) and decentralized decision-making (it was still up to individuals whether or not to engage in costly punishment) holds one key to solving problems of collective action in mixed groups of payoff-maximizers, reciprocators, and retaliators. It may also point the way towards a certain theory of government: if we combine Locke's suggestion that enforcement be in the hands of the state with the insight that democratic control of the state is likely to cause the perverse minority to be overruled, we may conclude that democratic choice of mandated contributions and penalties is yet another way to solve free rider problems.

To draw useful institutional lessons from the heterogeneity of individuals and the "moral ecology" of their interactions, it seems important to know something not only about a number of different agent types, but also about the relative numbers in which each tends to be found. The "demography of types" is a theme we hope to pursue in future research.

References

Anderson, Christopher M. 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).

Bochet, Olivier, Talbot Page and Louis Putterman, forthcoming, "Communication and Punishment in Voluntary Contribution Experiments," *Journal of Economic Behavior and Organization* (in press; full version, Working Paper No. 2002-29, Department of Economics, Brown University, available at http://www.econ.brown.edu/2002/).

Boyd, Robert, Herbert Gintis, Samuel Bowles and Peter Richerson, 2002, "Interdemic Group Selection can Lead to the Evolution of Group Beneficial Punishment in Large Groups," in Gintis, Bowles, Boyd and Ernst Fehr, eds., *The Moral Sentiments: Evidence, Models, and Policy*. Unpublished manuscript submitted for publication.

Boyd, Robert and Peter Richerson, 2002, "Group Beneficial Norms can Spread Rapidly in a Cultural Population," *Journal of Theoretical Biology* 215: 287-96.

Carpenter, Jeffrey and Peter Matthews, 2002, "Social reciprocity," Middlebury College Department of Economics Working Paper #29.

Cinyabuguma, Matthias, Talbot Page and Louis Putterman, 2003, "Cooperation Under the Threat of Expulsion in a Public Goods Experiment," unpublished paper, Brown University.

Davis, Douglas D. and Charles A. Holt, 1993, *Experimental Economics.* Princeton: Princeton University Press.

Ertan, Arhan, Talbot Page and Louis Putterman, 2003, "Public Choice and Private Sanctions in a Public Goods Game: An Experiment with Endogenous Institutions," unpublished paper, Department of Economics, Brown University.

Falk, Armin, Ernst Fehr, and Urs Fischbacher, 2001, "Driving Forces of Informal Sanctions," Working Paper No. 59, Institute for Empirical Research in Economics, University of Zurich, September.

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.

Henrich, Joseph, forthcoming, "Cultural Group Selection, Co-evolutionary Processes and Large-Scale Cooperation," *Journal of Economic Behavior and Organization* (in press).

Henrich, Joseph and Robert Boyd, 2001, "Why People Punish Defectors: Weak Conformist Transmission can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas," *Journal of Theoretical Biology* 208: 78-89.

Hoffman, Elizabeth, Kevin McCabe and Vernon Smith, 1998, Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology. *Economic Inquiry* 36: 335-52.

Ledyard, John, 1995, "Public Goods: A Survey of Experimental Research," pp. 111-94 in John Kagel and Alvin Roth, eds., *Handbook of Experimental Economics*. Princeton: Princeton University Press.

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

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, 2002, "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency," Working Paper No. 2002-19, Department of Economics, Brown University, available at http://www.econ.brown.edu/2002/.

Putterman, Louis and Umut Ones, "'Moral Ecology' and Sorting in a Public Goods Experiment," work in progress, Department of Economics, Brown University.

Saijo, Tatsuyoshi and Hideki Nakamura, 1995, "The 'Spite' Dilemma in Voluntary Contribution Mechanism Experiments," *Journal of Conflict Resolution* 38 (3): 535-60.

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.

Somerville, John and Ronald E. Santoni, eds., 1963, *Social and Political Philosophy: Readings from Plato to Gandhi*. Garden City, NY: Anchor Books.

Subjects' Instructions for the Second Order Punishment Experiment, OG Treatment

[Screen 1]
This is an experiment, funded by a research foundation, to study decision-making. You will be earning money in "experimental dollars" during the experiment. At the end of the experiment you will be paid in cash in real dollars (one experimental dollars converts to 8 real cents). The amount you will earn will depend on your and others' decisions. The maximum possible earning is $31.68 (real dollars) and the minimum possible is $5. You are likely to earn an amount in between. Please make sure you understand the decision process.

[Screen 2]

Structure of the Experiment

The experiment consists of eighteen distinct periods or rounds of decision-making. Each of these eighteen periods shares a common structure, consisting of two stages. After every three such periods, there will be an additional, third stage as explained later in the instructions.

[Screen 3]

Your Group

At the beginning of the experiment, you will be randomly matched with three other participants, to form a group of four that will remain together throughout the experiment. The other three people who are in your group will be identified to you as "B," "C," and "D," although the letters will be shuffled from period to period, so that the person identified as "B" in one period is equally likely to be called "C" or "D" in the next one. You will not know the actual identities of the other members of your group either while making your decisions or after the experiment.

[Screen 4]

Your First Decision: Assigning Money to Group and Personal Accounts

At the beginning of every period each person in your group will receive $10 (experimental dollars). Each of you must decide how to divide this amount between a group account and a personal account.

The money you assign to your personal account goes into your earnings.

An amount equal to 0.4 times the group's total assignment to the group account goes into your earnings.

Your earnings = (amount in your personal account) + (0.4)(total in group account)

[Screen 5]

The next four screens illustrate how the experiment works. Fill in the blanks of your worksheet first, then enter the information in the practice decision screen. The numbers you type in the practice screens are for practice only and will net affect your earnings from the experiment.

Practice Questions

Practice 1.

The four members of your group each have $10. Every member of your group has assigned $10 to the group account and $0 to their personal account. Fill in the blanks on the right.

(1) Amount you assigned to group account . . . . . $_____
(2) Amount you assigned to your personal account . . . . . $_____
[= $10 – group account assignment on line (1)]
(3) Total number of dollars assigned to your group account . . . . . $_____
(4) Income from the group account for a member of your group . . . $_____
[0.4 • group account total in line (3)]
(5) Your earnings after the assignment decisions . . . . . $_____
[group account income in line (4) + personal account income in line (2)]

Now, go back to the practice screen. Type in your assignment to the group account, press enter, and check your calculation.

Practice 2.

The four members of your group each have $10. Every member of your group has assigned $0 to the group account and $10 to their personal account. Fill in the blanks on the right.

(1) Amount you assigned to group account . . . . . $_____
(2) Amount you assigned to your personal account . . . . . $_____
[= $10 – group account assignment on line (1)]
(3) Total number of dollars assigned to your group account . . . . . $_____
(4) Income from the group account for a member of your group . . . $_____
[0.4 • group account total in line (3)]
(5) Your earnings after the assignment decisions . . . . . $_____
[group account income in line (4) + personal account income in line (2)]

Type in your assignment to the group account, press enter, and check your calculation.

Practice 3.

Person B assigned $10 to the group account and $0 to his or her personal account,
person C assigned $5 to the group account and $5 to his or her personal account,
person D assigned $0 to the group account and $10 to his or her personal account, and
you assigned $5 to the group account and $5 to your personal account.

Fill in the blanks on the right.

(1) Amount you assigned to group account . . . . . $_____
(2) Amount you assigned to your personal account . . . . . $_____
[= $10 – group account assignment on line (1)]
(3) Total number of dollars assigned to your group account . . . . . $_____
(4) Income from the group account for a member of your group . . . $_____
[0.4 • group account total in line (3)]
(5) Your earnings after the assignment decisions . . . . . $_____
[group account income in line (4) + personal account income
in line (2)]

Type in your contribution, press enter, and check your calculation.

[Screen 6]

Consider what would happen in practice 3 if you increase your assignment to the group account by $1.

Your personal account would go down by $1, reducing your earnings by $1.

Your group account would go up by $1, increasing your earnings by $0.40, for a net reduction of $0.60

But each of the other people in your group would increase their earnings by $0.40, for a total increase of $1.20 for the others in your group.

[Screen 7]

The Second Stage

After you learn the assignments to the group account by the others in your group, you have a chance to their earnings, and they have a chance to reduce your earnings. Suppose, in the last example, you decide to:

reduce B's earnings by $2
reduce C's earnings by $3
reduce D's earnings by $4

The total amount of reductions you make on others' earnings is $9.

It costs you $0.25 for each $1 you reduce others' earnings. So your own earnings are reduced by (0.25)($9) = $2.25 in this example.

Now, suppose

B reduces your earnings by $2
C reduces your earnings by $1
D reduces your earnings by $0

The total reduction of your earnings by others is ($2 + $1 + $0) = $3. Your screen will tell you how much your earnings have been reduced, but not who has reduced your earnings by what amount.

Similarly none of the others will learn by how much you have reduced their earnings. They will only learn their total reductions by others in the group as a whole.

Please fill in the sheet labeled practice 4 and the corresponding practice decision screen.

Practice 4.

You assigned $5 to the group account and $5 to your personal account,
person B assigned $10 to the group account and $0 to his or her personal account,
person C assigned $5 to the group account and $5 to his or her personal account, and
person D assigned $0 to the group account and $10 to his or her personal account.

You reduce person B's earnings by $2,
person C's earnings by $3, and
person D's earnings by $4.

You receive a total of $3 in reductions from other members of your group.

(1) Amount you assigned to group account . . . . . $_____
(2) Amount you assigned to your personal account . . . . . $_____
[$10 – group account assignment on line (1)]
(3) Total number of dollars assigned to your group account . . . . . $_____

(4) Income from the group account for a member of you group . . . $_____
[(0.4) • group account total in line (3)]
(5) Your earnings after the assignment decisions . . . . . .$_____
[group account income in line (4) + personal account income in line (2)]
(6) You reduced the earnings of others in your group by a total of . .$_____
(7) This cost you . . . . . $_____
[(0.25) • the sum of your reductions from line (6)]
(8) Other members of your group reduced your earnings by . . . . . $_____
(9) Your total earnings for this period . . . . . $_____
[Your earnings after the assignment decisions on line (5) minus your reduction cost on line (7) minus the amount by which your earnings were reduced on line (8)]

Enter your reductions, press enter, and check your calculation.

[Screen 8]

Your Net Earnings

Your net earnings for a period will be:

Amount in personal account
+ (0.4)(total in group account)
- (0.25)(total of your reductions of others)
- total of reductions of your earnings made by others.

If this results in a negative number in any period, your earnings for that period will be set to zero.

[Screen 9]

Each period you begin with a new $10 and each period's earnings are independent of the others. The right side of your decision screen will display a record your earnings, period by period.

[Screen 10]

The Third Stage

There is one last set of decisions to be made in this experiment. Every time that three periods of the type just described have been completed, there will be an additional decision stage. At this stage, your screen will provide information on the choices made by you and others during the three periods just past. In particular, for each group member, identified as "You," or the randomly chosen "B," "C," or "D," for the others, you will be shown

(1) the total number of dollars by which that person reduced the earnings of group members who assigned to the group account less than the group's average for the period
(2) the total number of dollars by which that person reduced the earnings of group members who assigned to the group account exactly the group's average for the period, and
(3) the total number of dollars by which that person reduced the earnings of group members who assigned to the group account more than the group's average for the period

Note that the determination of what is average and below average is made separately for each group and period, and the total reductions in each category in each of the three periods are then added together.

[Screen 11]

Possible Reductions

Under this information, there will be a row of boxes. In each box you can enter the number of dollars, if any, by which you would like to (further) reduce the earnings of others. It costs you $0.25 to reduce another's earnings by $1. You can spend, on such reductions, up to the net amount you have earned in the previous three periods. The maximum amount by which a group member's earnings can be reduced by the combined action of the other group members is the amount that causes their net earnings for the three periods to be zero. If the sum of the reductions initially chosen exceeds that amount, each member's reductions and reduction costs will be proportionately reduced until this maximum is reached. If you decide not to reduce another's earnings, enter zero in the appropriate box. For an illustration of this third decision stage, click next.

[Screen 12]

In this illustration, person "B," for example, has reduced the earnings of group members assigning below-average amounts by a total of $2 over the previous three periods, those of members assigning average amounts by a total of $0, and those of members assigning above-average amounts by a total of $0. The amounts by which you choose to reduce person "B"'s earnings would be entered in the box labeled b, and those for persons "C" and "D" in boxes c and d. Your total reductions and reduction costs would appear in boxes e and f. Click on next for additional information.

[Screen 13]

After all subjects have entered their decisions, you will be shown (in the second row of boxes), the total dollars of reductions received by each group member, including yourself.

[Screen 14]

Conclusion

During the experiment, there is to be no communication, apart from the entering of your decisions. Therefore, it is important that you understand the decision-making process fully before the experiment begins. Please raise your hand now if you have any questions. The experiment will begin once all questions have been answered.