

# Can second-order punishment deter perverse punishment?

Matthias Cinyabuguma · Talbot Page ·  
Louis Putterman

Received: 2 February 2005 / Revised: 5 December 2005 / Accepted: 10 February 2006  
© Economic Science Association 2006

**Abstract** Recent experiments have shown that voluntary punishment of free riders can increase contributions, mitigating the free-rider problem. But frequently punishers punish high contributors, creating “perverse” incentives which can undermine the benefits of voluntary punishment.

In our experiment, allowing punishment of punishing behaviors reduces punishment of high contributors, but gives rise to efficiency-reducing second-order “perverse” punishment. On balance, efficiency and contributions are slightly but not significantly enhanced.

**Keywords** Public goods · Collective action · Experiment · Punishment

**JEL Classification** C91 · C73 · C41 · D71

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

Punishment in experimental games is of considerable interest to economists exploring behavioral approaches, since it is a prime example of a behavior often inconsistent with the self-interested rational actor model and suggestive of a non-standard preference and/or of the influence of emotions. In studying the free-rider problem in a voluntary contribution

---

M. Cinyabuguma  
Department of Economics, University of Maryland, Baltimore County  
e-mail: matthias@umbc.edu

T. Page  
Economics and Environmental Studies, Brown University  
e-mail: talbot\_page@brown.edu

L. Putterman (✉)  
Department of Economics, Brown University, Providence, RI 02912  
e-mail: louis\_putterman@brown.edu

<sup>1</sup> P. 185 in Somerville and Santoni (1963)

mechanism, Fehr and Gächter (2000) allowed subjects to learn of others' contributions to a public good and then to have the opportunity to punish (reduce the earnings) of others. This opportunity led to considerable punishment, mostly targeted at low contributors. Such punishment was an effective incentive, leading low contributors to increase their contributions and mitigating the free rider problem.

But a substantial amount of punishment was directed at high contributors. The phenomenon of punishing high contributors has been noticed by virtually everyone who has studied punishment in voluntary contribution mechanisms (VCMs). In our own estimates, typically 20% or more of all punishment events are directed at the highest contributor in the group (or highest contributors if there are ties).

Punishment is commonplace in teams, firms, and partnerships, in terms of reduced salary raises and bonuses, exclusion from information loops and desired assignments, and more informally in terms of malicious gossip. If punishment of high contributors occurs at frequencies anywhere near what it is in VCM experiments, it likely would have a negative effect on cooperation and efficiency of organizations.<sup>2</sup> Experimentalists have suggested several possible reasons to explain the punishment of high contributors but as far as we know, the phenomenon has not been systematically studied in economic experiments.

In the experiment reported here, we studied two stages of punishment in a VCM. In the first stage the subjects in a group, having made voluntary contributions to a public good, learn the amounts of others' contributions and are allowed to punish each other (this is the standard VCM with a punishment option). In the second stage the subjects learn how much punishment each of the other subjects targeted at high contributors and at low contributors, and then the subjects are allowed to punish each other a second time. Thus in the first stage, individuals have the opportunity to punish each other on the basis of others' contribution behavior (first-order punishment); in the second stage individuals have the opportunity to punish each other on the basis of others' punishment behavior (second-order punishment). We compared results of our design allowing first-and-second-order punishment with results from the standard VCM with punishment, which is limited to a single stage of first-order punishment.

We call the *first-order punishment of low contributors* "normal" because it seems likely to strengthen incentives to increase contributions and aggregate earnings, and *first-order punishment of high contributors* "perverse" because it seems to work in the opposite direction. We call the *second-order punishment of punishers of high contributors* "normal" because it seems to create incentives toward less first-order punishment of high contributors, and this in turn improves incentives for higher contributions. And we call *second-order punishment of punishers of low contributors* "perverse" because it seems to weaken incentives toward punishing low contributors, and this in turn weakens incentives for low contributors to increase their contributions. (We refine these concepts later.)

We asked: Would second-order punishment be targeted on perverse first-order punishment, thus reducing it, leading to higher contributions and earnings? Or would second-order punishment be frequently targeted "perversely" on normal first-order punishers, thus discouraging normal punishment and its incentives to increase contributions and earnings? Would punishment be as mis-targeted in its second stage as it was in the first? How persistent is perverse punishment?

Our main result is that perverse punishment is quite persistent. We found that in the design with both first- and second-order punishment, contributions and earnings did increase, compared with the design with only first-order punishment, but mostly insignificantly so.

<sup>2</sup> In the experiment we define (earnings) "efficiency" as  $\frac{\text{observed total earnings}}{\text{maximum possible total earnings}}$ .

In the first- and-second-order design the total amount of punishment decreased, compared with the first-order design, but only slightly and mainly insignificantly. It appeared to some extent that in the first-and-second-order design the problem of perverse punishment migrated from the first stage of punishment to the second stage. That is, in the first-and-second-order design, the percentage of perverse punishment in the first stage was less than the percentage of perverse punishment in the first and only stage of punishment in the first-order design. But in the first-and-second-order design, the benefits of this decrease were offset by perverse punishment at the second stage of punishing punishers. Perverse punishment seemed like a hardy weed, pulled out in one place in the garden only to reemerge in another place.

The costs of revenge and vendettas are well known. In many cases you know exactly who has injured you, and you want to get back. Our experiment suggests that in organizations with a goal of cooperative work, when there are opportunities for individuals to punish each other at their own discretion, without being identified, a substantial amount of the punishment may be dysfunctional from the point of view of the organization. For some earlier VCM experiments we calculated that the presence of perverse punishment suffices to explain why earnings were not higher (even though contributions were significantly higher) when punishment was allowed than when it was not allowed.<sup>3</sup>

Our experiment is most closely related to that of Nikiforakis (2004). Both his and our experiments are on first-and-second-order punishment and revenge plays a large role in both, but the two have strikingly different results. In Nikiforakis's experiment, the opportunity to "counter-punish" brought the high contribution levels in VCM experiments with only first-order punishment (e.g. Fehr and Gächter, 2000) down almost to the level of the baseline VCM without any punishment. In our experiment, the opportunity to punish punishers sustains and possibly increases the high levels of contribution found in VCMs with first-order-punishment only. Key differences in information and rules governing punishment are likely reasons for this difference in results.<sup>4</sup>

Our experiment is related to the larger literature on voluntary contributions mechanisms, which are finitely repeated  $n$ -person prisoners' dilemmas.<sup>5</sup> In our design, for self-interested subjects who believe others are also self-interested, the game unravels from the end. Subjects have an iterated dominant strategy to not undertake second-order punishment, to not undertake first-order punishment, and to not contribute. In contrast, in a subject pool of heterogeneous preference types, including self-interested payoff maximizers, conditional cooperators, and "perverse" preference types, individuals may rationally (in their own views) contribute, punish normally (punish low contributors), and sometimes punish perversely (e.g. punish high contributors in revenge).

Saijo and Nakamura (1995) offered an explanation of perverse punishment as motivated by a desire to increase one's own earnings relative to others, which they called "spite." In a common pool resource experiment, Ostrom, Gardner and Walker (1992) introduced a pun-

<sup>3</sup> See our Working Paper 2004-012, where we also report calculations showing that average earnings are lower with than without punishment under all three matching protocols in Fehr and Gächter (2000). See also Carpenter and Matthews (2002) and Sefton, Shupp and Walker (2002).

<sup>4</sup> In Nikiforakis's design, if you are first-order punished you learn who punished you and by what amount. This makes targeted revenge easy. In our design, if you are first-order punished you don't learn who punished you, only that you were punished in an identified aggregate amount. Blind revenge is possible, but less effective as revenge. In Nikiforakis's design, you are only allowed to second-order punish those who punished you; in our design, you learn the pattern of punishing high, average, and low contributors in your group (or session) and then you can second-order punish anyone (including punishing those who had a pattern of punishing high contributors but who did not punish you in this period).

<sup>5</sup> Ledyard (1995) and Davis and Holt (1993) summarize early work on VCMs.

ishment option, finding it led to increased cooperation. They noticed the phenomenon of perverse punishment, offering the explanation of “blind revenge” (“blind” because individuals who were punished wanted revenge but did not know who their punishers were). In a debriefing of one of our own experiments, a perverse punisher explained his action by saying that when someone does something morally superior to you (contributing high while you contributed low) you feel bad and you want to get back at the person (he called this the “flash bastard” phenomenon, we might call it moral resentment). Fehr and Gächter (2000), noticing perverse punishment in their experiment, summarized several possible explanations, including some of the ones above.

The paper is organized as follows. Section 1 presents the design of the experiment. Section 2 gives the results. Section 3 provides a summary and brief discussion of an application to models of the evolution of cooperation.

## 1. Experimental design

We will explain our design by beginning with the basic VCM and building up from that. In each period of our basic VCM design, each subject is assigned randomly and anonymously to a group of 4 subjects in a session of 16 participants and is given an endowment of 10 experimental dollars ( $E\$1 = \$0.08$ ) to divide between a private and a group account. Subject  $i$ 's earnings in a period are

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

where  $C_i$  is  $i$ 's contribution to the group account and the summation is taken over all 4 members of  $i$ 's group, including  $i$ .

To this basic design, an opportunity to punish is added at the end of each period. After all four subjects made their contribution decisions, each was informed of the *contribution* levels of each of the others (identified by letters B, C, or D, which were randomly reshuffled every period), and then permitted to reduce the earnings of others in his or her group, at a cost of 0.25 experimental dollars to the punisher per experimental dollar of reduction of the other's earnings (this is the first-order punishment stage). Subject  $i$ 's period earnings after first-order punishment are

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

where  $R_{ij}$  is the number of dollars by which  $i$  reduced  $j$ 's earnings, and conversely for  $R_{ji}$ . Should the RHS of (2) be negative,  $i$ 's period earnings would be reset to zero.<sup>6</sup> Again a subject's total earnings are the sum of the period earnings plus the participation fee. Variations of this VCM design with punishment have now been done many times, and we used results from two of our own VCMs with only first-order-punishment to compare with the results of the VCM with first-and-second-order punishment.

<sup>6</sup> More generally, constraints on first-order punishment were: (i) a subject could not spend more than her/his pre-punishment earnings for the period on reducing the earnings of other subjects, (ii) a subject's post-punishment earnings for a period would be set to zero if earnings  $y_i$  in equation (2) were negative, and (iii) a subject  $i$  could not spend more on reducing the earnings of a subject  $j$  in any period than would single-handedly reduce  $j$ 's earnings according in (2) to less than zero.

Second-order punishment is built on the VCM with first-order punishment as follows. In every third period after the first-order punishment, each subject was shown a list of the amounts by which other subjects had reduced the earnings of below-average contributors, of average contributors, and of above-average contributors.<sup>7</sup> Given this information about others' *patterns of punishment*, each subject was permitted to reduce the earnings of others, at a cost of 0.25 experimental dollars to the punisher per experimental dollar of reduction of the other's earnings (this is the second-order punishment stage).<sup>8</sup> As before, a subject's total earnings were the sum of the period earnings plus the participation fee.

What information to show before second-order punishment was a central design problem for our experiment. Listing punishments for each of eleven possible contribution levels for each subject would create clutter, but any coarser grouping of the data could be "leading." We chose the above, below, and equal to average format because it seemed neutral and flexible. Moreover, Fehr and Gächter's analysis (2000, p. 991) suggested a discontinuity in subjects' own views of contributions below and those above the group average. Finally, the structure lends itself to a simple classification of second-order punishment: punishing those who punished above-average contributors can be called normal, since it tends to reduce punishment of above-average contributors and thus to encourage higher contributions, and punishing those who punished below-average contributors can be called, correspondingly, perverse.

Our design choice also affects our operational definition of perverse and normal *first-order* punishment. Conceptually, first-order punishment is perverse if it tends to reduce contributions and, typically, efficiency (which decreases with decreasing earnings). In another study, we found evidence that punishing a group's highest contributor has a strong negative effect on her contribution and punishing the second highest contributor has little effect. In Table 2 below we explore the use of a definition of perverse first-order punishment based on punishing the highest contributor, but in the rest of the paper we use the "average" definition, which, while less sharply focused, conforms to our experimental design and appears to lead to the same qualitative results (see our Working Paper and Ertan et al. 2005).

Table 1 summarizes the four treatments in the experiment. All four have ten or more periods that included contribution and first-order punishment stages played in fixed four person groups (i.e., a partner design) but there are variations in each design. The treatment **P10** has first-order **Punishment** only, with **10** periods in the treatment. **P20** has first-order **Punishment** only, with **20** periods in the treatment. **PPG** has first-order **Punishment** and second-order

**Table 1** Details of the Treatment Conditions

Treatment	# Sessions	# Groups	# Subjects	# Periods	Second-Order Punishment
PPG	4	16	64	18	punish others in the group every third period
PPS	4	16	64	18	punish others in the session every third period
P10	3	12	48	10	No
P20	4	16	64	20	No

<sup>7</sup> Here, contributing the "average" means contributing an amount equal to the average contributed by the others in one's group of four in the period in question.

<sup>8</sup> As stated in the 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 that common proportion just sufficient to cause the constraint to be observed.

**Punishment**, and before second-order punishment every member of the **Group** is provided with information on the first-order punishment pattern of other members of the group and individuals are allowed to second-order punish others in their **Group**. **PPS** also includes first-order **Punishment** and second-order **Punishment**, and before second-order punishment every member of the **Session** is provided with information on the first-order punishment pattern of other members of the session and individuals are allowed to second-order punish others in their **Session**. The wording of the instructions of all four treatments was the same except for the differences in adding second-order punishment and in the number of periods.<sup>9</sup>

In our design of the treatments **PPG** and **PPS**, we limited second-order punishment to every third period (of an 18 period session) for two reasons. First, the information on the punishment pattern of each other subject takes time to absorb and evaluate, and we did not want the subjects to lose their concentration. Averaged over the preceding three periods, there was less information to process. Second, we thought that desires for revenge might be blunted, and a longer-term and more deliberative view be encouraged, by allowing more time to elapse between each second-order punishment stage.

In the two first-and-second-order designs, **PPG** and **PPS**, we varied the scope of information about others' punishment patterns and the opportunity to second-order punish the 3 other group members or the 15 other session members because we wanted to see if the inability to identify one's own group members might alter punishment behavior.<sup>10</sup> We conjectured that a low contributor's fervor for blind revenge might be diluted by the need to punish many people to get revenge. In contrast, a more altruistic subject might be less discouraged in punishing perverse first-order punishers in other groups as well as his/her own group.

The sessions of each treatment were conducted in a computer classroom at Brown University. Sixteen students drawn from the entire undergraduate population of the university (about 5800 students) participated in each session, for a total of 240 subjects.<sup>11</sup>

## 2. Results

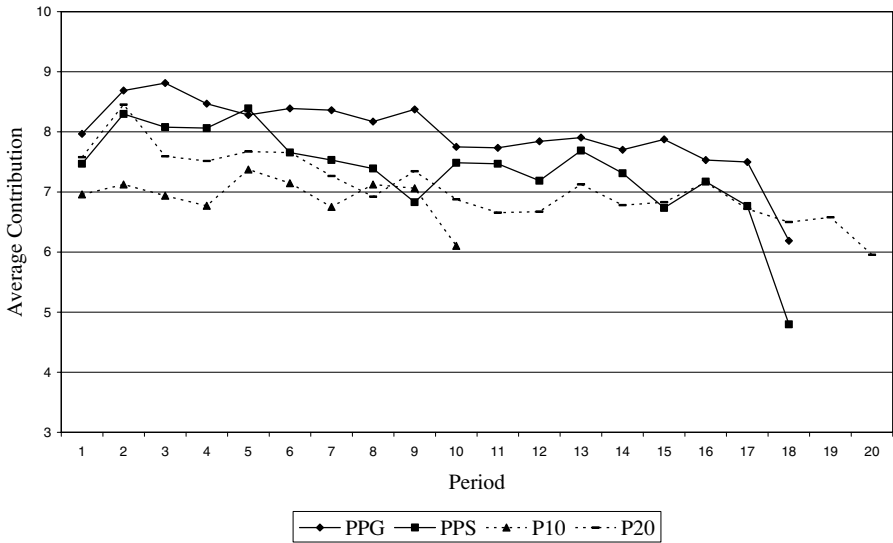
### 2.1. Contributions and earnings

Figure 1 shows the average contribution level in each period for **PPG** and **PPS**, and for **P10** and **P20** (the two first-order-only designs). The time paths of average contributions are similar, with initially high average contributions that are more sustained with repetition

<sup>9</sup> P10 is a treatment in Bochet, Page and Putterman (2006), there called Reduction or R treatment; P20 is a treatment in Page, Putterman and Unel 2005, there called Punishment treatment. Although those papers also consider treatments with communication or endogenous group formation, subjects in treatments P10 and P20 had neither involvement in nor information about such communication and regrouping.

<sup>10</sup> Carpenter and Matthews (2002) found evidence of altruistic behavior in a VCM of two groups where members of each group had the opportunity to first-order punish members of the other group, even though they had no material gain in doing so. See also Fehr and Fischbacher (2004). Our **PPS** treatment differs from these papers in that the opportunity for "out group" punishment arises only in our second-order punishment stage, and in that our subjects could not tell which individuals belonged to their own group, whereas own group and other group(s) were clearly differentiated in the punishment stage in their experiments.

<sup>11</sup> Subjects were recruited by flyers placed in their campus mailboxes or by an advertisement posted in an on-line magazine, the *Brown Daily Jolt*. 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 their 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.



**Fig. 1** Average Contribution

than in VCM experiments without punishment, until the familiar end-game declines. Average contributions in **PPG** and **PPS** are 12% and 4% higher than average contributions in **P20**, (see row 1 of Table 2), but neither difference is significant in Mann-Whitney tests.<sup>12</sup> Average contributions in **PPG** are significantly higher than average contributions in **P10**, but this difference may be due to the smaller number of periods in **P10**.<sup>13</sup> In summary:

*Result 1: Contributions were higher in treatments with second-order punishment than in treatments with only first-order punishment, but the difference is for the most part not significant.*

Figure 2 shows average earnings by period. In all four treatments earnings fluctuate between about 12 and 14 experimental dollars, which is about 80% of the maximum possible earnings. Earnings are higher in the two treatments with second-order punishment than in the two treatments with first-order punishment only, but the differences are small (see row 2 of Table 2). Mann-Whitney tests comparing average earnings<sup>14</sup> show no statistically significant differences among the **PPG**, **PPS** and **P20** treatments. Earnings in those three treatments are significantly higher than those in **P10**, but again this may be partly because of P10’s shorter duration.

<sup>12</sup> In the Mann-Whitney tests, each observation is the average contribution of all subjects in a group over all periods in a session for a total of 16 observations over the sessions from PPG, and so on, for the other treatments.

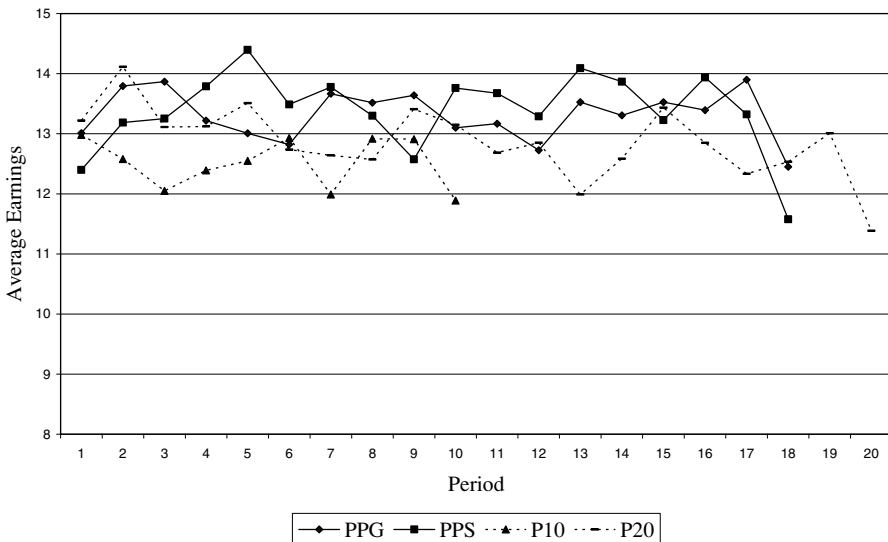
<sup>13</sup> Significant in a one-tailed Mann-Whitney test at the 5% level. But self-interested payoff maximizers have a weaker incentive to mimic conditional cooperators by signaling and building a reputation for cooperation, eliciting cooperation from actual conditional cooperators, when there are a smaller number of periods to benefit from the induced cooperation.

<sup>14</sup> An observation in these tests is the earnings of a group averaged over the periods of a session.

**Table 2** Contributions, Earnings and Punishment Costs Per Person Per Period

	PPG	PPS	P10	P20
(1) Contributions	7.97	7.35	6.94	7.09
(2) Earnings	13.31	13.38	12.52	12.86
(3) Costs of punishment	1.47	1.03	1.64	1.39
(3a) first-order punishment cost	0.96	0.74	1.64	1.39
(3b) second-order punishment cost	0.51	0.29	–	–
(4) Costs of perverse punishment	0.34	0.28	0.30	0.26
(4a) perverse first-order punishment cost	0.17	0.09	0.30	0.26
(4b) perverse second-order punishment cost	0.17	0.19	–	–

Notes: “Costs of punishment” are the direct costs to the punished plus the costs to the punisher. “Second-order punishment cost,” incurred only in periods 3, 6, etc., is given in per period terms by spreading it evenly over three periods, i.e., dividing by 3. In (4a), “perverse” punishment is defined as the punishment of a group’s highest contributor.



**Fig. 2** Average Earnings

Result 2: Earnings were higher in treatments with second-order punishment than in treatments with only first-order punishment, but the difference is for the most part not significant.

2.2. Frequency and costs of punishment

Seventy-five percent of subjects in **PPG** and 81% of subjects in **PPS** used the opportunity to first-order punish at least one time in their sessions. Sixty-four percent of subjects in the **PPG** treatment and 48% of subjects in the **PPS** treatment used the opportunity to second-order punish at least one time in their sessions.

Row (3a) of Table 2 shows that the (per person per period) direct costs<sup>15</sup> of first-order punishment in **PPG** and **PPS** are substantially lower than the direct cost of first-order punishment in **P20** and **P10**.<sup>16</sup> But when we add the cost of second-order punishment in **PPG** and **PPS** in row (3b), we find that the cost of first- and second-order punishment in **PPG** and **PPS** is roughly in the same range as punishment in **P20** and **P10**, which have no second-order punishment (see the aggregative row (3)). In this sense, punishment tends to “migrate” at least partially from first- to second-order punishment when the option of second-order punishment is allowed.

There is a similar and more pronounced migration of first- to second-order perverse punishment. To obtain a lower bound estimate in Table 2, we defined *perverse first-order punishment* in a period narrowly, including only punishing the highest contributor in that period (see our discussion in Section 1 above).

Row (4a) shows that the direct cost of perverse first-order punishment in **PPG** and **PPS** is about a third to a half of the cost of first-order perverse punishment in **P20** and **P10**. But the option of second-order punishment in **PPG** and **PPS** leads to perverse second-order punishing with cost about half of the cost of first-order perverse punishment in **P20** and **P10** (compare rows (4a) and (4b)). In this sense when the option of second-order punishment is allowed, perverse punishment “migrates” from the first- to second-order punishment opportunity. As we see in the aggregative row (4), the total (per person per period) cost of perverse punishment is about the same in the two treatments with first- and second-order punishment as in the two treatments with first-order punishment only.

*Result 3: A substantial number of subjects engaged in costly first- and second-order punishment in the PPG and PPS treatments. When the option of second-order punishment was allowed, first-order punishment “migrated” into second-order punishment, and first-order perverse punishment “migrated” into second-order perverse punishment.*

### 2.3. Who got punished in the PPG and PPS treatments

Table 3 estimates for PPG and PPS the amount of first-order punishment received by subject  $i$ , as determined by the same independent variables as in Fehr and Gächter (p. 991, 2000). All regressions include group and period dummy variables (not shown), and for each treatment, both OLS and tobit estimates are shown.<sup>17</sup>

In each of the four regressions, the absolute negative deviation<sup>18</sup> of subjects’ contributions from the average contribution of others in their group has a highly significant positive coefficient. The estimates suggest that in PPG for each dollar a subject’s contribution was below the group’s average the subject’s earnings were reduced by other group members by about 60 cents according to the OLS estimates, or \$1.10 according to the tobit estimates (a

<sup>15</sup> The *direct costs of punishment* are the monetary reductions of \$0.25 per unit of punishment to the punisher and \$1.00 per unit of punishment to the punished (see the right side of (2)).

<sup>16</sup> See our Working Paper where we also provide estimates of the indirect cost of perverse punishment.

<sup>17</sup> In principle, tobit estimates are preferred because in many observations a subject received no punishment, arguably a censored value. We include OLS estimates for comparability with estimates in other papers, including Fehr and Gächter (2000).

<sup>18</sup> Write  $C_i$  for  $i$ ’s contribution in a given period and  $C_{-i}$  for the average of others’ contributions in the group. Subject  $i$ ’s “positive deviation” is  $(C_i - C_{-i})$  if  $(C_i - C_{-i}) > 0$  and 0 otherwise. Subject  $i$ ’s “absolute negative deviation” is  $(C_{-i} - C_i)$  if  $(C_{-i} - C_i) > 0$  and 0 otherwise.

**Table 3** Determinants of First-Order Punishment Received in Treatments with Second-Order Punishment

	Dependent variable: First-Order Punishment Received by <i>i</i>			
	PPG		PPS	
	OLS	Tobit	OLS	Tobit
Positive deviation of <i>i</i> 's contribution from average in <i>i</i> 's group	0.070* (0.044)	−0.197 (0.132)	−0.008 (0.026)	−0.484*** (0.120)
Absolute negative deviation of <i>i</i> 's contribution from average in <i>i</i> 's group	0.599*** (0.048)	1.132*** (0.070)	0.588*** (0.044)	1.087*** (0.059)
Average contribution in <i>i</i> 's group	0.028 (0.050)	−0.126 (0.127)	−0.061* (0.037)	−0.258*** (0.084)
Constant	−0.691 (0.462)	−9.480*** (2.646)	0.162 (0.293)	−2.246*** (0.719)
No. of Observations	1152	1152	1152	1152
No. of Censored Observations		888		889
R_squared	0.51		0.53	
Pseudo R-squared		0.26		0.26

*Notes:* In this and later statistical tables the standard errors are in parentheses, and \*, \*\*, and \*\*\* indicate significance at the .05 level, the .01 level, and the .001 levels. See footnote 18 for definitions of “positive deviation” and “absolute negative deviation.”

60 cent punishment just offsets the marginal gain from the increase in free riding from shifting a dollar from the group account to the private account, for someone already contributing less than average). The estimates are less clear with respect to the likelihood of being punished when contributing above the group's average. The tendency for low contributors to be punished helps explain why contributions to the group account in **PPG** and **PPS** remained well above those typical of VCM experiments without punishment.

*Result 4: In the PPG and PPS treatments, the further below the group's average a subject contributed, the more first-order punishment the subject received.*

Table 4 estimates for **PPG** and **PPS** the amount of second-order punishment received by subject *j* in terms of his/her behaviors and the behaviors of others in his/her group. All six of the regressions in the table include group and period dummy variables (not shown) and tobit estimates only are shown to save space and in view of the large number of 0 observations for punishment.

In regressions (1) and (4) the total amount of second-order punishment aimed at subject *j* by others in her group (in the **PPG** treatment) or session (in the **PPS** treatment) is predicted by the three pieces of information on subjects' screens at the time of this decision, namely the amount by which *j* had reduced below-, above-, and average contributors in the previous three periods. These regressions indicate that giving first-order punishment of any kind, even to free riders, tended to attract second-order punishment. Moreover, the conjecture that subjects inclined toward perverse second-order punishment would do less of it when unable

**Table 4** Determinants of Second-Order Punishment

	Dependent variable: Second-Order Punishment Received by <i>j</i>					
	PPG			PPS		
	(1)	(2)	(3)	(4)	(5)	(6)
(a) <i>j</i> 's (normal) punishment of low contributors	0.291*** (0.067)	0.142*** (0.051)	0.120** (0.052)	0.854*** (0.093)	0.412*** (0.046)	0.394*** (0.045)
(b) <i>j</i> 's punishment of average contributors	0.520*** (0.112)	0.339*** (0.078)	0.303*** (0.077)	3.937 (2.691)	3.460*** (1.229)	3.350*** (1.232)
(c) <i>j</i> 's (perverse) punishment of high contributors	0.888*** (0.072)	0.429*** (0.05)	0.317*** (0.075)	0.640*** (0.214)	0.291*** (0.094)	0.280*** (0.093)
(d) average positive deviation of <i>i</i> 's contributions		-0.096 (0.199)	0.125 (0.209)		0.381*** (0.139)	0.411*** (0.139)
(e) average absolute negative deviation of <i>i</i> 's contributions		0.321** (0.146)	0.331** (0.168)		0.468*** (0.106)	0.357*** (0.125)
(f) <i>i</i> 's deviation times <i>j</i> 's punishment of low contributors			-0.033* (0.017)			-0.050*** (0.018)
(g) <i>i</i> 's deviation times <i>j</i> 's punishment of average contributors			4.814*** (1.614)			0.728 (0.617)
(h) <i>i</i> 's deviation times <i>j</i> 's punishment of high contributors			0.053** (0.025)			0.119*** (0.040)
Constant	-3.588*** (1.141)	-6.075*** (1.063)	-5.988*** (1.046)	-6.723*** (1.346)	-11.551*** (1.074)	-1.210*** (1.053)
No. of Observations	384	1152	1152	384	5760	5760
No. of Censored Observations	256	981	981	312	5620	5620
Pseudo R-squared	0.26	0.20	0.21	0.20	0.14	0.15

*Notes:* “Positive” and “absolute negative deviation” are defined as in Table 3 (see footnote 18). “Average” in (d) and (e) is the average over 3 periods between second-order punishments. Subject “*i*'s deviation” in rows (f)–(h) is *i*'s contribution minus the average contributed by others in *i*'s group. A “low contributor” is a lower-than-average contributor in a group and period and a “high contributor” is a higher-than-average contributor in a group and period.

to determine which of the potential targets belonged to their own group is not supported by the coefficients in rows (a) and (c).<sup>19</sup>

*Result 5: Second-order punishers targeted both those who punished high contributors and those who punished low contributors.*

#### 2.4. Who targets whom for second-order punishment?

Regressions (2), (3), (5), and (6) of Table 4 investigate for each treatment the impact of both punished and punisher characteristics on the amount of second-order punishment one gave to the other. These regressions have 3 (**PPG**) or 15 (**PPS**) times as many observations as regressions (1) and (4), because they include a distinct observation for each  $(i, j)$  pair as well as for each second-order punishment stage. Regressions (2) and (5) add, to the information about the targeted individual  $j$ 's first-order punishment behavior, information about second-order punisher  $i$ 's tendency to contribute above or below the average of his group. The coefficients in rows (a), (b), and (c), qualitatively consistent with the regression (1) and (4) estimates summarized in Result 5, indicate that all kinds of first-order punishment attracted second-order punishment. The added coefficients in rows (d) and (e) suggest that in the **PPG** treatment, lower contributors gave more second-order punishment than average contributors, while in the **PPS** treatment, both relatively low and relatively high contributors gave more second-order punishment than average contributors.

Regressions (3) and (6) add to regressions (2) and (5) a set of explanatory variables which are the average deviation of punisher  $i$ 's contribution from his group average multiplied by each of the first three explanatory variables. The original five coefficients remain qualitatively as before, while the added coefficients in rows (f), (g), and (h) indicate that it was mainly low contributors who (perversely) second-order punished those who engaged in punishment of low contributors, while it was mainly high contributors who punished those who engaged in perverse punishment of high contributors.

*Result 6: In PPG and PPS, high contributors tended to second-order punish normally, while low contributors tended to second-order punish perversely.*

#### 2.5. Effects of second-order punishment

Table 5 shows results of regressions to study the impact of receiving second-order punishment on giving first-order punishment. In regressions (1) and (2), the dependent variable is the change in subject  $i$ 's (perverse) punishment of above-average contributors from the previous three periods to the three periods following a second-order punishment stage. The explanatory variable is the amount of (normal) second-order punishment which  $i$  received for (perversely) punishing above average contributors in that stage. In regressions (3) and (4) the dependent variable is the change in  $i$ 's (normal) punishment of below average contributors, and the explanatory variable is the amount of (perverse) second-order punishment received for (normally) punishing below average contributors.<sup>20</sup>

<sup>19</sup> Specifically, take the ratio of the coefficient in (c) to the coefficient in (a). While our conjecture implies that this ratio should be larger in PPG than in PPS, the converse holds when comparing columns (4) and (1), columns (5) and (2), and columns (6) and (3) in Table 4.

<sup>20</sup> To define the dependent variable for regression (3) pick a particular subject  $i$  in a particular group of other members  $j, k,$  and  $l$ , for one of the second-order punishment stages other than the last one (the other observations for this and other dependent variables are defined correspondingly). If  $j$  made a below-average contribution

**Table 5** Impact of Second-Order Punishment on First-Order Punishment

	Change in $i$ 's first-order perverse punishing of above-average contributors		Change in $i$ 's first-order normal punishing of below-average contributors	
	(1) PPG	(2) PPS	(3) PPG	(4) PPS
(a) Second-order punishment $i$ received for perverse punishing above-average contributors	-0.030*** (0.003)	-0.204*** (0.014)		
(b) Second-order punishment $i$ received for normal punishing below-average contributors			-0.017*** (0.004)	-0.025*** (0.006)
Constant	0.008 (0.007)	-0.001 (0.009)	0.010 (0.011)	0.008 (0.011)
No. of Observations	320	320	320	320
Adjusted R Squared	.243	.382	.043	.057

Note: OLS regressions.

The regressions support the conjecture that second-order punishment discouraged both normal and perverse first-order punishers from persisting in those punishment behaviors. For the **PPG** treatment, the coefficients indicate that every dollar of second-order normal punishment given to a perverse first-order punisher led to a reduction of 3 cents in the first-order punisher's perverse punishing, for each of the next three periods (see row (a), regression (1)). And for the **PPG** treatment, every dollar of second-order perverse punishment given to a normal first-order punisher led to a reduction of 1.7 cents in the first-order punisher's normal punishing, for each of the next three periods (see row (b) regression (3)). For the **PPS** treatment, the corresponding numbers are 20.4 and 2.5 cents, respectively. Interestingly, the coefficients are evidence of a larger change in perverse first-order punishment (columns (1) and (2)) than in normal first-order punishment (columns (3) and (4)) per dollar of second-order punishment received.

in the period of the second-order punishment stage, divide  $i$ 's punishment of  $j$  (possibly zero) by  $j$ 's absolute negative deviation, getting " $i$ 's punishment of  $j$  per unit deviation" for that period. Do the same for each of the two periods just prior to this period, and make the corresponding calculations for  $k$  and  $l$ , if either made a below-average contribution. The average of these "punishment per unit deviation" values is " $i$ 's first-order normal punishing of below-average contributors before the second-order punishment stage." Correspondingly calculate " $i$ 's first-order normal punishing of below average contributors for the three periods after the second-order punishment stage." The difference between these two measures is the dependent variable "change in  $i$ 's first-order punishing of below average contributors." The "change in  $i$ 's first-order punishing of above average contributors" is calculated correspondingly. The independent variables are defined as follows. In case (a): if all of  $i$ 's first-order punishment of  $j$ ,  $k$ , and  $l$  in the 3 periods prior to this second-order punishment stage were normal, then we count all the second-order punishment  $i$  received in this stage as "second-order punishment  $i$  received for normal punishing below-average contributors," and in case (b) correspondingly if  $i$ 's punishment was all perverse in these 3 periods. In case (c) if  $i$  punished both perversely and normally in the 3 periods, the amount of second-order punishment  $i$  received is divided between punishment for normal punishing and punishment for perverse punishing in proportion to  $i$ 's first-order normal and perverse punishing in the 3 periods (surprisingly, only 5% of the cases were in case (c)). OLS is used because there are no common limiting values on the dependent variables.

*Result 7: The more second-order punishment a subject received, the more the subject reduced his/her first-order punishing in the following three periods.*

## 2.6. Evidence of non-strategic second-order punishment

Like first-order punishment, second-order punishment in a repeated interaction with fixed groups might be explained (in part, at least) as being intended to influence others' future choices for the benefit of one's own future payoff. Since no such strategic end can be accomplished in the last second-order punishment stage, the hypothesis that second-order punishment is strategically motivated can be tested by comparing the amount of punishment given in that period with amounts in earlier periods. Mann-Whitney tests using average second-order punishment in each group and stage as observations find no difference in the amount of second-order punishment given in the final second-order stage versus the other five second-order stages to those who first-order punished above-average contributors. Corresponding Mann-Whitney tests for second-order punishment of those who first-order punished low contributors – “perverse second-order punishment” – show that it was lower in the last period than in the other five second-order punishment stages.

*Result 8: Second-order punishment persisted into the final period, indicating that it was not (or not entirely) strategically motivated.*

## 3. Discussion and conclusion

In response to our title question “can second-order punishment deter perverse punishment?,” we find “only a little if at all” in this experiment. Allowing second-order punishment significantly reduced first-order perverse punishment, but this reduction was offset by perverse second-order punishment of those who (first-order) punished free riders. A main conclusion from the experiment is that perverse punishment is a persistent phenomenon, capable of migrating from one domain into another.

Interest in higher-order punishment is sparked by the theoretical literature on cooperation. Evolutionary theorists have long puzzled over how the tendency to engage in costly punishment of free riders could have evolved, since both punishers and non-punishers benefit from the increased contributions that the threat of punishment induces but only the punishers bear the cost. In response to this problem, Henrich and Boyd (2001) have demonstrated that second-order punishment can stabilize first-order punishment and both can be evolutionarily stable strategies under certain conditions.<sup>21</sup> Since our experiment permits group members to punish those who fail to punish free riders, it can be viewed as a laboratory test of whether higher-order punishment is used as Henrich and Boyd have modeled it, to punish those who fail to punish free riding. In the experiment, punishers of free riders and punishers of high contributors received more second-order punishment than did those who “free rode” on others' punishment, putting the latter at a further advantage, not disadvantage.<sup>22</sup>

<sup>21</sup> The argument is also presented in Henrich (2004), which serves as the focus for 17 commentaries in a special issue (January, 2004, vol. 53, no. 1) of the *Journal of Economic Behavior and Organization*. Earlier, Axelrod (1986) discussed the norm of punishing both those not conforming to a rule and those failing to punish them. He labeled “a norm that one must punish those who do not punish a defection” a *metanorm*.

<sup>22</sup> Note that not-punishing at all is the default condition relative to which the coefficients in rows (a), (b) and (c) of Table 4 are estimated. The finding that all of those coefficients are significantly positive implies that the way to assure that one received the least amount of second-order punishment in our experiment was not to first-order punish at all.

Although second-order punishment fails to significantly increase contributions and efficiency in our experiment, it still seems possible that second-order punishment might be part of a successful mechanism if supplemented by other elements. We conjectured that making individuals' punishing activities public and allowing punishers to be punished might allow norms regarding what is and is not appropriate sanctioning to arise. This may happen in the cases of real-world cooperation such as described by Ostrom (1990) more than it does in our lab because in the real world, group members can frequently engage also in face-to-face communication. Indeed, Ostrom, Gardner and Walker (1994) have obtained results from laboratory experiments concerning common pool resource problems that show that combining communication with sanctions can be very effective.

**Acknowledgments** 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 we are grateful to Xiaotong Wang, Ioannis Garos, and Ted Marr for their roles in various stages of the data analysis.

## References

- Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, 80(4), 1095–1111
- Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior and Organization*, 60, 11–26
- Carpenter, J., & Matthews, P. (2002). Social reciprocity. Middlebury College Department of Economics Working Paper #29
- Cinyabuguma, M., Page, T., & Putterman, L. (2004). On perverse and second-order punishment in public goods experiments with decentralized sanctioning. Brown University Department of Economics Working Paper 2004-12
- Davis, D. D., & Holt, C. (1993). *Experimental economics*. Princeton: Princeton University Press
- Ertan, A., Page, T., & Putterman, L. (2005). Can endogenously chosen institutions mitigate the free-rider problem and reduce perverse punishment? Department of Economics Working Paper, Brown University 2005-13
- Fehr, E., & Fishbach, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, 25, 63–87
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment. *American Economic Review*, 90, 980–994
- Henrich, J. (2004). Cultural group selection, co-evolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization*, 53(1), 3–35
- Henrich, J., & Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, 208, 78–89
- Ledyard, J. (1995). Public goods: A survey of experimental research. In J. Kagel and A. Roth (Eds.), *Handbook of Experimental Economics*, (pp. 111–94). Princeton: Princeton University Press
- Nikiforakis, N. (2004). Punishment and counter-punishment in public goods games: Can we still govern ourselves? unpublished paper, Royal Holloway, University of London
- Ostrom, E. (1992). *Governing the commons*. New York: Cambridge University Press
- Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self governance is possible. *American Political Science Review*, 86(2), 404–416
- Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, games and common-pool resources*. Ann Arbor: University of Michigan Press
- Page, T., Putterman, L., & Unel, B. (2005). Voluntary association in public goods experiments: reciprocity, mimicry, and efficiency. *Economic Journal*, 115, 1032–1053
- Saijo, T., & Nakamura, H. (1995). The 'spite' dilemma in voluntary contribution mechanism experiments. *Journal of Conflict Resolution*, 38(3), 535–560
- Sefton, M., Shupp, R., & Walker, J. (2002). The effect of rewards and sanctions in provision of public goods. Working Paper, University of Nottingham and Indiana University
- Somerville, J., & Santoni, R. E. (eds.) (1963). *Social and political philosophy: readings from Plato to Gandhi*. Garden City, NY: Anchor Books