ELSEVIER

Journal of Public Economics 92 (2008) 91–112

JOURNAL OF PUBLIC ECONOMICS

www.elsevier.com/locate/econbase

Punishment and counter-punishment in public good games: Can we really govern ourselves?[☆]

Nikos Nikiforakis [*]

Department of Economics, The University of Melbourne, Victoria 3010, Australia

Received 8 May 2005; received in revised form 25 April 2007; accepted 27 April 2007
Available online 5 May 2007

Abstract

A number of experimental studies have shown that the opportunity to punish anti-social behavior increases cooperation levels when agents face a social dilemma. Using a public good experiment, I show that in the presence of counter-punishment opportunities cooperators are less willing to punish free riders. As a result, cooperation breaks down and groups have lower earnings in comparison to a treatment without punishments where free riding is predominant. Approximately one quarter of all punishments are retaliated. Counter-punishments appear to be driven partly by strategic considerations and partly by a desire to reciprocate punishments.

JEL classification: C92; D70; H41
Keywords: Public goods; Counter-punishment; Revenge; Decentralized punishment; Experimental economics

1. Introduction

A topic of continuing interest in public economics is how cooperation can emerge when there is a conflict between social and individual interest. Standard economic theory, based on the assumption that all agents maximize their personal utility without regard for others, predicts that a socially sub-optimal outcome will be produced in these cases unless legally enforceable contracts exist to force agents to cooperate. This implies, amongst others, that a decentralized mechanism cannot be relied upon to efficiently provide public goods (Samuelson, 1954; Olson, 1965).

[☆] The title is a reference to Ostrom et al. (1992). Instructions, experimental software, and data can be downloaded from http://www.economics.unimelb.edu.au/nnikiforakis/research.htm.

[*] Tel.: +61 383449717; fax: +61 38344 5104.
E-mail address: n.nikiforakis@unimelb.edu.au.

0047-2727/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2007.04.008

Using the public good game as a testing ground, experimental economists have studied these predictions repeatedly. In the public good game agents have to make a decision concerning how much they wish to contribute to a public account; the higher an agent's contribution to the public account the higher the group payoff. However, every agent also has an incentive to free ride and not to contribute; hence the social dilemma. The experimental results indicate that the problem of free riding is indeed pervasive and leads to the under-provision of public goods (Ledyard, 1995).

Apart from testing theoretical predictions, experimental economists have contributed to the quest for a solution to the free-rider problem. One solution which has received considerable attention is the use of decentralized punishments (that is, punishments carried out by individuals without the intervention of a central authority) to discipline free riders in social dilemma experiments. Ostrom et al. (1992) show that the existence of such punishment opportunities in a common-pool resource game increases cooperation between appropriators significantly. The same result is reported in Fehr and Gächter (2000, 2002) who study public good games. In contrast to Ostrom et al. (1992), in some treatments of Fehr and Gächter (2000, 2002) strategic motives for punishing are removed as agents only meet each other once in the experiment. Since punishment is costly, a money-maximizing individual will never punish in equilibrium. Nevertheless, Fehr and Gächter (2000, 2002) report frequent punishments even in the last period of the experiment. The fear of punishment has a strong positive effect on cooperation as in Ostrom et al. (1992).[1]

The efficacy of decentralized punishments in improving cooperation prospects between agents is well-reported; a number of studies have used the experimental design of Fehr and Gächter (2000, 2002) to address closely related research questions and have replicated the original results.[2] A common feature of these studies is that they only allow for a single punishment stage. However, this deprives agents of the opportunity to take revenge for imposed punishments — an opportunity that exists in almost every decentralized interaction in practice. Indeed, Elster (1990 p.862) defines revenge as “the attempt, at some cost or risk to oneself, to impose suffering upon those who have made one suffer”. This definition implies that it is not only cooperators who might wish to punish free riders, but free riders might also be eager to counter-punish in response to a punishment.

The aim of this paper is to examine whether and why punished individuals counter-punish, and to test how the existence of counter-punishment opportunities affects cooperation prospects, welfare, and the willingness of agents to punish anti-social behavior in a public good game with no regulatory authority or formal contracts.

Omitting from the analysis the threat that counter-punishments pose might lead to an overestimation of the efficacy of decentralized punishment mechanisms in promoting cooperation, misleading conclusions and, perhaps, the implementation of inappropriate policies. To see how this is possible consider the following example. Nation *A* contemplates violating a treaty with Nation *B*. Nation *B* has decided that, should the violation occur, Nation *B* will consider imposing financial sanctions on Nation *A*. Nation *A* in anticipation of *B*'s sanctions has prepared for military counter-sanctions. If all threats are credible and the consequences are severe enough, ignoring the possibility of military counter-sanctions by Nation *A* leads to the conclusion

[1] Punishments carried out by a central authority have a similar positive effect on cooperation. See Samuelson et al. (1986), Sato (1987), and Yamagishi (1986, 1988).

[2] See for example Anderson and Putterman (2006), Bochet et al. (2006), Carpenter (in press, 2007), Masclet et al. (2003), Nikiforakis and Normann (in press), Nikiforakis et al. (2007), Noussair and Tucker (2005), Page et al. (2005), Sefton et al. (2005).

that Nation A, knowing that Nation B will punish its defection, will not violate the treaty. However, if Nation B recognizes the possibility of war, it will be unwilling to sanction Nation A if the conflict costs together with the cost of imposing financial sanctions outweigh the benefits from the treaty. In that case, Nation A will violate the treaty.[3]

The introduction of counter-punishment opportunities is also likely to have an effect on individual earnings. The direction of this effect is difficult to predict due to three competing forces. First, counter-punishments are costly for victims and culprits and, hence, tend to lower earnings, ceteris paribus. Second, as discussed in the previous paragraph, the threat of counter-punishment might limit the occurrence of punishments by making them unprofitable. That is, counter-punishment opportunities can have an indirect positive effect on earnings. As will be discussed in Section 3.4, this effect might not be negligible as participants often overuse punishments in public good experiments. Third, the weakening of the punishment threat is likely to make free riding profitable for some individuals. As a result, there will be a decline in cooperation levels which will influence earnings negatively. The difficulty in predicting the impact of counter-punishment opportunities on earnings, which is one way of measuring welfare, provides additional motivation for the present study.

The results from the experiment cast doubt over the prevailing optimism in the literature that “self-governance is possible” (Ostrom et al., 1992). Approximately one quarter of all punishments lead to retaliation. Under the threat of counter-punishment, individuals are less willing to punish free riders. This leads to the breakdown of cooperation. Groups in treatments where punishments (of any type) are allowed have lower earnings than groups in treatments without punishment opportunities where free riding is predominant. Counter-punishments appear to be the result partly of strategic behavior and partly of a desire to reciprocate punishments.

The remainder of the paper is structured as follows: Section 2 introduces the experimental design and discusses hypotheses for each of the treatments. Section 3 presents the results, while Section 4 discusses the robustness of the results and concludes.

2. Experimental design

The experiment consists of three treatments: One without any punishment – the standard public good game which is also known as the *voluntary contribution mechanism* (VCM); one with *one-sided punishment* – that is, a VCM followed by a single punishment stage (P); and one with *two-sided punishment* – that is, a VCM followed by a stage of punishment and a stage of counter-punishment (PCP). Instructions for all treatments are written in a neutral language and are adopted from Fehr and Gächter (2000).

Each treatment consists of 10 periods. The repetition of a task over 10 periods gives participants the opportunity to learn. However, it also creates a problem: If the composition of groups remains unchanged throughout the experiment (*fixed matching*) incomplete information about the motivation of other participants can change the nature of the equilibrium by giving rise to reputation effects (e.g. Kreps et al., 1982). To overcome this problem, and to test static game-theoretic predictions, experimental economists typically use *random matching*, where participants are randomly re-matched

[3] One of the referees offered an additional example. Two agents are trying to settle a case out of court. An agent who knows that he will lose the court case is likely to cooperate and settle out of court. However, if there exists a possibility of counter-litigation against the plaintiff (or any other means of threatening the plaintiff), then cooperation and out of court settlement becomes less likely.

Table 1
Experimental design

Session	Matching protocol	Number of participants	Order of treatments	Average	Contribution
				VCM	PCP/P
1	Fixed	12	PCP–VCM	2.25	7.69
2	Fixed	12	PCP–VCM	5.80	7.53
3	Fixed	12	VCM–PCP	4.83	9.05
4	Fixed	12	VCM–PCP	4.01	10.61
5	Fixed	12	P–VCM	6.50	13.90
6	Fixed	12	P–VCM	6.99	15.50
7	Fixed	12	VCM–P	6.35	14.78
8	Fixed	12	VCM–P	4.56	17.13
9	Random	12	PCP–VCM	3.97	6.80
10	Random	12	PCP–VCM	6.03	11.79
11	Random	12	VCM–PCP	3.55	2.46
12	Random	12	VCM–PCP	3.24	5.83
13	Random	12	P–VCM	6.48	14.77
14	Random	12	P–VCM	8.36	10.99
15	Random	12	VCM–P	6.94	10.41
16	Random	12	VCM–P	4.73	10.36

Each session had 10 periods of one treatment followed by 10 periods of a second treatment

in each period.[4] To have a better understanding of the motivations behind counter-punishment and its effect on individual behavior I run the treatments under both fixed and random matching.

In each session, 12 subjects are randomly divided into groups of 4 people and play a finitely repeated public good game for 20 periods. For the first 10 periods participants are exposed to one of the three treatments and for the last 10 periods to another.[5] To test for sequence effects, in half the sessions the order of the treatments is reversed. The experimental design is summarized in Table 1. I proceed by discussing each of the treatments in detail.

2.1. The VCM treatment

At the beginning of each of the ten periods, every participant receives a fixed amount of 20 Experimental Currency Units (ECU). Participants then decide simultaneously and without communication how much of their endowment to contribute to a public account, c_i , where $0 \leq c_i \leq 20$. The rest $(20-c_i)$ remains in the player's own account. In addition to the money that player i keeps, i receives a fixed return equal to 40% of the group's total contribution to the public account. The earnings for each subject in a given period are thus given by:

$$\pi_i = 20 - c_i + 0.4^* \sum_{h=1}^{4} c_h. \tag{1}$$

[4] See Botelho et al. (2005b) for a critical review of the literature on fixed and random matching. Botelho et al. also compare behavior under random and perfect-random matching (where individuals do not meet each other more than once). The authors report contamination effects even under random matching: Individuals are more likely to contribute to the production of a public good under random matching than under perfect-random matching.

[5] There is no session where both P and PCP are played. VCM is coupled with either P or PCP. Following Fehr and Gächter (2000), to prevent results from the first treatment being affected by the existence of a second treatment, subjects are not aware that a second treatment is to follow. In the beginning of the second treatment the participants are assured that the experiment will finish after the last period.

At the end of each period subjects are reminded of their contribution and are informed of the group's total contribution, their income from the project and their income for period *t* as given by Eq. (1).

2.2. The P treatment

The treatment with one-sided punishment is identical to the one by Fehr and Gächter (2000). In this treatment, one stage is added after the contribution stage of the VCM. At the beginning of this stage participants are informed of the individual contributions in their group and are given the opportunity to punish each other simultaneously. To punish, group member *i* has to assign *punishment points* to group member *j*, p_{ij} , $i \neq j$. The punishment of non-group members is not possible. Punishment has two distinct effects on the payoffs of *i* and *j*: Each point assigned to *j* reduces her income from the first stage, π_j^1 , by 10%; the first-stage income can never be reduced below zero, so, if player *j* receives more than 10 punishment points, her income is reduced by 100%. In addition, player *i* also faces a cost for distributing punishment points to player *j*, which is given by the convex cost function, $k(p_{ij})$, shown in Table 2. Player *i*'s cost from distributing points cannot exceed his income from the first stage, that is, $\Sigma_{i \neq j}\, k(p_{ij}) \leq \pi_i^1$. The earnings of subject *i* at the end of the second stage are equal to:

$$\pi_i^2 = \pi_i^1 * \left[\frac{\max\{0, 10 - \sum_{j \neq i} p_{ji}\}}{10}\right] - \sum_{j \neq i} k(p_{ij}). \tag{2}$$

Eq. (2) implies that a subject could have a negative payoff in a given period. Participants were warned of this possibility. However, this would only happen in some extreme cases in which subjects attracted considerable punishment and also decided to punish. Indeed, this happened only in 16 out of a possible 960 cases (96 subjects × 10 periods).

As in stage one, punishment decisions are made simultaneously and without communication. At the end of each period subjects are reminded of their income from the first stage, the punishment points they assigned in total and the associated cost, and are also informed of the punishment points they received in total from the group, the associated income reduction, as well as their total income from period *t* as given by Eq. (2).

2.3. The PCP treatment

PCP is the central treatment of this study as it gives participants the ability to counter-punish. In the PCP treatment, a third, and final, stage is added after the punishment stage. At the beginning of the third stage subjects are informed of the number of points each of the other group members assigned to them. They are then given the opportunity to reduce the income of the individuals who punished them during the second stage by assigning counter-points. The cost of the counter-points that *i* assigns to *j*, $k(cp_{ij})$, is also given by the cost function in Table 2, where cp_{ij} is the counter-points *i* assigns to *j*. The cost of punishing works cumulatively. That is, if

Table 2
Punishment points per player *j* and associated costs for punisher *i*

p_{ij}	0	1	2	3	4	5	6	7	8	9	10
$k(p_{ij})$	0	1	2	4	6	9	12	16	20	25	30

player *i* punishes player *j* with two points during the second stage of a given period and then with two further (counter-) points in the third stage, *i*'s total cost from assigning points will be equal to the cost of four points. The end-of-period income is given by the following equation:

$$\pi_i^3 = \pi_i^2 * \left[\frac{\max\{0, 10 - \sum_{j \neq i} cp_{ji}\}}{10} \right] - \sum_{j \neq i} [k(p_{ij} + cp_{ij}) - k(p_{ij})]. \tag{3}$$

Participants with a non-positive payoff at the beginning of the third stage are not allowed to counter-punish or to be counter-punished. To avoid strategic delay of punishment, only the subjects who are punished in the second stage are allowed to assign points in the third stage, and they are only allowed to assign points to those who punished them in the second stage. Allowing for strategic delay of punishment would have been inappropriate to study the net effect of counter-punishment as individuals could avoid retaliation by delaying punishment until the final punishment stage of a given period.

By posing restrictions on the types of punishments that can be carried out in the second punishment stage it is possible that I overestimate the negative effect of counter-punishment on cooperation and earnings. In the last section, I discuss recent experimental evidence by Denant-Boemont et al. (in press) indicating that this does not appear to be the case.

At the end of each period subjects in PCP are reminded of their income from the second stage, the counter-points they assigned in total and the associated cost. They are also informed of the total number of counter-points they received and the associated income reduction, as well as their end-of-period income as given by Eq. (3).

In all three treatments the endowment, the return from the project, the number of group members, the relevant payoff functions, the cost function for purchasing points (not available in VCM), and the duration of the treatment are common knowledge to the subjects. This is achieved by the experimenter reading aloud a pre-written summary of the instructions. The total earnings of a subject is equal to the sum of earnings over all periods. In the treatments with punishment opportunities (P and PCP), each subject is given a one-off lump-sum payment of 25 ECU at the beginning of the experiment to pay for any negative payoffs the participant might have incurred during the experiment.

To prevent the formation of individual reputation, at the beginning of each period subjects are randomly given a "name" (a number between 1 and 4) to distinguish their actions from those of the others within a period. Reassigning names in each period ensures that, even when group members remain the same, participants cannot create a link between the actions of other subjects across periods.

2.4. Procedures

There were three series of experiments which took place in the experimental laboratory of Royal Holloway, University of London: The first between December 2003 and March 2004, the second in February 2005 and the third in October 2005. Each session lasted approximately one hour forty-five minutes (the treatments without counter-punishment lasted slightly less).

In the beginning of each session subjects read the instructions before completing a control questionnaire. The experimenter (or an assistant) then checked and, if necessary, explained the answers to the subjects using a pre-written text. Once all subjects' answers were checked the experimenter read aloud a summary of the instructions.

Each of the 192 participants was recruited for one session via e-mail from a list of voluntary participants. None of them had previously participated in a public good experiment. Subjects were students of different nationalities and academic backgrounds, including Economics. The experiment was conducted using *z*-Tree (Fischbacher, 2007). To avoid an experimenter effect all sessions were run by the same individual. Participants earned on average £18. The exchange rate was 1 ECU=0.4 pence and no show-up fee was given apart from the 25 ECU.

2.5. Hypotheses

In this section I derive hypotheses for behavior across the three treatments. In the subgame perfect Nash equilibrium, under the standard assumption that individuals try to maximize their own monetary payoff, neither punishments nor counter-punishments will occur as they are costly for the punisher. In the contribution stage individuals will contribute zero as this is their dominant strategy given that punishments do not take place. That is, in the subgame perfect Nash equilibrium of all three treatments contributions will be equal to zero.

A number of public good experiments have shown that the subgame perfect Nash equilibrium can predict to some extent behavior in the VCM: Although contributions are typically higher than predicted in the early periods, there is a constant convergence towards the subgame perfect Nash equilibrium (Ledyard, 1995). However, this is not the case if punishment opportunities exist as in treatment P; a well-documented fact is that the introduction of punishment opportunities increases contributions significantly. This regularity can be rationalized by imperfect Nash equilibria, that is, Nash equilibria which are non-subgame perfect. Consider, for example, the following simple (symmetric) strategy for i. In every period contribute all of your endowment to the public account, i.e. $c_i=20$. In the event of a deviation, punish deviator j by $p_{ij}=\tilde{p}>0$, such that a defection is not profitable. This strategy is part of an imperfect Nash equilibrium as the punishment threat is non-credible if individuals are fully rational and care only for their own payoff. However, the existence of punishment opportunities might help to sustain cooperation in equilibrium if individuals have other-regarding preferences (for a survey see Fehr and Schmidt, 2003), or incomplete information about the players' types (Kreps et al., 1982; Kreps and Wilson, 1982; Milgrom and Roberts, 1982). This implies that, if punishment opportunities exist, then cooperation could be sustained at a high level. That is, contributions are likely to be higher in P than in VCM as observed in a number of experiments.

Given the multiplicity of imperfect Nash equilibria in a repeated game, it is hard to derive a definite hypothesis for contribution levels in the PCP treatment. However, one can reason as follows: As the cost of punishment increases, the quantity demanded of punishment will decrease, ceteris paribus, and, therefore, the punishment threat that sustains cooperation will become weaker. This hypothesis is in agreement with experimental findings (Anderson and Putterman, 2006, Carpenter, 2007; Nikiforakis and Normann, in press; Ostrom et al., 1992) and the model of inequity aversion by Fehr and Schmidt (1999, p.841).[6] This implies that cooperation will become less stable if the punishment cost increases. The threat of counter-punishment can be seen as raising the (expected) cost of punishment. Consequently, both imperfect Nash equilibria and Fehr and Schmidt (1999) predict less severe punishments in PCP than in P and, ultimately, lower contribution levels. I summarize the hypotheses.

[6] Counter-punishments are not predicted by the model of Fehr and Schmidt (1999) who construct punishments so that they nullify any payoff differences after the first punishment stage.

Hypothesis 1. The introduction of counter-punishment opportunities will decrease the punishment of a defection. That is, ceteris paribus, punishment will be higher in P than in PCP.

Hypothesis 2. If punishment is higher in P than in PCP, then contributions will be higher in P than in PCP.

3. Results

I begin by discussing contributions to the public account across treatments. I, then, examine punishment behavior in treatments P and PCP before analyzing the determinants of counter-punishment. Finally, I discuss welfare in each of the treatments.

For the statistical analysis, I use parametric tests to account for different factors that might be affecting the outcome and to control for panel effects. None of the paper's main results depend on the method used for hypothesis testing; non-parametric, as well as other parametric tests, yield the same results.

To control for the interaction of participants across periods all regressions include cluster random effects, where clusters are taken to be groups under fixed matching, and sessions under random matching (see footnote 5). To test the robustness of the results which are presented in the following section, I also estimated the models with individual *and* cluster random effects using Generalized Linear Latent and Mixed Models (Rabe-Hesketh and Skrondal, 2005). The results from the two estimations are very similar. In the following section I report the results from the regressions using only cluster random effects. This allows me to estimate the average marginal effects of the independent variables, their standard errors, and their *p*-values following Bartus (2005).

3.1. Contributions to the public account

Figs. 1 and 2 display the evolution of average contributions over the 10 periods by pooling data for each treatment under fixed and random matching, respectively. In line with previous

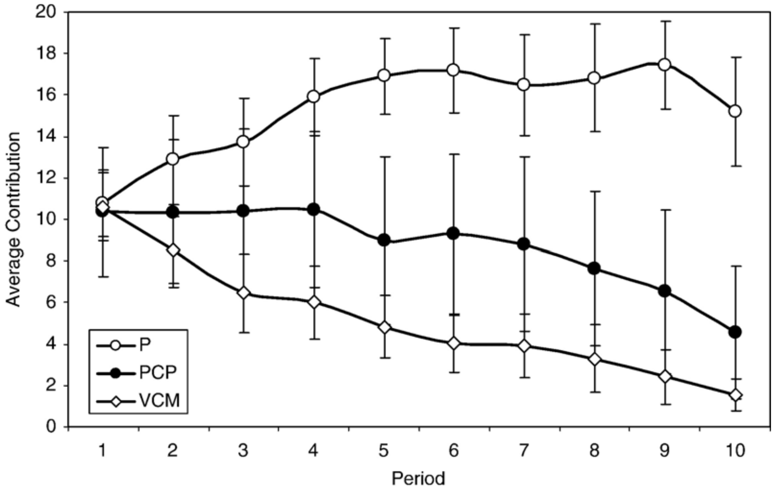

Fig. 1. Evolution of average contribution under fixed matching with 95% confidence intervals. Note: VCM is drawn from sessions 1–8, PCP from sessions 1–4, P from sessions 5–8.

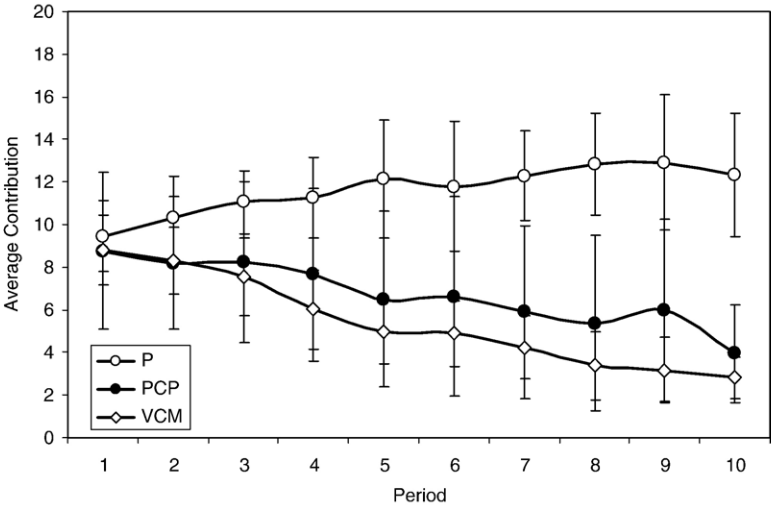

Fig. 2. Evolution of average contribution under random matching with 95% confidence intervals. Note: VCM is drawn from sessions 9–16, PCP from sessions 9–12, P from sessions 12–16.

experimental findings, contributions start at approximately 50% of the subjects' endowments under both matching protocols. From that point on, contributions in treatment P increase under both protocols, while they decline over time in VCM and PCP. Figs. 1 and 2, therefore, support the hypothesis that the introduction of counter-punishment opportunities would mitigate the disciplinary effect of one-sided punishment.[7] However, these raw results do not account for possible sequencing effects nor do they allow for comparison of the matching protocols. To control for these effects I turn to a thorough statistical analysis.

Table 1 suggests that a number of participants do not contribute to the public account (particularly in VCM). Therefore, using a Tobit model or OLS is inappropriate as these models constrain an individual's decision to contribute or not (*contribution decision*) and the decision of how much to contribute (*contribution level*) to have the same parameters (Johnston and DiNardo, 1997; p.440). The appropriate specification to capture the two-stage process is a hurdle model. The hurdle model is a parametric generalization of the Tobit model in which the decision to contribute to the public account and the level of contribution are determined by two separate stochastic processes. The hurdle is crossed if an individual decides to contribute. The likelihood function for the hurdle model is given by the product of two separate likelihoods. First, the likelihood that a subject will contribute a positive amount to the public account, which is captured by a standard Probit model, and second, the conditional likelihood of an individual contributing a certain number of ECU, which is estimated by using a truncated linear regression. The two parts of the hurdle model are estimated separately (McDowell, 2003).[8]

[7] The reader might have observed an "outlier" in Table 1. Session 10 has high contribution levels in both treatments relative to other sessions with the same matching protocol. At least a partial explanation for this is that on the day of the experiment some subjects arrived in a bad state at the lab due to a storm. The help offered by the experimenter provoked one of the subjects to say aloud "How nice? I feel in such a cooperative mood". This remark was met from the other subjects by laughter and further comments of the same nature.

[8] For another example of using a hurdle model to analyze contributions in a public good game see Botelho et al. (2005b).

Table 3 presents the maximum likelihood estimates of the hurdle model. The independent variables I use are: *PCP*, a dummy variable taking the value of one if the observation comes from PCP and zero otherwise; *P*, a dummy variable for treatment P; Fixed, a dummy variable that takes the value of one if the matching protocol is fixed and zero if the matching protocol is random; Before, a dummy variable which takes the value of one if the treatment with punishment (either P or PCP) preceded the VCM treatment; and *Period*, a variable to control for time effects. In order to test whether the trends observed in Figs. 1 and 2 are significant the second regression introduces interaction terms between *Period, PCP and P.*

Table 3 indicates that individuals in treatments PCP and P are 16.6 and 31.3% more likely to contribute a positive amount to the public account than individuals in VCM, respectively. Most importantly, in line with Hypothesis 2, individuals in PCP are significantly less likely to contribute a positive amount than individuals in P (p-value$<.01$). Individuals who decide to contribute to the public account contribute more in PCP and P than in VCM. The difference between PCP and P is also significant (p-value$<.01$). Neither the matching protocol nor the sequence of the treatments appears to affect contribution behavior.

The results from the second regression show that, at the beginning of the experiment, individuals in PCP and P are approximately 15% more likely to contribute to the public account than individuals in VCM. The difference between PCP and P is not significant (p-value$=.674$). In addition, the likelihood of contributing drops by 3.3% over time in VCM (as captured by *Period*). The insignificance of Period * PCP reveals that the contribution likelihood falls at similar rates in PCP and VCM. Conversely, individuals are more likely to contribute in P as the experiment

Table 3
Maximum likelihood estimates of the hurdle model of contributions

	(1)		(2)	
	Contribution decision	Contribution level	Contribution decision	Contribution level
PCP	0.166***	2.605***	0.141***	0.772
	(0.017)	(0.298)	(0.030)	(0.509)
P	0.313***	4.399***	0.156***	−2.017***
	(0.028)	(0.266)	(0.034)	(0.480)
Fixed	−0.026	1.880	−0.027	1.798
	(0.060)	(1.156)	(0.060)	(1.181)
Before	0.066	−0.704	0.066	−0.589
	(0.049)	(1.013)	(0.049)	(1.034)
Period	−0.028***	−0.250***	−0.033***	−0.765***
	(0.003)	(0.035)	(0.003)	(0.052)
Period*PCP			0.005	0.394***
			(0.005)	(0.084)
Period*P			0.040***	1.211***
			(0.008)	(0.077)
Constant		9.252***		11.739***
		(1.136)		(1.175)
N	3840	2765	3840	2765
Wald χ^2	486.10***	378.46***	526.20***	662.08***
Log likelihood	−1799.263		−1785.216	

Contribution decision is estimated using a Probit specification with cluster random effects, 'Cluster' refers to groups (sessions) under fixed (random) matching, Entries are the average marginal effects of the independent variables (Bartus, 2005), Contribution level is estimated using a truncated linear regression with cluster random effect, Estimation was done in STATA 9.2, Standard errors are in parentheses, ***Indicates significance at 1% level.

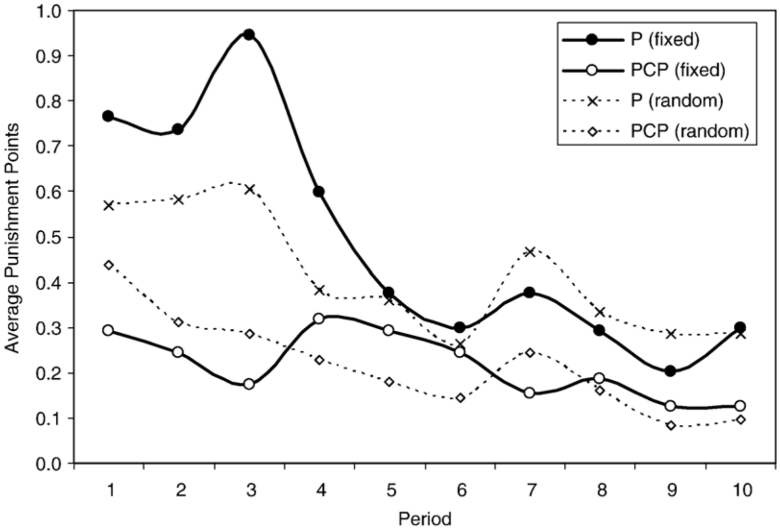

Fig. 3. Evolution of average punishment points assigned in P and PCP.

progresses. We also observe that those who decide to contribute to the public account contribute slightly more in PCP and P than in VCM over time. Result 1 summarizes.

Result 1. *Contribution levels are lower in the presence of counter-punishment opportunities. The highest contributions are in P followed by PCP and VCM. Contributions increase over time in P, while they decrease in PCP and in VCM.*

3.2. Punishment behavior

The next step is to analyze why the introduction of counter-punishment opportunities causes the unraveling of cooperation. I hypothesized earlier that the threat of counter-punishment might make people less willing to punish free riding. Fig. 3 illustrates the evolution of the average number of points assigned in the first punishment stage of P and PCP, and provides some initial support for Hypothesis 1. The average number of punishment points assigned in PCP is consistently lower than in P under both matching schemes. In particular there are 350 cases of punishment in PCP (153 under fixed and 197 under random matching) in contrast to 716 in P (357 under fixed and 359 under random matching); that is, punishment in treatment P is more than twice as frequent as in PCP.[9] The importance of this observation is amplified by the fact that contributions in PCP are significantly lower than in P giving participants additional reasons to punish.

To analyze formally whether counter-punishment influences negatively the willingness to punish free riders I employ a hurdle model. This approach is necessitated again by the large fraction of individuals that do not punish. In addition to the treatment variables, I include the following explanatory variables: $Own\_Neg\_Diff_{j,t} \equiv \max\{0, c_{i,t} - c_{j,t}\}$, and, $Group\_Neg\_Diff_{j,t} \equiv \max\{0, (\Sigma_{h \neq j}\, c_{h,t})/(n-1) - c_{j,t}\}$, where $c_{i,t}$ is the contribution of individual i in period t. The reason for including the first variable is straightforward: Individual i is likely to want to punish

[9] The maximum possible number of punishments is 2880 under each matching scheme (96 subjects, 10 periods, and 3 potential targets in each period).

Table 4
Maximum likelihood estimates of the hurdle model of punishments

	Punishment decision	Punishment level
PCP	−0.150***	−0.118
	(0.023)	(0.207)
Group_Neg_Diff$_{j,t}$	0.0001	0.053***
	(0.002)	(0.012)
Own_Neg_Diff$_{j,t}$	0.020***	0.034***
	(0.002)	(0.010)
Fixed	−0.008	0.379*
	(0.033)	(0.225)
Before	−0.004	−0.237
	(0.030)	(0.205)
Period	−0.015***	−0.018
	(0.002)	(0.011)
Constant		1.370***
		(0.247)
N	5760	1066
Wald χ^2	885.89***	151.10***
Log likelihood	−2108.44	

Punishment decision is estimated using a Probit specification with cluster random effects, 'Cluster' refers to groups (sessions) under fixed (random) matching, entries are the average marginal effects of the independent variables (Bartus, 2005), Punishment level is estimated using a truncated linear regression with cluster random effects, Estimation was done in STATA 9.2, Standard errors are in parentheses, ***Indicates significance at 1% level, *Indicates significance at 10% level.

individual *j* more if *j* contributes less than i to the public account. The variable *Group_Neg_Diff$_{j,t}$* is included to test the hypothesis that individuals are punished for deviating from group norms. The results from the regression are reported in Table 4.

Table 4 reveals that, in line with Hypothesis 1, individuals in PCP are 15% less likely to punish than they are in P. On the other hand, the threat of a counter-punishment does not significantly affect the punishment intensity. An additional unit of negative deviation from the contribution of one's peers does not increase significantly the likelihood of punishment, but increases the punishment intensity.[10] On the other hand, every unit of negative deviation of *j* from *i*'s contribution increases the likelihood of a punishment by 2% and also the punishment intensity. The punishment likelihood does not seem to be affected by the matching protocol or the sequence of the treatments. However, punishments are more severe if the group composition remains constant; this effect is significant at the 10-percent level. All else equal, the probability of punishment decreases in each period (by 1.5%), while the punishment intensity remains constant. Result 2 highlights the main finding from this section.

Result 2. *The willingness to punish decreases in the presence of a counter-punishment threat.*

Result 2 might be qualified by the fact that in almost all experiments utilizing one-sided punishment there is a percentage of *anti-social punishments*; that is, punishments of individuals

[10] The Probit coefficient of *Group_Neg_Diff$_{j,t}$* is significant at the 1-percent level. This implies that the likelihood of an individual being punished increases with his negative deviation from his peer's average contribution. However, the magnitude of this effect is small, and, therefore, an additional unit of negative deviation is not expected to increase the punishment likelihood significantly. The marginal effect of *Group_Neg_Diff$_{j,t}$* remains insignificant even if *Own_Neg_Diff$_{j,t}$* is not included in the regression.

who contribute more than the average of their peers. Anti-social punishments can be attributed to the anticipation by some free riders of the forthcoming punishment by cooperators and their willingness to retaliate these sanctions or the desire to avenge sanctions that took place in previous periods. It could be, therefore, that the lower demand for points in the first punishment stage of PCP is not due to a decreased willingness for punishing free riders, but simply due to anti-social punishments being relocated at the final stage as counter-punishments. However, this is not the case. In treatment P the percentage of anti-social punishments as a fraction of the total number of punishments is 21.5% (20 and 23% under fixed and random matching respectively). This is not dissimilar to the 18.1% for PCP (19 and 17.3% under fixed and random matching respectively).[11] Therefore, I conclude that the introduction of counter-punishment opportunities decreases individuals' willingness to punish free riders which in turn leads to the breakdown of cooperation.[12]

3.3. Counter-punishment behavior

Counter-punishments, similar to first-stage punishments, are acts of revenge. However, while first-stage punishments can be regarded as pro-social given that they are mostly aimed towards free riders, counter-punishments are mainly anti-social as they tend to target cooperators and discourage the enforcement of cooperation. Even though one can easily think of motivations behind punishments in the first stage, it seems harder to imagine why an individual would seek revenge for a punishment triggered by his decision to free ride. In this section, I investigate the driving forces of counter-punishments.

To understand the motivations of counter-punishment, one has first to examine who punished whom. In the majority of cases, the punisher contributed more than the average of her peers (72.1% — 73.2 and 71.1 under fixed and random matching respectively). Similarly, 79.8% of the punishments can be classified as pro-social punishments, that is, the target of the punishment is an individual who contributed less than the average of his peers (77.8 and 81.7 under fixed and random matching respectively). On the other hand, 18.1% of the punishments can be classified as anti-social (19 and 17.3 under fixed and random matching respectively), while 2.1% of the punishments were targeted to individuals contributing the same percentage of their endowment as their peers on average. In 80.1% of the cases (79.7 and 80.7 under fixed and random matching respectively) punishment is *justified* as the punisher contributed more than the punished individual, while in 19.9% of the cases (20.3 and 19.3 under fixed and random matching respectively) the punishment is *unjustified* as the punished individual contributed at least as much as the punisher. It seems, therefore, that punishments follow a logical pattern as they are mostly aimed towards those who contribute less than either the punisher or their peers on average.

[11] Cinyabuguma et al. (2006) conduct a public good experiment to investigate whether anti-social punishments can be mitigated by second-order punishment. Every three periods individuals are allowed to mete out punishments to those who punished cooperators or to those who punished free riders. Anti-social punishments, instead of disappearing, "migrate" to the second punishment stage. Unlike in this experiment, participants in Cinyabuguma et al. (2006) can only take revenge "blindly" as they are not informed about who punished them. Overall, the addition of a second punishment stage in every third period has no significant effect on either cooperation or individual earnings.

[12] The results in Fig. 3 can be interpreted, perhaps, in a different way. Some participants might wish to 'save' money to retaliate and, hence, punish less. Though this possibility cannot be excluded, I think that saving is not the driving force behind Result 2 for the following reasons: First, participants are probably aware that running out of money prior to stage three is not very likely as this requires a combination of receiving and giving punishment. Second, in a debriefing questionnaire, when asked about their punishment decisions, none of the participants reported 'saving' as a concern.

Nevertheless, there is a substantial amount of anti-social and unjustified punishments which might provoke retaliation.

Out of the 350 punishments that occurred in the first punishment stage of PCP, 25.7% triggered a counter-punishment (25.5 and 25.9% under fixed and random matching respectively). In general, victims of anti-social punishments appear to be more likely to counter-punish (47.4% — 44.8 and 50% under fixed and random matching respectively) than victims of pro-social punishments (21.5 — 21.8 and 21.1% under fixed and random matching respectively) as one might expect. However, only 33.3% of the total counter-punishment cases (same percentage under both protocols) are carried out by victims of anti-social punishments.

Similarly, victims of unjustified punishments retaliate in 46.4% of the cases (47.8 and 45.5 under fixed and random matching respectively), while victims of justified punishments retaliate 22.1% of the time (21.3 and 22.6 under fixed and random matching respectively). Therefore, it seems that victims of anti-social and unjustified punishments are more likely to counter-punish, although, a considerable percentage of counter-punishments is carried out by free riders and victims of justified punishments.

While these percentages reveal valuable information regarding counter-punishment behavior, they do not take into consideration factors which might be important. For example, an anti-social punishment towards the second highest contributor in a group might trigger a counter-punishment if the culprit is the lowest contributor in the group, but it might be accepted if the punishment is justified, in the sense that the culprit is the highest contributor of the group. To examine the determinants of counter-punishment I turn to a thorough statistical analysis.

Table 5 reports the maximum likelihood estimates of a hurdle model on counter-punishment behavior. In addition to the previous regressors I include p_{ij}, the punishment points that i assigned to j in the first punishment stage. I also include $Group\_Pos\_Diff_{j,t} \equiv \max\{0, (c_{j,t} - \Sigma_{h \neq j}\, c_{h,t})/(n-1)\}$, and $Own\_Pos\_Diff_{j,t} \equiv \max\{0,\ c_{j,t} - c_{i,t}\}$ to capture the reaction of individuals who are victims of anti-social and unjustified punishments, respectively.

The results indicate that each punishment point increases the likelihood of counter-punishment by 3.3%. The severity of j's counter-punishment also increases with the number of points i assigned to j, although, on average, subjects appear to be giving back fewer points than they receive. A punishment victim is more likely to counter-punish if the preceding punishment is unjustified. More specifically, every unit of positive difference between the contributions of victim and culprit increases the counter-punishment likelihood by 0.2%. In contrast, an additional unit of positive deviation from the average of one's peers increases significantly neither the counter-punishment likelihood nor the counter-punishment severity.[13] The likelihood of counter-punishment decreases by 0.2% in every period, while the severity of counter-punishment does not decrease significantly over time. Participants are equally likely to counter-punish under fixed and random matching although counter-punishments are more severe in the first case. Result 3 summarizes the key findings.

Result 3. *Both the likelihood and the intensity of counter-punishment increase in the severity of the preceding punishment. The likelihood of counter-punishment decreases slightly over time, while the greater the contribution of the punishment victim compared to the culprit the more likely counter-punishment is.*

[13] The Probit coefficient of $Group\_Pos\_Diff_{j,t}$ is significant at the 10-percent level (p-value = .086). The average marginal effect of $Group\_Pos\_Diff_{j,t}$ does not become significant if $Own\_Pos\_Diff_{j,t}$ is not included in the regression.

Table 5
Maximum likelihood estimates of the hurdle model of counter-punishments

	Counter-punishment decision	Counter-punishment level
p_{ij}	0.033***	0.264**
	(0.006)	(0.122)
Group_Pos_Diff$_{j,t}$	0.0001	0.023
	(0.001)	(0.068)
Own_Pos_Diff$_{j,t}$	0.002**	0.053
	(0.001)	(0.042)
Fixed	−0.013	0.678***
	(0.009)	(0.243)
Before	0.031*	−0.453*
	(0.016)	(0.269)
Period	−0.002**	0.025
	(0.001)	(0.042)
Constant		1.082***
		(0.400)
N	2880	90
Wald χ^2	189.32***	24.70***
Log likelihood	−278.24	

Counter-Punishment Decision is estimated using a Probit specification with cluster random effects, 'Cluster' refers to groups (sessions) under fixed (random) matching, Entries are the average marginal effects of the independent variables (Bartus, 2005), Counter-punishment Level is estimated using a truncated linear regression with cluster random effects, Estimation was done in STATA 9.2, Standard errors are in parentheses, *** Indicates significance at 1% level, ** Indicates significance at 5% level, * Indicates significance at 10% level.

What do these results imply for the motivations behind counter-punishments in this experiment? Subjects can use counter-punishments strategically to signal that future sanctions will not be tolerated. This will allow them to contribute less in subsequent periods. If this is the case we should observe more frequent and heavier counter-punishments under fixed matching. In addition, we should observe a progressive fall in the likelihood and the intensity of counter-punishment given that the expected benefits from counter-punishing decrease as we approach the end of the experiment. The results lend some support to the existence of strategic motivations behind counter-punishments: Counter-punishments are more severe under fixed matching, even though they are not more frequent. In addition, the likelihood of counter-punishment decreases slightly over time, even though the counter-punishment intensity does not.

Another explanation is that individuals counter-punish because they receive utility from retaliation, or, in Elster's (1990, p.862) words, "to impose suffering upon those who made [them] suffer". The results also lend support to this explanation. First, the greater the punishment severity the more likely punishment is to be followed by counter-punishment and the larger the severity of counter-punishment is predicted to be. Second, individuals counter-punish as frequently under random matching as they do under fixed matching, despite the absence (by and large) of strategic motivations. Those who do counter-punish assign the same number of points over time.[14]

[14] Individuals counter-punish even in the final period where strategic motivations are totally absent under both matching protocols. Six of the 22 (27.3%) punishments occurring in the last period of PCP are answered back.

Table 6
Earnings by treatment and matching protocol

	Average earnings	Average contribution	Average earnings after contribution	Punishment-associated costs[a]	Counter-punishment-associated costs[a]
	(1)	(2)	(3)	(4)	(5)
VCM fixed	23.10	5.16	23.10	–	–
P fixed	22.68	15.33	29.20	6.52	–
PCP fixed	21.93	8.72	25.23	2.70	0.60
VCM random	23.25	5.41	23.25	–	–
P random	21.81	11.63	26.98	5.16	–
PCP random	20.94	6.72	24.03	2.55	0.54

[a] "Associated Costs" include the amount of ECU paid by the victim of the (counter-) punishment *and* the culprit.

The results also indicate that counter-punishments are triggered by unjustified punishments: Individual j is more likely to counter-punish i if j has contributed more than i. One can also not exclude the possibility that counter-punishments are sometimes driven by *spiteful behavior*. Spiteful behavior has been observed in experiments (e.g. Fouraker and Siegel, 1963) and refers to actions taken by an agent that hurt others more than they hurt the agent. While at odds with standard economic assumptions, spiteful behavior can be justified by the fact that natural selection favors the fittest (Schaffer, 1989). As a result, individuals might be better off in the long run maximizing their relative instead of their absolute payoff. Unfortunately, this experiment is not suitable for isolating spiteful behavior from other motivations. For example, a counter-punishment of i by j could be driven by spiteful behavior if i had a higher income than j at the beginning of the final stage. This, however, in most cases would imply that j was the victim of an unjustified punishment. As a result, it is impossible to tell whether j was motivated by spiteful behavior or by the desire to take revenge for the unjustified punishment.

3.4. Welfare

So far we have seen that the threat of punishment increases contribution levels in all treatments (although the effect is significantly weaker if agents are allowed to counter-punish). From an organizational perspective, however, this is arguably not the most important issue. What is of interest is whether decentralized punishments can improve welfare.

One way to measure welfare is by adding up individual earnings. For decentralized punishments to increase earnings the benefits from higher contributions must outweigh the punishment and counter-punishment-associated costs. I begin by examining earnings in the absence of counter-punishment opportunities.

The experimental evidence is mixed on whether one-sided punishment increases individual earnings. Most studies find that one-sided punishment lowers earnings (Botelho et al., 2005a; Fehr and Gächter, 2000, 2002; Ostrom et al., 1992; Sefton et al., 2005).[15] Some studies find that the introduction of punishment opportunities does not affect earnings significantly (Bochet et al.,

[15] The welfare comparison for the experiments of Fehr and Gächter (2000) and Fehr and Gächter (2002) can be found in Cinyabuguma et al. (2004) and Botelho et al. (2005a), respectively.

Table 7
Determinants of earnings

	(1)	(2)	(3)
PCP	−1.141***	−4.104***	−0.118
	(0.269)	(0.505)	(0.586)
P	−1.518***	−9.168***	2.796***
	(0.269)	(0.505)	(0.586)
Fixed	0.390	0.390	−0.239
	(1.054)	(1.054)	(0.850)
Before	0.276	0.275	1.132
	(0.922)	(0.922)	(0.791)
Period	0.003	−0.479***	
	(0.034)	(0.046)	
Period*PCP		0.539***	
		(0.079)	
Period*P		1.390***	
		(0.079)	
Constant	22.822***	25.476***	20.860***
	(1.039)	(1.053)	(0.836)
N	3840	3840	384
Wald χ^2	49.11***	365.37***	25.65***

Dependent variable in (1) and (2) is the earnings of *i* at the end of a given period, In (3) the dependent variable is the earnings of individual *i* in the final period, The model is estimated using a linear regression with cluster random effects, ‘Cluster’ refers to groups (sessions) under fixed (random) matching, Estimation was done in STATA 9.2, Standard errors are in parentheses, *** Indicates significance at 1% level.

2006; Page et al., 2005; van Soest and Vyrastekova, 2007). Nikiforakis and Normann (in press), and Yamagishi (1986, 1988) find that, if the punishment threat is severe enough, then punishments can increase earnings, while Masclet et al. (2003) find that the introduction of punishment opportunities unambiguously increases earnings.

Table 6 provides an overview of how average earnings are shaped in each treatment of this experiment. Comparing columns (3) and (4) one sees that the higher contributions enforced by one-sided punishment in P increase average earnings by 6.10 (3.73) ECU under fixed (random) matching compared to the VCM. This increase, however, is not enough to cover the average punishment-associated costs that amount to 6.52 (5.16) ECU under fixed (random) matching. As a consequence groups in P have lower earnings on average than groups in VCM.

Given the fact that individuals tend to overuse punishments one wonders whether the introduction of counter-punishment opportunities can actually lead to an increase in earnings. The answer is no: Despite the fact that the average punishment-associated costs are reduced by more than half under both matching protocols (58.6 and 50.1% under fixed and random matching respectively), this decrease is not sufficient to compensate for the decline in contributions. As a result, groups in PCP have lower earnings than groups in VCM and P. Interestingly, the sum of punishment and counter-punishment-associated costs in PCP amounts to only 50.1 and 59.9% of the punishment-associated costs in P under fixed and random matching, respectively.

Table 7 examines the relation between treatment variables and individual earnings. The first regression shows that earnings are significantly lower in PCP and P than in VCM on average. The difference between *P* and *PCP* is not significant (p-value = .317). Neither the matching protocol nor the sequence of the treatments has a significant effect on earnings.

The second regression examines the evolution of earnings across treatments. Earnings are lower in P and PCP at the beginning of the experiment compared to VCM. This is to be expected as contribution levels are similar across treatments, but individuals in P and PCP have to pay the punishment-associated costs. Earnings are decreasing over time in VCM as indicated by *Period*. The decline is due to the decay in contribution levels. Variable *Period* * *PCP* captures the time trend in PCP with respect to that in VCM. Due to the higher contribution levels in PCP and the drop in the number of punishments over time, earnings are steadily improving in PCP compared to the VCM. However, it should be noted that there is no significant trend in PCP; that is, earnings are neither increasing nor decreasing over time in PCP, while they are significantly increasing in P.[16] The latter is the result of the decline in the number of punishments and the increase in contribution levels.

Given the different trends, it is interesting to compare earnings in the final period of the experiment. At the beginning of the experiment, earnings in VCM are higher than in PCP. However, earnings are falling in VCM over time, while they remain constant in PCP. In the third column of Table 7 one can see that, as a result of these trends, earnings in the final period are not significantly different in PCP and VCM. On the other hand, P has significantly higher earnings in the final period than both PCP and VCM. Note, however, that earnings in P are still far from the socially optimal payoff of 32 ECU per person which would be acquired if all group members contributed the whole of their endowment to the public account and refrained from any type of punishment. Result 4 summarizes.

Result 4. *The highest earnings are found in VCM followed by P and PCP. Earnings are increasing over time in P, are constant in PCP, and are decreasing in VCM.*

Earnings are only an approximate measure of welfare as they do not take into account other factors which might affect an individual's utility such as the satisfaction of punishing a free rider or the disutility from being counter-punished. A natural question to ask, therefore, is which institution would individuals choose if they were given the option: VCM, P or PCP?

To answer this question some assumptions are necessary. If individuals decide to punish/counter-punish at a cost (and they do so even in the final period of the experiment), it must be that they enjoy a non-material benefit from their action (see also de Quervain et al., 2004). The fact that cooperators punish much more frequently than free riders in the first punishment stage and are also more likely to retaliate if they are punished implies that, not only do cooperators receive more utility by punishing/counter-punishing than free riders, but also that if they are punished/counter-punished they suffer a greater loss in utility than free riders. One explanation why this might happen is that individuals punish/counter-punish to protect their personal identity or self-esteem. Indeed, Benabou and Tirole (2006; p.4) write that "challenges to a strongly held [identity] ... elicit forceful *counterreactions* aimed at restoring the threatened beliefs" (emphasis in original). This view is in line with the conjecture in Fehr and Gächter (2000; p. 984) that "subjects strongly dislike being the 'sucker', that is, being those who cooperate while other group members free ride. This aversion against being the 'sucker' might well trigger a willingness to punish free-riders". The free rider, on the other hand, knowing that he outwitted the cooperator, might not suffer as much from the punishment (punishment might even be taken to be the acknowledgement that the free rider outwitted the cooperator). Based on these observations, it seems reasonable to assume that (a) the

[16] If I remove *Period* from the second regression, variables *Period* * *PCP* and *Period* * *P* capture the trends in PCP and P, respectively. The former is not significant (p-value = .356) while the latter is (p-value < .01).

utility a cooperator receives by punishing/counter-punishing a free rider exceeds the loss in utility the free rider experiences; (b) a cooperator who is punished/counter-punished suffers a loss in utility which exceeds the utility received by the free rider; and (c) individuals who have the opportunity to punish but decide not to, fearing a counter-punishment, are likely to suffer a loss in utility as their self-esteem will be lowered. This is not the case when individuals cannot punish free riders as in VCM. Given these assumptions I proceed to compare the three institutions.

Consider first the choice between VCM and PCP. There are 350 punishments in PCP. From assumption (a), these punishments have a net positive effect on welfare. However, from assumption (c), this effect is likely to be offset by the equally large number of punishments that are repressed under the threat of counter-punishment (remember that there are 716 punishments in total in P and only 350 in PCP despite the lower contribution levels). Further, while some individuals in PCP might enjoy the fact that they can counter-punish, from assumption (b), this effect is likely to be overcompensated by the loss in utility suffered by the victims of counter-punishment (remember that 66.7% of all counter-punishments are carried out by free riders). Given the fact that earnings are on average higher in VCM than in PCP, it seems plausible that more individuals would select VCM than PCP.

Consider next the choice between P and PCP. Earnings in P are not higher than in PCP due to the excessive investments in punishment. Individuals in both treatments can punish free riders, but more do so in P. Therefore, from assumption (a), the utility from punishing will be higher in P. From assumption (c), the utility enjoyed by the individuals who punish in PCP might be offset by the disutility of those who abstain from punishing fearing a counter-punishment. Similarly, from assumption (b), the utility of the counter-punishers will be counter-balanced by the loss in utility of their victims. Given the greater number of punishments in P, from assumption (a), I conclude that more individuals would choose P over PCP. In other words, it seems that, given the option, most individuals would choose VCM and P over PCP; that is, PCP would be the least preferred institution.

The comparison between VCM and P is less straightforward: Individuals in VCM might have higher earnings on average than individuals in P, but they cannot take revenge for anti-social behavior as individuals in P. Four experimental studies have tried to answer which effect is dominant. The studies have a number of methodological differences and obtain contrasting results: Ostrom et al. (1992) and Gürerk et al. (2006) report that the majority of the participants prefers institutions like P over institutions like VCM. On the other hand, Botelho et al. (2005a) and Sutter et al. (2005) find the opposite result.

4. Discussion and conclusion

Oliver (1980, p.1373) writes that punishment is “essential for ensuring unanimous cooperation in costly collective action, but has the potential side effects of disharmony and discord”. The efficacy of punishment in sustaining cooperation has been repeatedly tested in the laboratory and has received wide support. However, Oliver’s second point has been largely overlooked in experimental studies although scholars including Axelrod (1984) and Schelling (1960) have warned about agents’ inclination to avenge sanctions.

In this paper, I examine how cooperation and group welfare are affected when agents are given the ability to take revenge for punishments. Using a public good game as a testing ground, I find that one quarter of all punishments are retaliated. Individuals who counter-punish seem to be motivated by a desire to hurt those who hurt them, but also use counter-punishments strategically to discourage future punishments. The threat of revenge weakens cooperators’ willingness to punish free riders and leads to the breakdown of cooperation. The benefits from higher cooperation are insufficient to offset the punishment costs and, as a result, earnings in both the treatment with punishment (P) and the treatment with punishment *and* counter-punishment opportunities (PCP) are lower than in the treatment where punishments are not allowed (VCM) and free riding is predominant. Earnings are not significantly different between treatments PCP and P. Nevertheless, earnings are increasing in P over time, while they are decreasing in VCM and remain stable at a low level in PCP.

What conclusions can we draw from these results? In a frequently cited passage, Olson (1965, p.2) writes that "unless the number of individuals is quite small, or unless there is a coercion or some other special device to make individuals act in their common interest, *rational*, *self-interested individuals will not act to achieve their common or group interest*" (emphasis in original). In the present study, the group size is small and a coercive device exists. Furthermore, some individuals appear to deviate from standard assumptions of rationality/selfishness by contributing and punishing free riders even in the final period. However, even under these favorable conditions, cooperation cannot be sustained in the presence of counter-punishment opportunities. As counter-punishment opportunities exist almost in every decentralized interaction where punishment opportunities exist, the results question the belief that individuals can govern themselves through punishments, and lend support to the widespread existence of central authorities. As Thomas Hobbes famously wrote in 1651 "without a *common* power to keep them all in awe [men] are in that condition which is called war" (Hobbes, 1994; p.76 — emphasis added).

The results are drawn from a specific environment and generalizations should be made with care. In a naturally occurring environment a number of factors might help foster cooperation even in the presence of counter-punishment opportunities. Communication (Ostrom et al., 1992), rewards (Sefton et al., 2005), and indirect reciprocity (Milinski et al., 2002) are factors which can positively influence cooperation among individuals. What the results from the present experiment indicate is that punishment *by itself* improves neither cooperation nor individual earnings in the presence of counter-punishment opportunities.

Of course, as mentioned earlier in the paper, certain restrictions have been imposed on the types of punishments that can be carried out. These restrictions might qualify the conclusion above. One can argue, for example, that, by not providing information about the punishment activities of other group members, I have deprived participants of the opportunity to establish other pro-social norms like the punishment of non-punishers. Such norms might help sustain decentralized cooperation at high levels even in the presence of counter-punishment opportunities. Denant-Boemont et al. (in press) show that this does not appear to be the case.

Denant-Boemont et al. (in press) is an interesting complementary study based on a similar experimental design. By manipulating the information available to participants at the second punishment stage the authors can compare the effect of *sanction enforcement* (that is, the punishment of those who failed to punish free riders) to that of counter-punishment: If the former effect is larger, then cooperation might be sustained even in the presence of counter-punishment opportunities. In one treatment (*Revenge*) participants receive information only about who punished them — as do participants in PCP. In another treatment (*Full Information*) individuals receive information about the punishment activity in the whole group. In other words, individuals in Full Information can also punish those who failed to punish free riders. Denant-Boemont et al. find that sanction enforcement increases contributions slightly above the level observed in Revenge, although this effect is not significant. The level of cooperation in both treatments is significantly lower than in the treatment with one-sided punishment. These results indicate that the effect of counter-punishment is indeed strong and that the restriction of information at the second-punishment stage does not affect significantly the results in this paper.

Summarizing, this paper shows that individuals retaliate punishments. As a result, the introduction of counter-punishment opportunities makes individuals less willing to punish free riders. The decrease in the number of punishments leads to the breakdown of cooperation and lowers group earnings in comparison to a treatment where punishments are restricted and low levels of cooperation are observed.

Acknowledgments

I am indebted to Hans Normann, Dirk Engelmann and Simon Gächter for their continuous support and comments. I wish to thank two anonymous referees for their helpful and detailed comments, as well as Dan Anderberg, Andres Carvajal, Marco Casari, Mick Coelli, John Creedy, Ernst Fehr, Herbert Gintis, Liana Jacobi, Johan Lagerlof, Manisha Shah, Martin Sefton, Anne van den Nouweland, Jonathan Wadsworth and Brian Wallace all of whom gracefully shared some of their wisdom and limited time and helped improve this paper. The paper also benefited from comments at the International ESA Meeting in Amsterdam (June 2004), the North American ESA Meeting in Tucson (November 2004) and the International Meeting on Experimental Economics in Cordoba (December 2004). I gratefully acknowledge support from Leverhulme (project F/07537/S) and ESRC (project RES-000-22-0948).

References

Anderson, C., Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games and Economic Behavior 54 (1), 1–24.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, Inc., New York.

Bartus, T., 2005. Estimation of marginal effects using margeff. The Stata Journal 5, 309–329.

Benabou, R., Tirole, J., 2006. Identity, dignity and taboos: beliefs as assets. IDEI Working Paper, vol. 437.

Bochet, O., Page, T., Putterman, L., 2006. Communication and punishment in voluntary contribution experiments. Journal of Economic Behavior and Organization 60 (1), 11–26.

Botelho, A., Harrison, G.W., Pinto, L.M.C., Rutstrom, E.E., 2005a. Social Norms and Social Choice, Working Paper 05-23, Department of Economics, College of Business Administration, University of Central Florida.

Botelho, A., Harrison, G.W., Pinto, L.M.C., Rutstrom, E.E., 2005b. Testing Static Game Theory with Dynamic Experiments: A Case Study of Public Goods, Working Paper 05-25, Department of Economics, College of Business Administration, University of Central Florida.

Carpenter, J., 2007. The demand for punishment. Journal of Economic Behavior and Organization 62 (4), 522–542.

Carpenter, J., in press. Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. Games and Economic Behavior.

Cinyabuguma, M., Page, T., Putterman, L., 2004. On Perverse and Second-Order Punishment in Public Goods Experiments with Decentralized Sanctions, Working Papers 2004-12, Brown University, Department of Economics.

Cinyabuguma, M., Page, T., Putterman, L., 2006. Can second-order punishment deter perverse punishment? Experimental Economics (9), 265–279.

de Quervain, D.J.F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. Science 305 (5688), 1254–1258.

Denant-Boemont, L., Masclet, D., Noussair, C., in press. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Economic Theory.

Elster, J., 1990. Norms of revenge. Ethics 100 (4), 862–885.

Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90, 980–994.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.

Fehr, E., Schmidt, K., 1999. A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114, 817–868.

Fehr, E., Schmidt, K., 2003. Theories of fairness and reciprocity: evidence and economic applications. In: Dewatripont, M., et al. (Ed.), Advances in Economic Theory. Eighth World Congress of the Econometric Society, vol. I. Cambridge University Press, Cambridge, pp. 208–257.

Fischbacher, U., 2007. *z*-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10 (2), 171–178.

Fouraker, L.E., Siegel, S., 1963. Bargaining Behavior. McGraw-Hill, New York.

Gürerk, O., Irlendbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. Science 312, 108–111.

Hobbes, T., 1994. Leviathan. Hackett Publishing Company Inc., Indianapolis.

Johnston, J., DiNardo, J., 1997. Econometric Methods, Fourth edition. McGraw-Hill.

Kreps, D., Wilson, R., 1982. Reputation and imperfect information. Journal of Economic Theory 27, 253–279.

Kreps, D., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated prisoners' dilemma. Journal of Economic Theory 27, 245–252.

Ledyard, J., 1995. Public goods: a survey of experimental research. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press.

Masclet, D., Noussair, C., Tucker, S., Villeval, M.C., 2003. Monetary and non-monetary punishment in the voluntary contributions mechanism. American Economic Review 93, 366–380.

McDowell, A., 2003. From the help desk: hurdle models. The Stata Journal 3 (2), 178–184.

Milgrom, P., Roberts, J., 1982. Predation, reputation and entry deterrence. Journal of Economic Theory 27, 280–312.

Milinski, M., Semmann, D., Krambeck, H.J., 2002. Reputation helps solve the tragedy of the commons. Nature 415, 424–426.

Nikiforakis, N., Normann, H.T., in press. A comparative statics analysis of punishment in public-good experiments. Experimental Economics.

Nikiforakis, N., Normann, H.T., Wallace, B., 2007. Asymmetric Enforcement of Cooperation in a Social Dilemma. University of Melbourne Economics Working Paper, vol. 982.

Noussair, C., Tucker, S., 2005. Combining monetary and social sanctions to promote cooperation. Economic Inquiry 43 (3), 649–660.

Oliver, P., 1980. Rewards and punishments as incentives for collective actions: theoretical investigations. The American Journal of Sociology 85 (6), 1356–1375.

Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge, Massachusetts.

Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: self governance is possible. American Political Science Review 86, 404–417.

Page, T., Putterman, L., Unel, B., 2005. Voluntary association in public goods experiments: reciprocity, mimicry and efficiency. The Economic Journal 115 (506), 1032–1053.

Rabe-Hesketh, S., Skrondal, A., 2005. Multilevel and Longitudinal Modeling Using Stata. Stata Press, College Station, TX.

Samuelson, P., 1954. The pure theory of public expenditure. The Review of Economics and Statistics 36 (4), 387–389.

Samuelson, C.D., Messick, D.M., Wilke, H.A.M, Rutte, C.G., 1986. Individual restraint and structural change as solutions to social dilemmas. In: Wilke, H.A.M, Messick, D.M., Rutte, C.G. (Eds.), Experimental Social Dilemmas. Lang, Frankfurt.

Sato, K., 1987. Distribution of the cost of maintaining common resources. Journal of Experimental Social Psychology 23, 19–31.

Schaffer, M.E., 1989. Are profit maximisers the best survivors? A Darwinian model of economic natural selection. Journal of Economic Behavior and Organization 12, 29–45.

Schelling, T.C., 1960. The Strategy of Conflict. Harvard University Press, Cambridge, Massachusetts.

Sefton, M., Shupp, R., Walker, J., 2005. The Effect of Rewards and Sanctions in Provision of Public Goods. Ball State University, Department of Economics, Working Paper, vol. 200504.

Sutter, M., Haigner, S., Kocher, M.G., 2005. Choosing the Stick or the Carrot? Endogenous Institutional Choice in Social Dilemma Situations, Centre for Economic Policy Research, DP 5497.

van Soest, D., Vyrastekova, J., 2007. Peer enforcement in CPR experiments: the relative effectiveness of sanctions and transfer rewards, and the role of behavioral types. In: List, J. (Ed.), Using Experimental Methods In Environmental And Resource Economics. Edward Elgar.

Yamagishi, T., 1986. The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology 51 (1), 110–116.

Yamagishi, T., 1988. Seriousness of social dilemmas and the provision of a sanctioning system. Social Psychology Quarterly 51 (1), 32–42.