Discussion Paper Series

2004 – 05

Department of Economics
Royal Holloway College
University of London
Egham TW20 0EX

Punishment and Counter-punishment in Public Goods Games: Can we still govern ourselves? *

Nikos S. Nikiforakis

Royal Holloway University of London

Abstract

Recent public goods experiments have shown that free riding can be curtailed through mutual monitoring and sanctioning between members of a group. However, often we can not allow for punishment and exclude the possibility of counter-punishment occurring. We design a public goods experiment, where we allow for both punishment and counter-punishment. We find that in both partner and stranger treatments cooperation declines over time. The reason is that people are less willing to punish under the threat of counter-punishment. Participants squander their endowment in costly confrontations leading to a relative payoff loss, in comparison to a treatment without punishments.

JEL Classification: C91, C92, H41, D64

Keywords: public goods, punishment, counter-punishment, mutual monitoring, free-riding, experiments

---

* The title is a reference to Ostrom, Walker and Gardner (1992): "Covenants with and without a sword: Self governance is possible". I would like to express my gratitude to Hans Normann for his invaluable help at every stage of this paper and Brian Wallace. I would also like to thank Simon Gaechter, Ernst Fehr, Dirk Engelmann, Herbert Gintis and Marco Casari for the helpful comments. The author can be contacted at: n.nikiforakis@rhul.ac.uk

1. Introduction

Contrary to the predictions of standard economic theory that people will not contribute voluntarily for the production of a public good, a considerable amount of experiments have shown that, initially, people give on average between 40 and 60 percent of their endowment. However, the contribution level decreases with repetition under the influence of free-riders [Davis and Holt (1993), Ledyard (1995), Fehr and Fischbacher (2003)]. This leads to the under provision of public goods.

In an attempt to deal with this problem, Fehr and Gaechter (2000) (hereafter F&G) designed an experiment, where participants played a two-stage public goods game. In the first stage, they were asked to divide their endowment between a public and a private account. The returns from each account were designed so that group earnings were maximized when participants contributed all their money in the public account. However, each individual had an incentive to keep his endowment for himself.

In the second stage, participants were allowed to assign punishment points to the other members in their group after they were notified about individual contributions in the public account. Punishment was costly for both the punisher and its receiver. The ability to punish non-cooperators led to significantly higher contributions in comparison to a treatment, where sanction opportunities did not exist.

The design of F&G has become a standard and a considerable number of more recent studies have used it since to address further issues in public goods literature [Bowles and Gintis (2002), Page and Putterman (2000), Sefton, Shupp and Walker (2002), Carpenter (2002), Masclet, Noussair, Tucker, Villeval (2003)].

In every day life, however, there exists an abundance of evidence that people are willing to engage in costly counter-punishment, such as the infamous vendettas in Italy and the recently revived honour-related blood feuds in Albania. At the same time, one can observe many cases of free-riding coupled with the cooperators' unwillingness to punish. An example of

the latter is the reluctance to punish countries that refused to sign the Kyoto Protocol for the reduction of the emissions of greenhouse gasses in the atmosphere[1].

In the light of this, we wish to test whether the threat of counter-punishment can be an explanation in cases where free-riding is observed. To do this we designed a public goods experiment with two treatments: one without any form of punishment, the familiar voluntary contribution mechanism (VCM), and one with punishment and counter-punishment (P&CP). The two treatments were run under both the partner and the stranger protocol.

In the VCM treatment, as we will see, average contribution exhibited a similar behaviour to the one reported in other experiments, by starting between 40 and 60 percent of the endowment and decreasing over time. The introduction of counter-punishment opportunities in the P&CP treatment seems to cancel out, to a large extent, punishment's reported disciplinary effect and participants behave similarly to the VCM treatment with average contribution declining with repetition. In the words of Girard (1979): "Reciprocal violence now demolishes everything that unanimous violence has erected". We show that an explanation for this is that under the counter-threat, people are less willing to punish and as a result, participants are almost free to free ride.

To our knowledge, there is no other paper testing for the effect that the existence of counter-punishment opportunities has on the level of cooperation. Although in our experiment no explicit coordination opportunities exist, in the partner treatment, the fact that the composition of

[1] Air is a textbook case of a pure public good.

the groups remains the same might lead to the formation of behavioural norms that will alleviate free-riding more effectively[2].

The remaining of the paper is structured as follows: section 2 introduces the experimental design and the procedures of the experiment, while section 3 presents the theoretical predictions for our model. Section 4 discusses the experimental results and section 5 concludes.

2. The Experiment

2.1 *The experimental design*:

To have a clear picture of the effect that counter-punishment has on cooperation we based our design on F&G. In general, we will refer to the type of punishment, like the one found in F&G, as "*one-sided punishment*", in contrast to the "*two-sided punishment*" when counter-punishment is allowed.

Using a related sample design, the experiment consists of two treatments: one without any punishment (VCM), and one with two-sided punishment i.e. with punishment and counter-punishment (P&CP). We run the treatments both under the partner protocol, where the composition of each group remains unchanged throughout the experiment and under the stranger protocol, where the participants where randomly re-matched in each period. For each treatment there were 12 subjects who were randomly divided in groups of 4 people and played a finitely repeated public goods game for 10 periods.

All participants were aware that each treatment would last exactly 10 periods. However, they were not aware that a second treatment was to

[2] Masclet et al. (2003) show that when the same group of people play a finitely repeated public goods game the expression of disapproval towards anti-social behavior can also play a significant role in decreasing free-riding.

follow[3]. The related sample design, i.e. each subject participates in both treatments, has the advantage that additionally to across-subjects comparison we can make within-subjects comparisons of the average level of contribution, which have much more statistical power. To test for sequence effects, in session 1 (stranger) and session 4 (partner) the participants played the P&CP treatment first and the VCM second, whereas in sessions 2, 3 (stranger) and 5 (partner) the order was reversed. All this are summarised in table 1. In addition, we run two control sessions, one under each protocol, using one-sided punishment.

Table 1: Treatment Conditions

	P&CP / VCM	**VCM / P&CP**
Stranger	Session 1: 3 groups of 4 participants	Session 2 & 3: 3 groups of 4 participants
Partner	Session 4: 3 groups of 4 participants	Session 5: 3 groups of 4 participants

2.1.1 *The VCM treatment*:

The first treatment is the standard voluntary contribution mechanism as presented first by Isaac, Walker and Thomas (1984) and served as a control for the P&CP treatment. In the beginning of each of the ten periods, every participant received a fixed amount of 20 Experimental Currency Units (ECUs)[4]. The participant had then to decide how many ECUs to keep for himself and how many to invest into a project. All the participants made their decision simultaneously and without being aware of the others' decisions. The monetary payoff for each subject in each period was given by:

[3] This was done following the example of F&G, to keep the results from the first treatment unaffected by the existence of a second treatment.

[4] The ECU was exchanged at a rate of: 1 ECU = 4 p.

(1) $$\pi_i^{VCM} = 20 - g_i + 0.4 * \sum_{j=1}^{n} g_{j,}$$

where 20 is the endowment in ECUs, g_i is the amount of ECUs subject *i* invests in the project ($0 \leq g_i \leq 20$) and 0.4 is the marginal return per capita (MRPC) from the project. The payoff function implies that each player's income comes from two sources: the money he keeps for himself, as indicated by $20 - g_i$ and a fraction of the total amount that the group invested in the project, $0.4 * \sum_{j=1}^{n} g_j$.

Equation (1) also suggests that full free-riding (g_i=0) is a dominant strategy in the stage game. This follows from $\partial \pi_i^{VCM} / \partial g_i$=-1+0.4<0, which means that the more an individual contributes to the project the less her income will be in that stage. However, the aggregate payoff, $\sum_{i=1}^{4} \pi_i^{VCM}$ is maximized if each group member fully cooperates (g_i=y), since $\partial \sum_{i=1}^{4} \pi_i^{VCM} / \partial g_i$=-1+4*0.4>0.

In the first treatment, the payoff function (1), the amount of the endowment, the MRPC, the number of the subjects and the duration of the treatment were all common knowledge between the players. The total payoff from the VCM treatment is equal to the sum of the 10 period payoffs as given by (1) i.e. $\sum_{n=1}^{10} \pi_i^{VCM}$.

2.1.2 *The P&CP treatment*:

In the second treatment, two more stages were added to the simple voluntary contribution mechanism, which now became the first of three stages. In the second stage subjects were given the opportunity to simultaneously punish

each other after being informed of the individual contributions[5]. To do so, group member *i* had to assign *punishment points* to group member *j*. This had two different effects in the payoffs of members *i* and *j*: for each point received by player *j* his income from the first stage, π_i^1, was reduced by 10%. Note that the first stage income could never be reduced below zero, so if player *j* received more than 10 punishment points her income was reduced by 100%. Additionally, player *i* also faced a cost for distributing punishment points to player *j*. This cost was given by the following convex cost function, $\underset{i\neq j}{c}(p_i^j)$:

Table 2: Punishment points per player and associated costs for the punishing subject

p_i^j	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$\underset{i\neq j}{c}(p_i^j)$	0	1	2	4	6	7	12	16	20	25	30

Given the above information, the payoff at the end of the second stage for subject ***i*** is equal to:

(2)
$$\pi_i^2 = \pi_i^1 * \left[\frac{\max(0, 10 - \sum_{j\neq i} p_j^i)}{10} \right] - \sum_{i\neq j} c(p_i^j)$$

Up to the end of the second stage, the experiment is identical to the one by F&G. In the third and final stage, the subjects were informed about how many points each of the other members in their groups assigned to them. They were then given a last opportunity to reduce the income of the participants who punished them during the second stage by buying counter-points[6]. To avoid strategic punishing, which would be inappropriate to study

[5] For the whole experiment we used neutral framing. Punishment was referred to as "assigning points" in order to "reduce" another participant's income. The public good itself was named "project".

[6] The cost of the counter-points was equal to the cost of points, i.e. $c(cp_i^j) = c(p_i^j)$.

the effect of counter-punishment, only the subjects who were punished were allowed to punish back[7].

The cost for assigning points works accumulatively i.e. if player *i* punished player *j* with 2 points during the second stage and then with 2 further (counter-) points in the third stage, his total cost from points would be equal to 6 i.e. the cost of 4 points. The end-of-period income is given by the following equation:

$$\textbf{(3)} \qquad \pi_t^3 = \pi_t^2 * \left[\frac{\max(0, 10 - \sum_{j \neq i} cp_j^i)}{10} \right] - \sum_{i \neq j} c(p_j^i + cp_i^j) + \sum_{i \neq j} c(p_j^i)$$

where cp_i^j is the number of counter-points that player *i* assigns to player ***j*.**

The payoff functions (2) and (3), the cost function ($\underset{i \neq j}{c}(p_i^j)$), the amount of the endowment, the MRPC, the number of the subjects and the duration of the treatment were all common knowledge.

To prevent the possibility of forming an individual reputation, in the beginning of each period, every player received a number between 1 and 4 to distinguish their actions from the others' within a period. This number, however, changed from period to period.

Due to the restriction we impose on punishment, our design is expected to be a better predictor in cases where the punishment of non-punishers is not possible, relevant or significant. Amongst others, such cases can include social exclusions, where individuals cannot observe each other's sanctions, blood feuds and one-off interactions.

2.2. *Procedures*

The experiment took part between December 2003 and March 2004 in the experimental laboratory of Royal Holloway, University of London. It

[7] By strategic punishment here we mean the preference of a subject to punish in the last stage, to avoid counter-punishment, instead of the second stage.

consisted of five sessions (2 partner, 3 stranger and 2 control[8]), which lasted approximately an hour and forty-five minutes[9]. The participants were recruited via e-mail. The total number of subjects was 84. The sample consisted of students with different nationalities and backgrounds including Economics[10].

At the beginning of each of the treatments, the participants were given a different set of instructions explaining in detail what was to happen[11]. They were then given as much time as they needed to read the instructions and to fill in a brief control questionnaire. In addition they were read a summary from a pre-written text. A trial period was used, where the participants were introduced to the computer screens they would have to use to make their decisions. Again, for this purpose, a pre-written text was used, to ascertain, as before, that all subjects would receive the same explanations regardless of the session they participated. The experiment was programmed and conducted with the software z-Tree (Fischbacher [1999]). Participants earned on average £17.90. No show up fee was given.

3. Theoretical Predictions

The subgame-perfect Nash equilibrium prediction in all treatments is that participants should contribute nothing to the project, i.e. g_i=0, for every i. In specific, in the VCM treatment the dominant strategy is to free ride. Using backward induction for the ten periods we find that the dominant strategy is to contribute nothing in the project.

[8] Controls were used to test for differences in behaviour across countries based on cultural characteristics (Burlando and Hey [1997]).

[9] The control treatment lasted slightly less.

[10] Contrary to other findings (Marwell and Ames [1981]) the economists-to-be were arguably the strongest supporters of cooperation.

[11] The instructions for stages one and two were adopted from F&G. Instructions are available from the author upon request.

In the P&CP treatment, the subgame-perfect Nash equilibrium prediction is that people will neither punish nor counter-punish since this is costly and yields no material benefits. At the first stage, the participants understand that no one is going to punish them no matter whether they cooperate or not, and therefore they have no reason to contribute to the project, thus choosing to contribute zero. Applying backward induction for the ten periods we arrive at the prediction that g_i=0, p_i^j=0 and cp_i^j=0.

As we saw earlier, experimental findings contradict these predictions. People are willing to contribute substantial fractions of their endowments in public accounts and to engage in costly confrontations. This is the first paper looking at people's behaviour in the presence of counter-punishment.

4. Experimental Results

We will begin by analyzing the effect of counter-punishment under the stranger protocol and then continue with the partner protocol. In the stranger condition, a 25% of the 136 sanctions were answered back[12]. Out of them, 55.3% were answered back with as many counter-points as the punishment points received. A 13.1% of them were answered back with even more points than those received.

4.1 *The impact of counter-punishment under the stranger protocol*

If the introduction of counter-punishment is of no importance, then we should observe no difference in the behaviour of the participants in comparison to other experiments that studied one-sided punishment. This means that in the P&CP treatment average contribution should increase in comparison to the VCM treatment and continue to do so under the threat of punishment. However, there is a significant difference between this behaviour and the one observed when counter-punishment was possible.

[12] The maximum number of sanctions possible was: 1080.

Result 1: *The simultaneous introduction of punishment and counter-punishment causes only a minor aggregate increase in the average contribution level, which is considerably smaller than the one caused by one-sided punishment.*

Support for the first result comes from table 3. On the upper part of table 3, comparison of columns 2 and 3 shows that in sessions 1 and 3 we had an increase on the average contribution level, whereas in session 2 (when the VCM was played first) counter-punishment led to a decrease[13]. To have a basis for comparison, next to the results of our treatment with one-sided punishment, we present the aggregate results from F&G, as well as the ones from their third session, which was identical to our control[14].

The first thing that one should notice is the striking similarity of the results in the VCM treatment between the two experiments (3.59-3.7). On average, the contribution level increases from 3.59 to 3.77, that is by 5%, which is substantially different from the 211% increase that the introduction of one-sided punishment caused in F&G or even the 51% in our control session. Additionally, one has to notice, in the last rows, the similarity of average contribution between the two samples in the one-sided punishment treatment across all periods (10.4-10.7). However, whereas in F&G average

[13] It has been shown that the outcome of a public goods game is largely dependent on the mixture of selfish and altruistic individuals, and the environment in which the game is played (Fehr and Fischbacher [2003]). In session 2, four participants could be characterized as "perfect free riders" as they contributed zero in all periods. These subjects were able to drag down cooperation very quickly.

[14] Fehr and Gaechter also had three independent observations, each with 24 subjects. Dufwenberg and Gneezy (2000) have shown that there is no difference in the results when using 12 or 24 subjects.

contribution was higher in the final period, in our case, there was an end-of-treatment effect[15].

Table 3: Mean contributions in the stranger-treatment

	mean contribution in all periods		**mean contribution in the final periods**	
Session	VCM	P&CP	VCM	P&CP
1	3.97	6.80	2.17	3.83
	(1.66)	(1.71)	(2.69)	(3.13)
2	3.55	2.47	0.58	0.92
	(3.23)	(1.86)	(1.44)	(1.51)
3	3.24	5.83	2.58	4.50
	(1.2)	(0.93)	(5.74)	(3.94)
mean	3.59	3.77	1.78	3.08
	(1.85)	(1.41)	(3.75)	(3.34)
	VCM	Punishment	VCM	Punishment
FG mean	3.7	11.5	1.9	12.
	(5.7)	(5.9)	(4.1)	(5.6)
FG session 3	4.5	10.7	2.0	13.1
	(6.0)	(4.9)	(3.8)	(4.0)
NSN control	6.9	10.4	2.83	9.25
	(2.29)	(1.14)	(4.20)	(5.83)

Note: The numbers in parentheses indicate standard deviations. In session one the treatment with P&CP was played first and then the VCM whereas in sessions two and three the roles were reversed. NSN refers to the author's initials.

These findings support our hypothesis that counter-punishment would eliminate, to a large extent, punishment's positive effect on cooperation.

Our next result deals with the evolution of average contribution over time.

Result 2: *Average contribution exhibits a similar behaviour in the VCM and the P&CP treatments, by staying at very low levels and declining over time.*

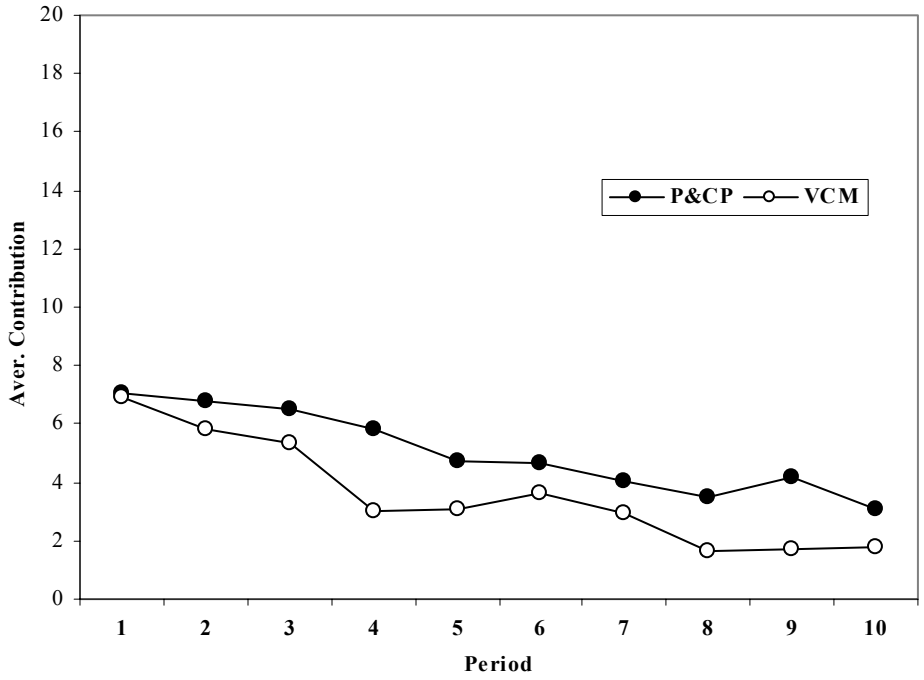
A first indication for result 2 can be found in table 3 by comparing columns 2 to 3, and 4 to 5: we can see that there is only a small difference between

[15] The evolution of average contribution in the control treatment can be seen in figure 8 in the appendix.

the two treatments. By focuing at columns 3 and 5, we observe the decline in average contribution with repetition.

Figure 1, illustrates result 2 better and shows how strong the effect of counter-punishment is on the one of punishment in the stranger-treatment. In experiments with one-sided punishment, average contribution was increasing over time or at least was non-decreasing. The same behaviour arose in our control treatment. However, when counter-punishment is possible, average contribution is decreasing in both treatments and is very similar. This suggests that in the stranger-treatment counter-punishment balances off the punishment effect to a great extent.

Figure 1: Average contribution over time in the stranger-treatment (session 1, 2 and 3)

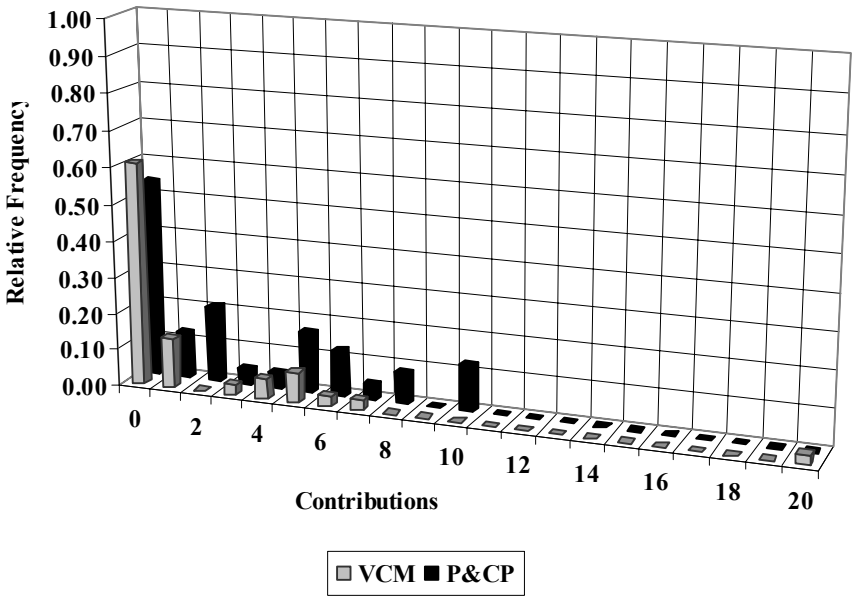


Results 1 and 2 deal only with average contribution. To have a deeper understanding we take a look at the behavioural regularities at the individual level. Result 3 summarizes the findings.

Result 3: *There is very similar behaviour in the final period of both treatments and free riding emerges as the modal action.*

The aforementioned result comes from figure 2. Although there appears to be a greater variation in the final period in the P&CP treatment, complete free-riding arises as the modal action and there is a total absence of participants who contributed more than 10 ECUs. This is in total antithesis of full cooperation being the mode in experiments with one-sided punishment.

Figure 2: Distribution of contributions in the final period of the stranger-treatment.



4.2 *The impact of counter-punishment in the partner-treatment*

Under the partner protocol, there were 91 sanctions, 30% of which were answered back[16]. Of the latter, 40.7% counter-punished with more points than originally received and 44.4% with just as many.

[16] The maximum number of sanctions possible was: 720.

The first result in the partner-treatment deals with the average contribution over all periods.

Result 4: *The introduction of punishment and counter-punishment opportunities causes a rise in the average contribution level.*

Table 4: Mean contributions in the partner-treatment

	mean contribution in all periods		**mean contribution in the final periods**	
Group	VCM	P&CP	VCM	P&CP
1	4.45	13.03	0	10
	(2.55)	(1.44)	(0)	(3.56)
2	0.73	2.33	0.25	0
	(1.51)	(3.09)	(0.5)	(0)
3	1.58	7.73	0.25	0.5
	(3.20)	(6.30)	(0.5)	(0.58)
4	3.7	7.15	0	3.25
	(3.90)	(2.84)	(0)	(3.95)
5	2.95	7	0	5
	(3.24)	(1.07)	(0)	(5.77)
6	7.85	13	0	0.25
	(5.52)	(5.68)	(0)	(0.5)
Mean	3.54	8.37	0.07	2.71
	(4.1)	(5.32)	(0.28)	(4.61)
	VCM	Punishment	VCM	Punishment
FG mean	7.5	17	3.2	18.2
	(6.8)	(4.5)	(4.4)	(2.3)
FG session 5	7.59	17.58	2.57	18.33
	(6.8)	(4.67)	(4.79)	(5.35)
NSN Control	6.35	14.78	3	12.5
	(2.71)	(2.15)	(6.16)	(8.27)

Note: The numbers in parentheses indicate standard deviations. In session four (groups 1, 2, 3) the treatment with P&CP was played first and then the VCM whereas in session five (groups 4, 5, 6) the roles were reversed. NSN refers to the authors initials.

Evidence for result 4 can be found in table 4. By comparing column 2 with column 3 we notice that contribution has increased on average in all the groups. According to a Wilcoxon matched pairs test, with group averages as observations, this difference is statistically significant (p=0.028, two-tailed). On average, subjects contribute from 1.7 (group 6) to 4.9 (group 3) times more than in the no-punishment condition. In the P&CP condition, participants contribute on average 42 percent of their endowment. The

increase in contribution, in comparison to the VCM treatment (136%), is similar in amount to the one found by F&G, although the aggregate levels in both conditions seem to be half in our case.

On the lower part of the table we can see that in both F&G and in our control, which is identical to session 5 of F&G, the introduction of one-sided punishment rises contribution on average[17]

If we compare column 2 with column 4 and column 3 with column 5 we find again that in both treatments and for all 6 groups there has been a decline on the average level of contribution. In the final period of the P&CP, participants contribute on average only 2.71 ECUs. This can be summarized by result 5.

Result 5: *Both in the VCM and the P&CP conditions of the Partner-treatment average contributions decreased sharply over time.*

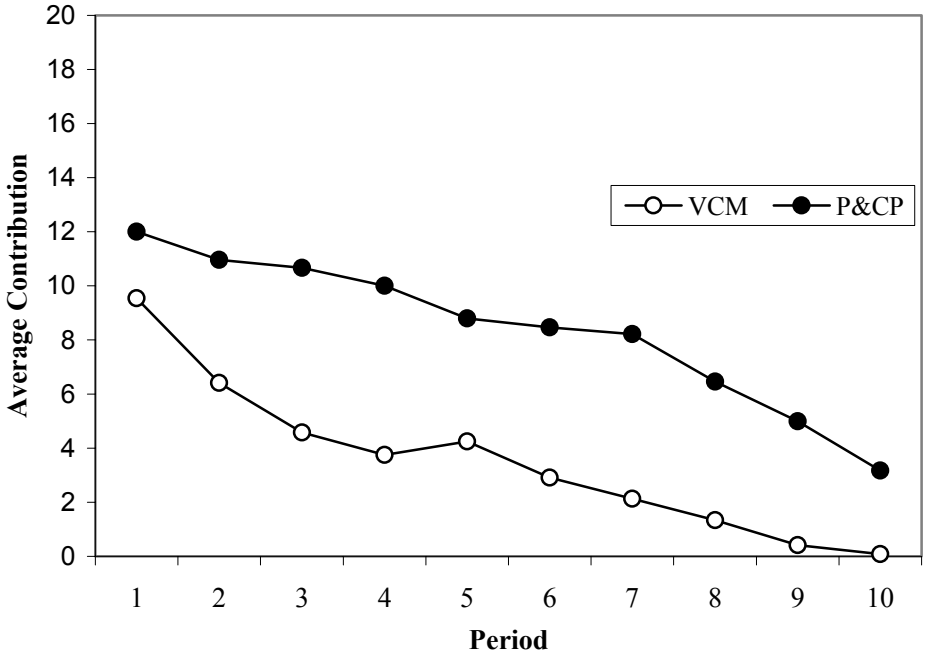
Result 5 is better illustrated by figure 3, which once again shows that counter-punishment draws away most of the power that punishment had to discipline free riders. In both sessions, the average contribution to the public good in the P&CP treatment initially is roughly 12 ECUs and then follows a similar negative trend until it settles at approximately 3.5 ECUs. The VCM treatment has the same characteristics as in most reported experiments. People are conditionally cooperative and begin by contributing a significant fraction of their endowment which varies between 40 percent (session 1) to 60 percent (session 2). However, soon the free-riders drug the cooperation in both cases down until it reaches almost complete free riding.

The rate at which average contribution declines in figure 3 is similar for the two treatments. In the P&CP treatment there is on average a higher

[17] Although this is true for all 3 groups in our control, cooperation in group 3 remained at low levels. The explanation is the same as the one given in footnote 13. There was also an end-of treatment effect. All these can be found in the appendix.

level of contribution, which was less obvious in the stranger-treatment and might be attributed to the willingness to avoid disapproval (Masclet et al.[2003]) or at the repeated interaction between the participants (Fehr and Fischbacher [2003]) . Still in both cases, the subjects start contributing less as they become more experienced and cooperation falls at very low levels.

Figure 3: Average contribution over time in the partner-treatment (session 4 & 5)



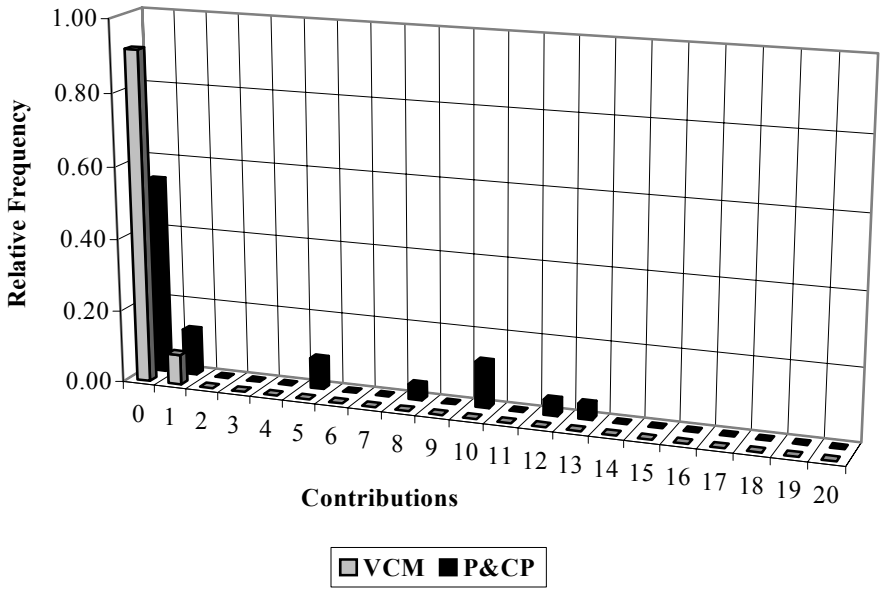
Our last result concerning the partner-treatment has again to do with the behavior at the individual level in the final period.

Result 6: *In both treatments, free-riding emerges as the modal action.*

Evidence for result 6 is drawn from the histogram in figure 4, which shows the relative frequency of contributions in the final period. As we can see, for both treatments zero contribution is the mode. In the P&CP condition, 54 percent of the participants choose to free-ride completely and 13 percent

more to contribute just one ECU[18]. There are some individuals with higher contributions. In the VCM treatment, 92 percent decide to free-ride completely and the remaining 8 percent contribute one ECU.

Figure 4: Distribution of contributions in the final periods of the partner-treatment.



4.3 *Willingness to punish*

So far we have shown that the introduction of counter-punishment opportunities has a drastic effect to the level and the evolution of average contribution if compared to treatments employing one-sided punishment. The initial contributions in F&G are very similar to ours, however, as the experiments proceed the results diverge: in F&G, as well as in other experiments with one-sided punishment, average contribution increases with repetition, whereas in our experiment, average contribution decreases and tends towards full free riding. The question that arises therefore is what triggers this different behaviour?

[18] This is a vast departure from the 82.5 percent of participants who chose to cooperate completely in F&G when counter-punishment was absent.

Punishment is a second order public good, since everyone benefits from its existence, but every individual would rather avoid its cost. The possibility of counter-punishment and the uncertainty of its harshness make punishment more costly and people less willing to punish. If this is the case indeed, we should observe a decline in the number of sanctions, which would then explain the existence of free riding.

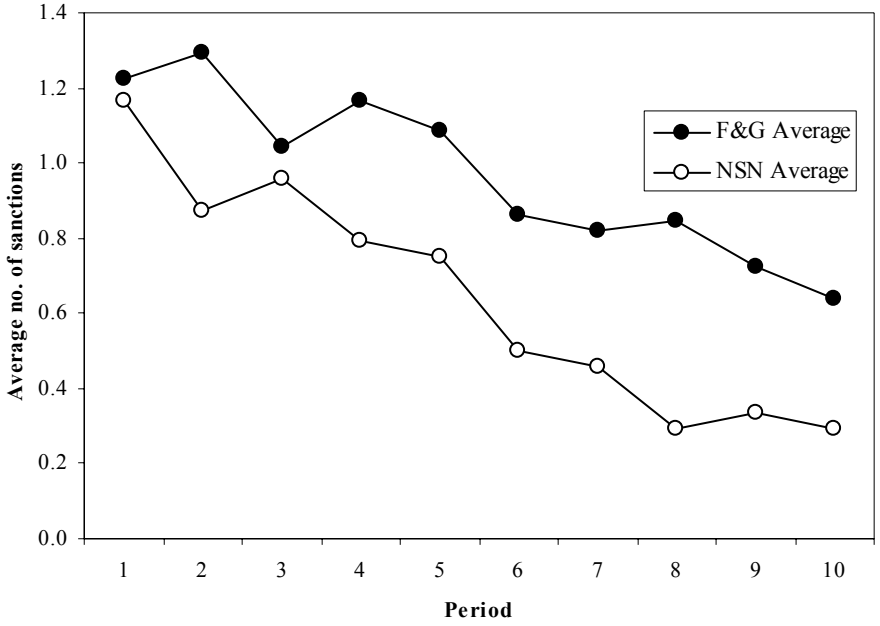
To have a basis for comparison, we will juxtapose the evolution of the average number of sanctions from this experiment and the one of F&G. Our findings are summarized by result 7.

Result 7: *Even though average contribution declines, the average number of sanctions decreases significantly in both the partner and the stranger-treatment when we allow for counter-punishment.*

Evidence for Result 7 is drawn from figures 5 and 6, which depict the evolution of the average number of sanctions over time. As we can see in figure 5, in the stranger-treatment of F&G there is a decline in the average number of sanctions over time reflecting mainly the increase on the level of contribution. The average number settles at approximately 0.65[19]. This implies that the participants, having realized the effectiveness of punishment, try to push the last non-cooperators to contribute more until the last moment.

[19] An average of "0.25" implies that on average there was one sanction per group. An average of "1" implies that on average there were 4 sanctions per group i.e. one per player.

Figure 5: Evolution of the average number of punishments sanctions in the stranger-treatment

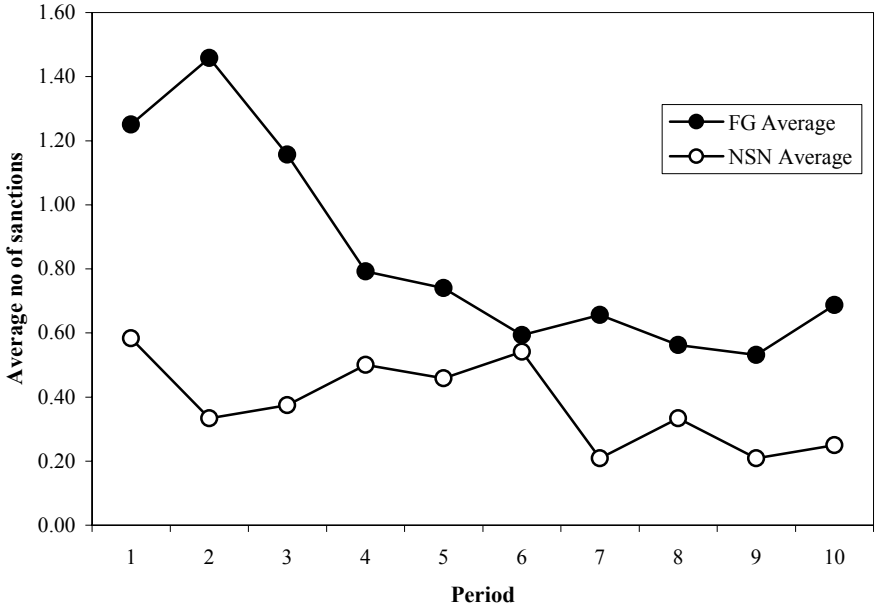


In our experiment, the average number of sanctions starts and remains at a lower level, while it pursues a similar course, which should now be attributed to the realization that the threat of punishment can not alleviate free-riding and also that punishment can be punished.

In the partner-treatment depicted in figure 6, after the second period, the average number of sanctions in F&G falls sharply following the increase in the subjects' cooperation levels and continues to do so with minor increases until it is finally stabilized around 0.6. In our experiment, the average number of sanctions in the partner-treatment is not downward sloping and it moves at much lower levels. Between periods 3 and 5 there appears to be an effort from the cooperators side to discipline the free riders in order to stop declining contribution. However, after period 5 there is a decline on the average number of punishments possibly reflecting the "surrender" of the cooperators and the average settles at very low levels (less than one punishment action per group). The difference of the average number of sanctions across treatments is statistically significant under both

protocols according to a Mann-Whitney U test with the average number of sanctions per group as observations[20].

Figure 6: Evolution of the average number of punishments sanctions in the partner-treatment



Result 7 becomes even more remarkable when we take in consideration

the fact that in our experiment, where average contribution was at a much lower level, participants had a more serious reason to want to punish. On the other hand, in the experiment by F&G, average contribution was constantly increasing approaching full cooperation eliminating the reasons for punishment. These findings lend support to the hypothesis that counter-punishment makes people less willing to punish

[20] The accuracy of these results is supported by our findings in the control treatment. Figures 10 and 11 in the appendix compare our control treatments with their F&G equivalent. The difference is not significant.

4.4 *Effectiveness of punishment*

The effect that counter-punishment has on the willingness to punish is not the only one: counter-punishment appears to diminish the effectiveness of punishment.

In F&G, 89 (78) percent of the participants increased their contribution in the partner (stranger) treatment, after they were punished. The average increase was 4.6 ECUs (3.8 ECUs). In this experiment only 30 (29) percent increased their contribution level by an average of 3.6 ECUs (4 ECUs), following a punishment. So why are people less responsive to punishment?

First, we have to see whether the actual size of the punishments is now different i.e. do people punish more lightly in order to avoid retribution? In the partner-treatment of F&G, the weighted average size of punishment was 1.71, whereas in this experiment it was equal to 2.20. So, if anything, participants punished even more on average when counter-punishment was present. The answer, therefore, to the previous question can not be found here.

The situation is reversed in the stranger-treatment, where the weighted average size of punishment in F&G was 1.90, in contrast to the 1.47 of our experiment. In this case, therefore, part of the observed lack of reaction to punishment might be attributed to the lower average size of punishment.

The decreased responsiveness to punishment might seem surprising, since even people who did not counter-punish were reluctant to increase their contribution. An explanation might be that participants, observing the modest willingness of cooperators to punish free riders, pre-emptied the decay of cooperation and chose not to raise their contribution when they themselves were punished.

Result 8: *In the presence of counter-punishment, people react less to punishment.*

4.5 *Payoff Consequences of Two-Sided Punishment*

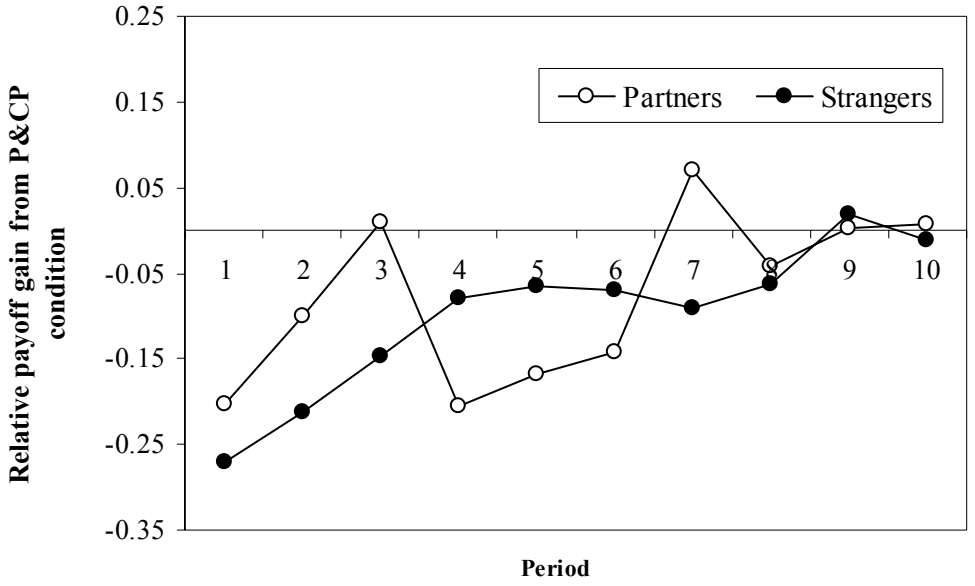
We saw earlier that the Nash equilibrium in the voluntary contribution mechanism with zero contribution (i.e. g_i=0) and an individual payoff of 20 ECUs, is not the Pareto-dominant, welfare-maximizing solution, where g_i=20 and the individual payoff equals to 32 ECUs. It has been shown [Fehr and Gaechter (2000)] that punishment alone can force people to cooperate and, though it comes with a cost (i.e. money given to buy punishment points and the income loss that punishment implies), it eventually leads to an improvement for the society as whole. Counter-punishment, on the other hand, implies additional costs and is a weapon on the hands of the free riders. One key question to be answered, therefore, is: how does the punishment option together with counter-punishment affect the average group payoff? Is the group better off now than before?

To answer this question we calculate the relative payoff gain of the punishment and counter-punishment, which is equal to the difference of the average group payoff of between two treatments normalized by the average group payoff of the no punishment treatment. In mathematical terms:

$$relative.gain = \frac{aver.group.payoff.P\ \&\ CP - aver.group.payoff.VCM}{aver.group.payoff.VCM}$$

Figure 7 depicts the payoff consequences that counter-punishment and punishment have over time in both the stranger and partner treatment. As we can see, the squander of the endowments in sanctions and counter-sanctions leads to a relative payoff loss; in 9 out of 10 periods in the stranger treatment and in 6 out of 10 in the partner.

Figure 7: Payoff consequences of punishment and counter-punishment in the partner and stranger treatment



Under the stranger protocol there is an almost constant convergence towards 0, which is the result of the declining number of punishments (see figure 9) and of the almost identical contributions between the VCM and the P&CP treatment (figure 3). In the partner-treatment, however, where punishment is more effective in raising contributions and there are implicit opportunities for coordination, relative payoff follows a more turbulent path. In the last two periods, in both conditions the relative difference approaches zero, which implies that the average payoff in the different treatments is approximately identical. This is the result of similar contributions and the declining number of sanctions.

This finding indicates how harmful mutual monitoring can be to a society. It also demonstrates that in the presence of counter-punishment, where controlling the free riders is harder, participants might be better off free riding and avoiding costly confrontations. In combination to the previous results it serves as a sign that counter-punishment might lead, eventually, to similar outcomes to the treatment where no punishment was possible i.e. similar contributions, no punishments and similar payoffs.

Result 9: *Under both protocols, punishment with counter-punishment leads to a relative payoff loss for most of the experiment until the participants learn to behave as in the VCM treatment i.e. not contribute and not punish.*

4.6 *Selfish vs. Altruistic individuals*

The careful reader might have noticed in table 3 differences in the contributing behaviour between the different groups. This observation in combination to the limited number of counter-sanctions that preclude us from a regression analysis behind the counter-punishment driving forces, makes a deeper look at the individual actions essential.

In general, in contrast to the experiments with one-sided punishment there seems to be a big variation in individual activities that seems to decrease as we approach the end. Under both protocols, the initial contributions vary from 0 to 20 ECUs. Most of the subjects appear to decrease their contribution over time, whereas some keep it relatively constant at either high or low levels of contribution and some appear to be undecided about whether to contribute a lot or little. Some individuals contribute zero throughout the P&CP treatment[21].

Table 5 summarizes the results from the partner treatment and is particularly useful since we can observe how the actions of a participant affect the future decisions of the other group members[22]. It appears that it takes only one determined free-rider to bring cooperation down. This cannot be better illustrated than in the case of group 6 (participants 21-24), where 3 participants were strong supporters of cooperation contributing for most of the experiment 20 ECUs. Subject 22, who contributed not more than 13 ECUs at any instance, forced the other three members to drop substantially

[21] It is interesting to observe that most of these participants also spend no money on punishment activities.

[22] For space economy, the respective table from the stranger treatment is available upon request.

their contributions from period 7 onwards. Note that none of the cooperators used punishment extensively. The ability of the free riders to obliterate cooperation under this set up can also be seen in the cases of group 3 (subjects 9-12), group 4 (subjects 13-16), and in lesser extent, group 2 (subjects 5-8).

Another notable case is group 5: subject 20, a strong reciprocator[23], spent most of his money in the experiment to sanction the other group members. However, his 77 points (!) were not enough to increase cooperation within the group. Consequently, by the end of the experiment he had also decreased his contribution.

An enlightening exception to this behaviour is group 1 (subjects 1-4). All four members were like-minded people whose initial contributions did not vary greatly. As a result, though they could not increase cooperation, they were able to sustain it at the initial levels. All these are summarized in result 10.

Result 10: *The level of cooperation, when counter-punishment is allowed, depends on whether or not selfish individuals exist: one determined selfish individual can obliterate cooperation like in the VCM treatment. Cooperation seems possible only between like minded individuals.*

5. Conclusion

In the last years, there has been a considerable amount of papers indicating the efficiency of mutual monitoring and sanctioning among the members of a group in providing public goods. These papers show that contrary to standard economic theory people are willing to punish and under this threat contribution levels raise significantly. However, in most cases, we can not allow for punishment and exclude counter-punishment. Our hypothesis is

[23] A "strong reciprocator" is an individual willing to engage in costly activities, even when they yield no future material benefits for him (Herbert Gintis [2000]).

that punishment elicits negative emotions amongst the punished, which in turn might lead to counter-sanctions.

Our results show that when we introduce counter-punishment, punishment stops being a valid mechanism for the discipline of selfish individuals and the efficient provision of public goods. Under both the stranger and the partner protocol, contributions decrease over time and in some cases approach full defection.

The reason behind this behaviour is the decreased willingness of cooperators to turn into punishment activities in order to alleviate free riding. In this environment, one determined free rider appears to be enough to bring down cooperation.

Mutual monitoring amongst individuals is now a harmful devise since it leads to a large squander of resources without any beneficiary result until the point where participants actually realise that they can not control the free-riders and give up cooperating. In our opinion, this serves as a warning that, in many cases, people are unable to achieve cooperation and a formal independent body is needed to enforce it.

The situation might even be understated. We believe that one of the characteristics of the individuals who chose to free ride in the real world is often their relative "strength" to the cooperators. In that case, people might be even less willing or not willing at all to punish free riders in fear of a severe counter-punishment.

An additional reason, which affects the willingness to punish negatively might be the group size; punishment is a second order public good, counter-punishment, however, is not. As a result, we believe that the greater the group size, the weaker the incentive to punish will be.

On the other hand, if people are willing to punish cooperators who did not punish free-riders this might lead to higher levels of cooperation than the ones reported in this paper.

Our results are mostly related to that of Carpenter (2002) who shows that when the price of punishment increases the demand for it decreases.

This diminishes the threat of punishment and leads to a raise in free-riding. In an indirect way, the threat of counter-punishment increases the price an individual has to pay in order to punish. However, in our view, punishment comes always at an (expected) high cost since it cannot be separated from punishment.

References:

Bowles, S. and. Gintis, H. (2002). 'Social Capital and Community Governance', *Economic Journal,* vol.113 (483), pp. 419-436.

Burlando, R. and Hey, J. (1997). 'Do Anglo-Saxons free ride more?', *Journal of Public Economics,* vol.64, pp. 41-60.

Carpenter, J. (2002). 'The Demand for Punishment', mimeo.

Davis, D. and Holt, C. (1993). *Experimental Economics,* New Jersey, Princeton University Press.

Dufwenberg, M. and Gneezy, U. (2000). 'Price competition and market concentration: an experimental study', *International Journal of Industrial Organization*, vol. 18, pp. 7-22.

Falk, A., Fehr, E. and Fischbacher, U. (2000). 'Informal Sanctions', mimeo, University of Zurich

Fehr, E. and Fischbacher, U. (2003). 'The nature of human altruism', *Nature*, vol. 425, pp. 785-791.

Fehr, E. and Gaechter, S. (2002). 'Altruistic Punishment in Humans', *Nature*, vol. 415, pp.137-140.

Fehr, E. and Gaechter, S. (2000). 'Cooperation and Punishment in Public Goods Experiments', *American Economic Review,* vol. 90 (4), pp. 980-994.

Fehr, E., Gaechter, S. and Kirchsteiger, G. (1997). 'Reciprocity as a constant enforcement device- Experimental Evidence', *Econometrica,* vol. 65 (4), pp. 833-860.

Fischbacher, U. (1999). 'z-Tree: Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual', mimeo, Institute for Empirical Research in Economics, University of Zurich.

Gintis, H. (2000). 'Strong Reciprocity and Human Sociality', *Journal of Theoretical Biology*, vol. 206, pp. 169--179.

Girard, R. (1979). *Violence and the Sacred*, Johns Hopkins University Press.

Isaac, R. M., Walker, J. and Thomas, S. (1984). 'Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations', *Public Choice,* vol. 43, pp. 113-149.

Ledyard, J. (1995). 'Public Goods: A survey of experimental research' in: Handbook of Experimental Economics ed. by J. Kagel and A. Roth, Princeton University Press

Marwell, G. and Ames, R. E. (1981). 'Economists Free Ride, Does Anyone Else? Experiments of the Provision of Public Goods,' *Journal of Public Economics*, vol.15, pp. 295-310.

Masclet, D., Noussair, C., Tucker, S. and Villeval M.C. (2003). 'Monetary and Non-monetary Punishment in the Voluntary Contributions Mechanism', *American Economic Review*, vol. 93 (1), pp. 366-380.

Ostrom, E., Walker, J. and Gardner, R. (1992). 'Covenants with and without a sword: Self governance is possible', *American Political Science Review,* vol. 86 (2), pp. 404-417

Page, T. and Putterman L. (2000): "Cheap talk and punishment in voluntary contribution experiments", mimeo.

Sefton, M., Shupp, R. and Walker J. (2002). 'The effect of rewards and sanctions in provision of public goods', mimeo.

A.1 The control treatment

Figure 8: Average Contribution over time in the treatment with one-sided punishment (Stranger)

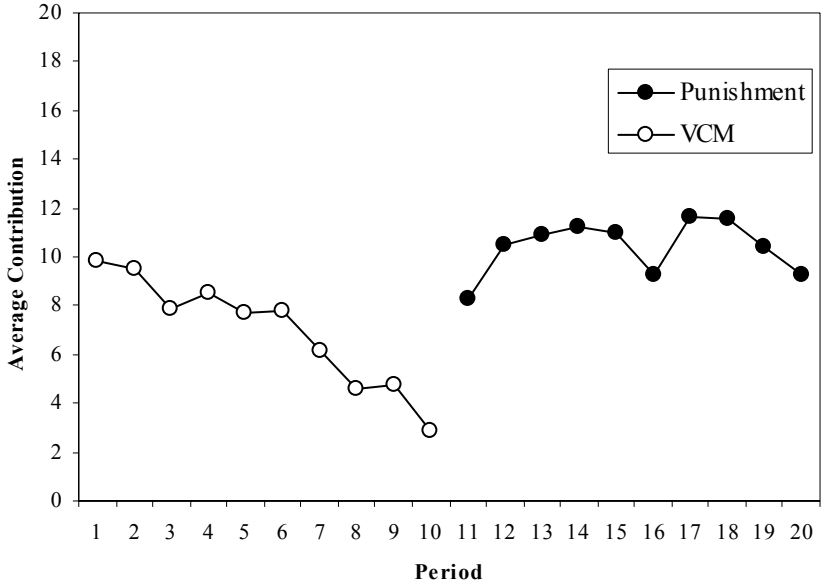


Figure 9 : Average Contribution over time in the treatment with one-sided punishment (Partner)

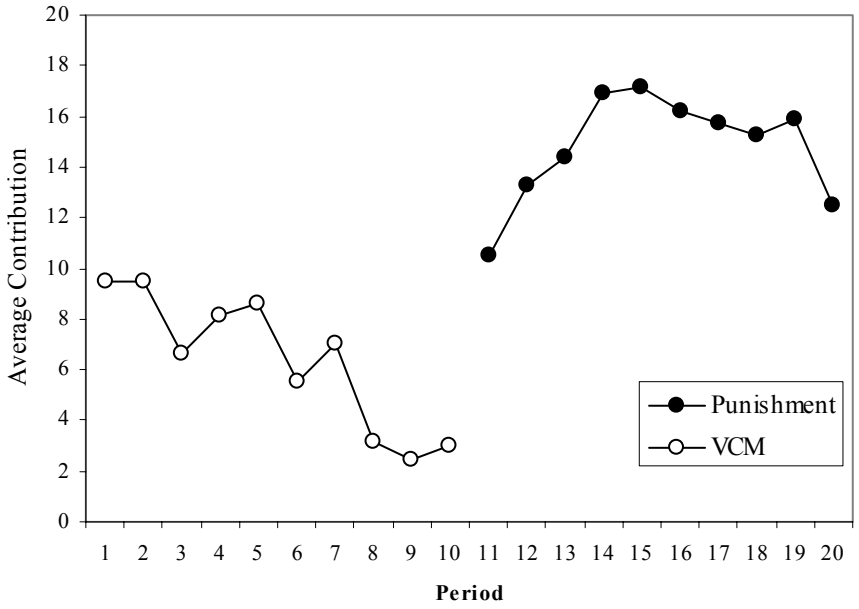


Figure 10: Evolution of the average number of sanctions with one-sided punishment (Partner)

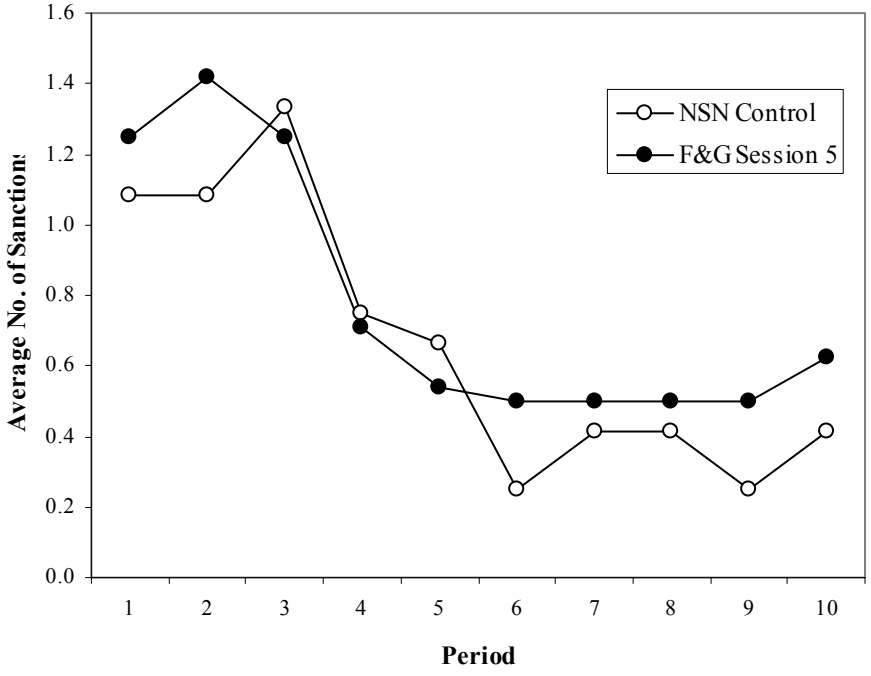


Figure 11: Evolution of the average number of sanctions with one-sided punishment (Stranger)

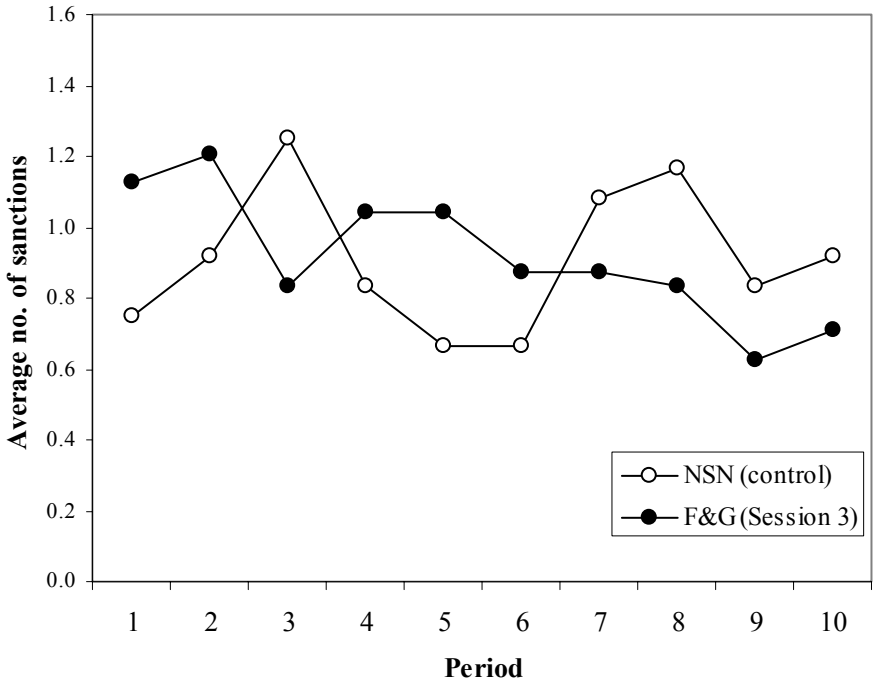


Table 5- PARTNERS

Subject	Average contribution	Evolution of contribution	Punishments given		Punishments received		Counter-punishments given	
			No of sanctions	Total points	No of sanctions	Total points	No of sanctions	Total points
1	13.4	11,12,11,15,13,15,15,15,14,13	1	1	4	4	2	3
2	14.1	13,10,15,15,15,15,16,16,14,12	6	9	1	2	0	0
3	10.7	10,10,9,11,11,10,12,12,12,10	6	8	8	14	1	1
4	13.9	14,15,15,15,15,15,15,15,15,5	3	6	3	4	2	3
5	3.3	3,10,0,5,0,15,0,0,0,0	2	2	0	0	0	0
6	4.0	5,10,8,5,2,0,0,0,0,0	1	1	1	1	1	1
7	2.0	20,0,0,0,0,0,0,0,0,0	0	0	1	1	0	0
8	0.0	0,0,0,0,0,0,0,0,0,0	0	0	1	1	1	1
9	7.4	20,20,12,13,1,1,3,4,0,0	2	4	4	12	1	3
10	8.9	15,10,20,20,15,1,5,1,1,1	3	5	7	17	5	12
11	5.1	8,9,10,2,12,5,2,1,1,1	10	27	6	9	6	16
12	9.5	20,20,20,10,10,10,1,3,1,0	2	3	1	2	1	2
13	7.0	6,8,8,8,0,8,8,8,8,8	8	25	7	14	0	0
14	9.3	20,10,12,12,12,12,15,0,0,0	4	20	4	10	2	10
15	9.8	10,11,12,12,12,15,10,11,0,5	8	13	2	4	0	0
16	2.5	5,12,0,0,8,0,0,0,0,0	4	6	11	34	0	0
17	5.6	5,7,10,7,5,0,8,8,6,0	1	2	10	25	2	5
18	0.5	0,0,0,0,1,0,4,0,0,0	1	1	10	46	0	0
19	9.9	8,10,11,10,10,11,9,10,10,10	0	0	7	9	0	0
20	11.9	15,14,13,12,11,11,11,11,11,10	25	77	0	0	0	0
21	14.0	20,20,20,20,10,20,20,10,0,0	0	0	1	1	0	0
22	7.3	10,5,10,8,7,9,13,0,10,1	3	3	1	2	1	4
23	14.7	20,20,20,20,20,10,20,10,7,0	1	2	1	1	1	2
24	17.0	20,20,20,20,20,20,20,20,10,0	0	0	1	1	1	1

Subjects 1-12 took part in session 3 and subjects 13-24 in the session 4. Subjects 1-4 formed group 1, 5-8 group 2 etc.
Contributions refer to the P&CP treatment.