MASARYK UNIVERSITY
FACULTY OF ECONOMICS AND ADMINISTRATION

The decentralization of punishments in experiments with public goods

Zuzana Berná, Jiří Špalek

Working paper WP KVE – 05/2013

The decentralization of punishments in experiments with public goods

Zuzana Berná and Jiří Špalek

Masaryk University

Faculty of Economics and Administration

Department of Public Economics

Working paper

WP KVE – 05/2013

Acknowledgements: The authors are grateful to Fabio Padovano, Yvon Rocaboy and participants of PEARL seminar in Helsinki. This research was supported by grant from Masaryk University.

Zuzana Berná: Zuzana Berná is PhD student in last year in Public Economics. She deals within her thesis with experimental testing of effectiveness of selected modifications of classical voluntary contribution mechanism. She is author of several papers on the topic.

Jiří Špalek: Jiří Špalek is associate professor at Department of Public Economics, Masaryk university. He specializes in experimental and behavioral economics, particularly experiments with public goods, charitable lotteries.

Abstract

This paper deals with the effects of introducing adequate punishment opportunities in experiments with public goods. Decentralized punishment means that the contributing subjects have a possibility to sanction free riders without the intervention of an external authority. The very first experiments demonstrated a significantly positive effect of a punishment opportunity on enhancing cooperation in situations of social dilemma. Following studies, however, pointed at limited effectiveness of this mechanism.

The first part of the paper summarizes selected literature on the topic and presents its principal findings. The second part is dedicated to the presentation of the results of an experimental series on decentralized punishment realized in the Czech Republic. The last part introduces possible questions and topics which may be subject of future research within this area.

Key words: economic experiment, cooperation, public goods, decentralized punishment, partner and stranger matching

JEL classification**:** C92, D70, H41

CONTENT

1. Introduction

Behavior of individuals facing situations called social dilemmas is a popular topic of laboratory testing. The term social dilemma was first used by Dawes (1975, 1980) and it represents the situation in which individual interest is in conflict with the social one. A typical example of such situation is voluntary contribution to public goods.

Standard economic theory predicts that social dilemma situations result in inefficient (sub-optimal) outcome. Typically, in case of voluntary contribution to public goods this means that rational individuals (in economic terms) seeking to maximize their own utility, won't contribute anything to public goods (hoping that other individuals do so). As a result, the public good will not be provided (or will be provided at a sub-optimal level). However, reality as well as results of economic experiments shows that by far not all people are selfish and act rationally in this sense.

Experimental economists (and not only them) have been searching for the meaning of cooperation in social dilemma situations and the identification of individual incentives leading to its emergence. The related challenge of laboratory testing has been to define factors having the capacity to influence the level of cooperation. Concerning voluntary contribution to public goods, an exhaustive list of such factors and their impacts has been presented in Ledyard (1995).

Those who experiment employ different schemes and modifications of the classical Voluntary Contribution Mechanism (hereafter "VCM") to study the effects of various factors. One of such modifications is VCM, with opportunity to punish free riders. The mechanism is as follows: After all individual decisions are made, information about individual levels of contribution to a public good is published and individuals are given an opportunity to sanction their co-players, as this means sanctioning without the intervention of an external authority, for example *decentralized punishment* (Nikiforakis, 2007).

The given sanction reduces the current income of a penalized subject and, at the same time, the act of punishment brings also an inevitable cost to the sanctioning subject. As it doesn’t ensure any future financial benefit to the latter, we also speak about the so-called *altruistic punishment* (Fehr and Gächter, 2002). Other appellations such as *costing* or *peer punishment* (see e.g. Guala, 2012, Casari, 2012) are used to represent the same concept of decentralized punishment. The basic model of VCM with punishment opportunities are introduced in section 2.

The original experiments (Ostrom et al., 1992, Fehr and Gächter, 2000) demonstrated a considerable and positive effect of the decentralized punishment opportunities on cooperation. Following studies have shown, however, that the effectiveness of this mechanism has several limitations. A summary of some principal findings related to the concept of decentralized punishment is explained in section 3.

Section 4 presents the design of an experiment executed with Czech students replicating the experiment of Denant-Boèmont et al. (2007). The results are discussed in section 5.

The last section (6) opens a discussion about the limits of a decentralized punishment mechanism and of the possible orientation of future researches within this area.

2. Decentralized punishment: The model

As aforementioned above, VCM experiments associated with a decentralized punishment add one (or more) stage(s) to the classical Public Goods Game (hereafter “PGG”). After all the individual contributions are made, they are published and the players are given the opportunity to punish free riders by imposing points on them. Each point received reduces the income of its receiver by a certain percentage; however, even imposing points is costly. In this sense the punishment represents a *second-order public good* (see e.g. Fehr and Gächter, 2002, or Guala, 2012). This means that while a sanction is imposed on the punisher, the whole group of contributors benefit from this act. Or, as Fehr and Gächter (2002) state: "*everybody in the group will be better off if free riding is deterred, but nobody has an incentive to punish the free riders*". The authors further claim that this problem can be solved if enough people have a tendency for altruistic punishment.

In classical PGG, individual payoffs at the end of a period are given by equation (1) (according e.g. to Fehr and Gächter, 2000).

$$\pi_i^1 = y - g_i + a \sum_{j=1}^{n} g_j \qquad (1)$$

π_i^1 represents payoff of individual *i* in given round, *y i*'s endowment in experimental monetary units *tokens* (usually the same amount for all players), g_i is *i*'s contribution to a public good[1] and *a* is marginal payoff of the public good (or Marginal Per Capita Return – MPCR), knowing that $0 < a < 1 < na$, where *n* is number of players(size of group).

Equation (1) together with the condition regarding *a*, implies two phenomena typical for the problem of voluntary provision of public goods (or social dilemma):

Dominant strategy of player *i* is full free-riding, i.e. $g_i = 0$.

Aggregate payoff is maximized if everyone fully cooperates, i.e. $g_i = y \ \forall i$.

Introduction of punishment the opportunity changes the individuals' payoffs in a way given by equation (2).

$$\pi_i^2 = \pi_i^1 \left[max\left\{0{,}1 - (1/10) \sum_{j \neq i} p_{ji} \right\}\right] - \sum_{j \neq i} c\left(p_{ij}\right) \qquad (2)$$

Here π_i^2 is i's payoff after distribution of sanctions, π_i^1 her initial payoff after investment stage, p_{ji} is number of punishment point assigned to player *i* by

[1] Public good in a game is usually represented by some common project or group account.

player j and $c(p_{ij})$ is a convex function representing cost paid by *i* associated with the r sanction of *j*.

As punishment means a financial penalty for the punisher, the model predicts that a rational individual won't engage in an altruistic punishment. All individuals know this and as the threat of punishment doesn't become credible, they will have no reason to change their behavior in investment stage. They will contribute zero ($g_i = 0$) and the game will end in the same equilibrium outcome as the classical PGG.

However, the same as in case of a classic public goods game, experimental results have shown that not all individuals are rational (in economical terms); the punishments occur and have the power to influence the level of contributions. The following section presents the principal findings of experiments dealing with decentralization of punishments carried out on an up to date basis.

3. Principal findings up to date

The original authors who experimentally tested the impacts of the opportunities of decentralized punishment were Ostrom et al. (1992). Their experiment focused on common pool resources. The motivation for their survey was the disparity between the economic theory predicting that individuals are not able to negotiate and apply a common strategy leading to efficient exploitation of a common pool resource (without the intervention of an external subject), and the empirical experience. The latter has demonstrated repeatedly that fishers, herders and the other "users" of a common pool resource (called by the authors "*appropriators"*) are able to organize themselves in order to create binding commitments, to control their respect and sanction potential non-respect. The experiment combined use of communication among subjects with sanction opportunity. The results showed especially that subjects who employed communication to create an agreement on common investment strategy and opted for their own sanctioning mechanism, achieved close-to-optimal results. The authors concluded that individuals are able to find an optimal strategy which enables them to efficiently use the common pool of resources. To accomplish this task they needed sufficient information, an arena for discussion and (alternatively) monitoring and sanctioning. Under these circumstances it is not necessary to create a sovereign who governs, monitors and sanctions other subjects (for example , as stated in Hobbes, 1960).

Fehr and Gächter (2000) executed the first experiment studying decentralized punishment in situation of voluntary contribution to public goods. The authors compared the results of experimental treatment which was a good classical public game, because of its modification in which players got an opportunity to punish (on a one-time basis) free riders. The effect of punishment was observed separately for partner and stranger matching treatments[2]. The results demonstrated significantly the positive effect of a punishment opportunity on the level of an average contribution towards a public good; under both matching treatments it led to a considerable increase of cooperation. The Players' responses to different contribution scenarios (which were presented to them during the experiment by experimenters) showed that free riding causes strong negative emotions and in relation will lead to punishing free riders. These negative emotions are, at the same time, anticipated by the majority of subjects. These findings were confirmed also in Fehr and Gächter (2002). These authors concluded that under the possibility of the decentralized punishment of free riders a very high (or even full) cooperation may be achieved and maintained, while without this possibility the same subjects resume to the act of full free riding.

In the furtherance of their work on this topic (Fehr and Gächter, 2002), the authors show experimentally that an altruistic punishment is a key motive for the explanation of cooperation in situations where other theories (such as the theory of kin selection, theories of direct and indirect reciprocity or costly signaling) lack arguments. These situations include cooperation among genetically unrelated people, often in large groups where they don't meet each other more than once and gains from reputation are minuscule. Punishment is

[2] The meaning of different matching types will be presented further in the paper.

considered as an altruistic act as it provides a material benefit for the future interaction of the partners of the punished subject and not for the punisher , to whom however, bears the costs of imposing the sanction. The authors stated that "*altruistic punishment is a key force in the establishment of human cooperation*" and from this point of view, the presented evidence has important implications for the evolutionary study of human behavior.

A common feature of the studies mentioned above was that the experiments used only one punishment stage (IE. after the investment stage subjects were given one-time opportunity to reduce payoffs of their co-players and once this happened, the experiment proceeded to next level). However, as in Nikiforakis (2008) they argued, this makes it impossible for subjects to take revenge for imposed punishments, although such a possibility exists in almost all decentralized interactions in reality. According to the author, the omission of the threat of counter-punishments from an analysis could lead to the overestimation of the effectiveness of the decentralized punishment, misleading conclusions or even to the implementation of unsuitable policy. In his experiment using a modified PGG, Nikiforakis (2008) studied the effects of the introduction of the possibility to counter-punish on cooperation and welfare. The results showed that in the presence of counter-punishment opportunities the willingness to punish free riders decreases which in turn results in a breakdown of cooperation and of lower earnings. Approximately one quarter of punishments are retaliated while the counter-punishments seem to be driven partly by strategic considerations and partly by a desire to avenge punishments. The results of the experiment cast doubt over Ostrom's et al. (1992) concept saying that self-governance is possible.

Dunant-Boèmont et al. (2007) enriched the existing literature on decentralized punishment with a new feature: the use of punishments for repeated sanctioning of low contributors in order to enhance cooperation. Such an act is called by authors *sanction enforcement* and it takes two forms: sanctioning of those who fail to punish low contributors and those who punish high contributors. In their study, the authors raised a question on whether the effect of sanctioning enforcement on cooperation would be stronger than the effect of counter-punishment or reversely. The results showed that the significant negative effect of counter-punishment prevails over the positive one of sanction enforcement, which is not statistically significant, and the overall effect is negative. The second issue the authors examined was the effect of multiple opportunities to sanction. In regard to this topic, the results showed that in addition of the multiple rounds of sanctioning reducing the level of contributions (compared to the setting of the single punishment stage) as well as welfare (compared to the setting of no sanctions or of one part of the sanctions). The authors concluded that punishments are the most likely to help increase the level of cooperation if the sanctioning subjects stay anonymous.

A different experimental design enabling the multiple stages of sanctioning was introduced by Nikiforakis and Engelmann (2011). The authors examined how the threat of *feuds* (defined as long-running arguments between parties) influences individuals' willingness to engage in an altruistic punishment. In their (modified) PGG the number of sanctioning stages was determined by the subjects' actions and there were minimal restrictions on who may sanction whom and when. According to the authors this extended the PGG with punishment opportunities to its natural limit. The results showed that subjects recognize the threat of feuds and when punishment can lead to a long (inter-period) feud, their willingness to engage in altruistic punishment is considerably reduced. As result, the level of cooperation decreases over time which leads also to decline in earnings (compared to treatment without any possibility to act in retaliation with the sanctions). This implies, however, a reduced role of the altruistic punishment in explaining cooperation.

Some authors focus also on the role of the effectiveness of punishment, IE. the ratio between cost and the impact of the punishment. Nikiforaktis and Normann (2008) examined in their experiment of four levels of punishment effectiveness (four different levels of cost born by the punished subject) and they found out that meant the contributions increased monotonously in the effectiveness of the punishment. The trend of the contributions was likewise influenced by the different effectiveness. As for earnings, the authors concluded that the punishment effectiveness of minimally the ratio of 1/3 (which means that of at least one point that is obtained, this reduces the payoff of the punished subject by three) is required to obtain a welfare improvement (compared to PGG without punishment).Their results concerning level of contribution were in accordance with Egas and Riedl (2005). These authors examined four different effectiveness levels of punishment; they varied in both impact and of the cost of punishment. Their results regarding welfare, however, showed an inverse relationship than of the above mentioned (which may be partly due to the fact of different matching types[3] in experiments). The authors concluded that the altruistic punishers also take into account the costs and effects of their actions, and altruistic punishment needs to be combined with other cooperation-enhancing mechanisms (e.g. reputation, reciprocity or the possibility of opt-outs).

Bochet et al. (2006) experimented in the combination between decentralized punishments with the possibility of communication. Experimental results demonstrated mainly the strong positive effect of verbal (primarily face-to-face) communication; the increase in the amount of cooperation was at such a high level (compared to treatment without communication) that completing it with additional punishment opportunities didn't lead to a significant increase in the level of contribution. This is in line with the results of Ostrom's et al. (1992) experiments which pointed especially at the positive effect of communication (combined with punishment). The net effect of punishment on efficiency (i.e. welfare) was zero.

4. Experiment

4.1 Motivation

Our experiment replicated four treatments of Denant-Boèmont's et al. (2007) experiments studying the effects of counter-punishment and sanction

[3] Egas and Riedl (2005) employed in their experiment absolute stranger matching while Nikiforakis and Normann (2008) used partner matching. Discussion of different matching types is to be found in the following section.

enforcement[4]. Each treatment constituted a single session in which 24 subjects took part. The participants played in groups of four. The modification vis-à-vis the original experiment was that the composition of these groups changed within every round (so-called *stranger matching*).

The original experiment was executed in so called *partner matching,* which means that the subjects interacted with the same co-players in every round of an experimental session. Alternatively, the setting that involves the so called stranger matching, implies that the group composition changes randomly before each round, and such a setting represents a good approximation to the single-shot experiments, since reputation effects are eliminated (a "perfect approximation" would be under *perfect stranger matching* ensuring that two subjects don't meet more than once during a single session). If reputation matters one would expect partners to cooperate significantly more than strangers (Andreoni and Croson, 2008). However, the first study dealing with this question (Andreoni, 1988) showed just the opposite. Starting with a previous paper, there has been an intensive discussion whether cooperation is higher under partner settings or not. Andreoni and Croson (2008) bring a synthesis of replications and studies on this topic. According to it the picture remains quite unclear, as *"four studies find more cooperation among strangers, five find more by partners and four fail to find any difference at all"*. The aim of the Czech experiment was to enrich and complete the data acquired by Denant-Boèmont et al. (2007) by results obtained under stranger matching. The motivation was the question whether a different matching type would influence individual contributions and the willingness to engage in costly punishment or not.

4.2 Overview

A set of experiments took place at Masaryk University in Brno during academic year 2009-2010. The participants were recruited among undergraduate students of different faculties of Masaryk University by means of an

[4] In our experiment we didn't focus on studying effects of multiple stages of punishment which was the second issue considered in the original experiment.

advertisement published in university's information system. In total, 96 subjects participated in the experiment. Average individual earnings were 230.5 CZK (about 9 euro). All experimental sessions were run on computer terminals using z-tree[5] program.

Each of the four treatments consisted of 20 identical rounds (repetitions). In the beginning of each treatment, participants played one trial round so that they make certain that they understood the instructions[6].

The basic treatment called *Baseline* consisted only of two stages in each round. The first stage which we may call the *investment stage* was of a classical VCM. Within this stage the participants were given a certain amount of disposable income and they had to decide which part of it they would keep on their personal account and which part they would invest to a group account. Then a punishment stage followed at the beginning of which the players learned about individual investments to the group account and they received a subsequent opportunity to assign points to their co-players, reducing their current income. At the end of the punishment stage players were informed about their original income (after the first stage), number of points received and of the total payoff from a round. The generators of received sanctions were hidden from the players.

The three other treatments contained one more punishment stage and the only difference among them was the character of the published information about punishments assigned. In this second punishment stage the players had the opportunity to punish again all of their co-players in a group. In *Revenge Only* treatment all players learned after the first punishment stage who and by which amount sanctioned were given to them personally. In the *No Revenge* treatment, on the other hand, they were informed about all the punishments excepting those assigned to them. Whilst in the *Full Information* treatment they learned about all the assigned punishments and their generators.

[5] See Fishbacher (1999).

[6] Instructions are available upon request to the authors.

Different treatments allowed for the use of various punishment strategies. In *Baseline*, the subjects used punishments only in response to contribution decisions made at the first stage. *Revenge Only* treatment allowed, in addition to above mentioned, the use of counter-punishment. *In No Revenge* the possibility of counter-punishment was eliminated while the subjects were allowed to engage in sanction enforcement, as well as punishing their co-players in response to first-stage contributions. *Full Information* treatment allowed all of the sanction strategies mentioned above. Therefore, the difference in the contribution levels between Baseline and Revenge Only treatments, as well as difference between No Revenge and Full Information measures the marginal effect of counter-punishments on cooperation. On the other hand, the difference in contributions between Baseline and No Revenge, as well as between Revenge Only and Full Information represents the marginal effect of sanction enforcement (Denant-Boèmont et al., 2007).

4.3 Calculation of payoffs

During the experiment the payoffs were calculated in experimental monetary units - *token*. At the end of each session the total sum acquired was converted into CZK, using the exchange rate of 1 token = 0.50 CZK, and subsequently paid to participants. The calculation of profits was based on Fehr and Gachter's (2000) design.

In the beginning of each of the investment stages, the subjects were given 20 tokens and were also asked to decide how many tokens they would keep (on their personal account) or invest to a group account, which is common to all players in a given group. Each token kept on private account maintained its value (ratio 1:1), while each token invested to a group account yielded 0.4 tokens[7] to every player of a group. Calculation of payoffs at the end of each investment stage is given by equation (3).

$$\pi_i^1 = 20 - g_i + 0.4 \sum_{j=1}^{n} g_j \quad (3)$$

[7] This means that MPCR of a public good is equal to 0.4.

At the end of this investment stage subjects learned about their current profits and individual contributions (of their co-players) towards the group account. Then a punishment stage followed whereas each player had the opportunity to reduce payoffs of their co-players by assigning them points (0-10 points to each co-player). Each point received reduced its owner's profit by 10 % while 10 and more points received meant reduction by 100 % (not more). Assignment of points caused an increase of costs also to the punishing subject; he or she bears the cost from punishing each of co-players and these costs (for each co-player) are added up. The costs born by punishing subjects were a convex function, punishment points and their amount is given by Table I[8].

Table I. Cost function of points assigned

Points assigned	0	1	2	3	4	5	6	7	8	9	10
Costs of points assigned by player	0	1	2	4	6	9	12	16	20	25	30

Source: Fehr and Gächter (2000).

The calculation of individual payoffs at the end of the first punishment stage was given by equation 2.

$$\pi_i^2 = \pi_i^1 \left[max\{0,1 - (1/10) \sum_{j \neq i} p_{ji} \} \right] - \sum_{j \neq i} c\left(p_{ij}\right) \qquad (2)$$

where $c\left(p_{ij}\right)$ is convex cost function defined in Table 1, assigning cost to player *i* for punishing player *j*.

This payoff represented the total of the payoff within the *Baseline* treatment. In three other treatments one more punishment stage followed. Each point received reduced again the current profit of its receiver by 10 %. The costs of punishments assigned were (again) calculated on the basis of Table 1. Total profit at the end of the second punishable stage (i.e. total profit per round for the four treatments) was given by equation 5.

$$\pi_i^3 = \pi_i^1 \left[max \left\{ 0,1 - (\tfrac{1}{10}) \left[\sum_{j \neq i} p_{ji}^2 + \sum_{j \neq i} p_{ji}^3 \right] \right\} \right] - \sum_{j \neq i} c\left(p_{ij}^2\right) - \sum_{j \neq i} c\left(p_{ij}^3\right) \quad (4)$$

where p_{ji}^2 is the punishment of player i assigned by player j in the second stage and p_{ji}^3 is also the punishment of player i assigned by player j in the third stage.

[8] Subjects had identical table on their disposition and were able to calculate the financial consequences of their actions.

4.4 Hypothesis

Based on previous findings (e.g. Fehr and Gächter, 2000) we hypothesized that (1) the contribution levels would be considerably lower and (2) subjects would assign less punishment points under stranger matching than under partner matching (as used in Denant-Boèmont et al., 2007).

This assumption is based on the so called strategies hypothesis introduced by Andreoni (1988). According to this hypothesis, subjects – if they are rational - play in order to influence their partners' actions. As strangers they play an actual repetitive single shot game, there is no reason for them to play strategically and we can thus expect that they would contribute less than partners. The same reasoning may be used in relation to punishments: the single shot equilibrium presumes zero punishment (see e.g. Fehr and Gächter, 2000) and there is no reason to expect another result in repetitive one-shot interaction where in addition the subjects' experience plays a significant role.

In accordance with original experiment we focused in our analysis also on effects of punishments on level of contribution and welfare, and motivations of sanction behavior. As we considered there was no reason to expect fundamentally different results under stranger matching we advanced no related hypothesis ex-ante.

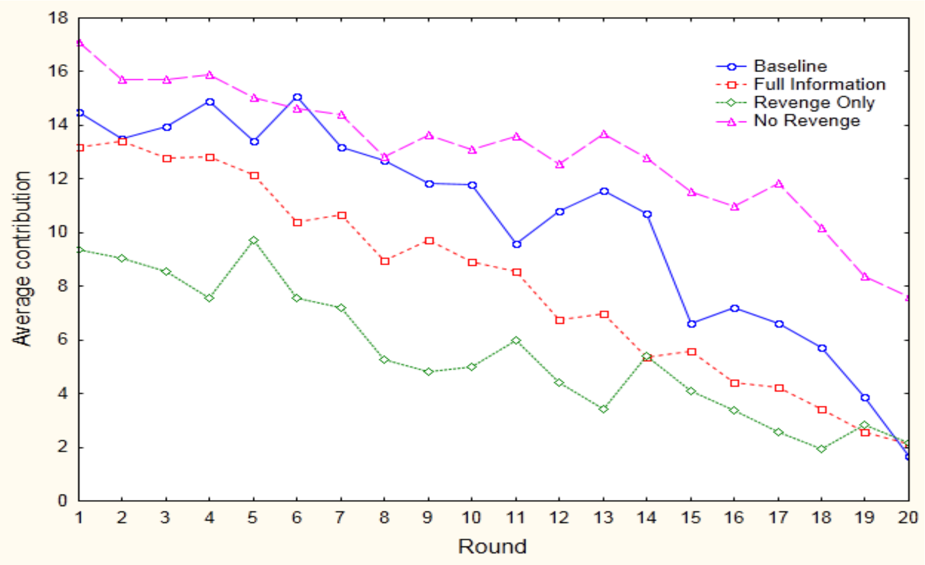
5. Results and discussion

5.1 Average contributions and individual earnings

The average individual contributions in four experimental treatments are captured by Figure I. The highest average contributions were attained in the No Revenge treatment (13.05 tokens), followed by Baseline (10.46) and Full Information (8.15), and the lowest average contributions were reached in the Revenge Only treatment (5.52). This sequence copies the results of Denant-Boèmont et al. (2007). However, the contribution levels in our experiment were actually considerably lower than in the original experiment (where average contributions were 16.17 under the No Revenge treatment, 15.49 under Baseline, 10.59 under Full Information and 7.21 under Revenge Only treatment).The difference in contribution levels between the two experiments (or the two matching types) varied from 1.69 to 5.03 tokens which supports the first part of our hypothesis saying that subjects contribute considerably less under stranger matching.

Another significant difference in relation to the original experiment was that in our experiment, the average contributions were in a decreasing trend under all four treatments[9]. This finding demonstrated that under stranger matching cooperation was not a sustainable solution. Initially despite high contribution levels (mainly in No Revenge treatment) the average contributions in all treatments tended over time to zero off, which is in line with the theoretical game predictions.

Figure I. Average individual contributions



Source: Authors

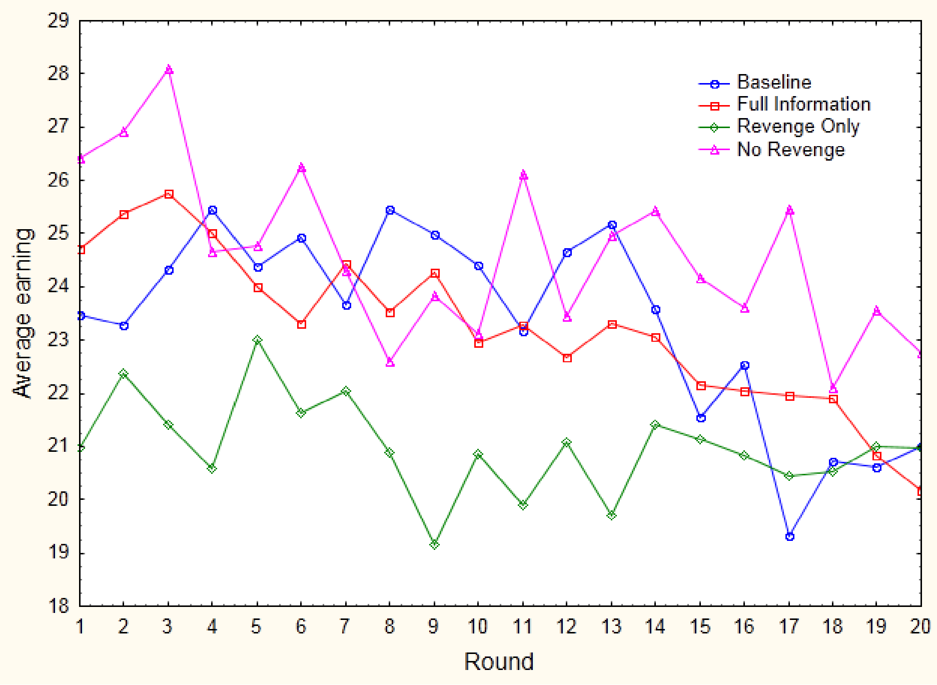
The same as in the case of the original experiment our results show that the introduction of an opportunity to counter punish has a negative effect upon the level of cooperation. The differences in contributions between the Baseline and Revenge Only treatments, as well as between the Full Information and No Revenge treatments, are statistically significant (at $p < 0.01$). This finding

[9] In Denant-Boemont's et al. (2007) experiment, the average contribution level didn't change appreciably during the game in any treatment. The exception was the decline over time in Revenge Only and an initial increase in the first periods of Baseline and No Revenge.

confirms the conclusions of Nikiforakis (2004) and Denant-Boèmont et al (2007) stating that the threat of counter punishment decreases considerably the level of contributions to a group account. The possibility to enforce sanctions has, by contrast, a positive effect on the cooperation level. The differences in contributions between the Full Information and Revenge Only treatments, as well as between Baseline and No Revenge, are statistically significant (at $p < 0.05$). There exists a demonstrable and positive effect of the possibility of a sanction enforcement to the level of contribution; however, it is weaker than the negative effect related to the threat of a counter punishment. The difference in the level of contributions between the Baseline and Full Information is statistically significant (at $p < 0.05$) which means that overall, the effect of counter punishment and sanction enforcement is negative, IE. the positive effect of a sanction enforcement is not strong enough to counterbalance the negative effect of a counter punishment. This finding is in accordance with the original experiment.

The level of average earnings under the four treatments may be observed in Figure 2. The highest individual earnings were attained in the No Revenge treatment (24.62 tokens), followed by Baseline (23.33) and Full Information (23.24). The lowest average earnings were gained from the Revenge Only treatment (21). These results are again in line with findings of original experiment of Denant-Boèmont et al. (2007); we can conclude likewise that the average earnings under the Revenge Only treatment were comparable to those corresponding to the full free riding without the threat of punishment (20 tokens).

Figure 2: Average individual earnings



Source: Authors

Our data demonstrates the positive effect of a sanction enforcement possibility, on average individual earnings. The variances in earnings are statistically significant ($p < 0.01$) in the case of Revenge Only and of the Full Information treatments and bordering significantly ($p < 0.1$) in the case of Baseline and No Revenge.

Concerning the effect of the counter punishment opportunity, the difference in average earnings is statistically significant between the Revenge Only and Baseline treatments ($p < 0,01$) but not between the Baseline or the Full Information treatments. We can conclude that the possibility to counter punish has a negative effect on individual earnings, although this phenomenon is confirmed by data only in part. The overall effect of counter punishment and sanction enforcement to average earnings (measured by the difference in earnings between the Baseline and Full Information treatments) is not statistically significant.

5.2 Sanction behavior

The second part of our hypothesis was related to the intensity of sanctions. The average quantity of sanctions assigned under the two matching types may be observed in Table II.

Table II. Average quantity of sanctions assigned in each stage of a period

		Treatment			
		Baseline	Full Information	No Revenge	Revenge Only
Average points assigned in Denant-Boèmont et al. (2007)	Stage 2	1.512	0.46	0.65	0.73
	Stage 3	-	0.57	0.37	0.38
	Both stages	**1.512**	**1.03**	**1.02**	**1.11**
Average points assigned in our experiment	Stage 2	0.68	0.15	0.52	0.30
	Stage 3	-	0.24	0.18	0.31
	Both stages	**0.68**	**0.39**	**0.70**	**0.61**

Source: Authors

As it is clearly visible from Table II, the subject is sanctioned considerably more heavily in original experiment, i.e. under fixed matching. In the Baseline and Full Information treatments the average punishment points overall were even more than double, compared to our results. This supports again our hypothesis explaining that the subjects punish less under stranger matching (i.e. when the composition of the groups changes in each round).

For the studying of motivations and of the incentives for sanction behavior, Denant-Boèmont et al. (2007) have introduced two regression functions, which are explained below. Equation 5 records sanction behavior at the first stage of punishment; the number of punishment points distributed by a subject in the first punishment stage is expressed as an elemental function of others' contributions and of a deviation in the contribution level of a sanctioned subject from the group average.

$$p_i^{j2t} = \beta_0 + \beta_1 c_{-i}^{-t} + \beta_2 max\{0, c_{-j}^{-t} - c_j^t\} + \beta_3 max\{c_j^t - c_{-j}^{-t}\} + \beta_4 t \qquad (5)$$

Variables employed in the equation 5 are indicated as follows: the dependent variable p_i^{j2t}represents the quantity of points assigned by player *i* to player *j* in the second stage in period *t*. c_j^t is the contribution of player *j* in period t while c_{-j}^{-t}

(c_{-i}^{-t}) signifies the average contribution of players in given group other than *j* (*i*). The regressions are made separately for all of the 20 periods of a game and the first period[10].

Table III. Motivations for second stage punishment

	All periods			First period		
	Full informatio n	Revenge only	No revenge	Full informatio n	Revenge only	No revenge
Constant (*β₀*)	0.061	0.242***	0.300***	-0.160	-0.026	0.148
	(0.042)	(0.046)	(0.085)	(0.368)	(0.281)	(0.602)
Others' average contribution (β_1)	-0.005*	-0.019***	-0.012***	-0.005	-0.008	-0.007
	(0.003)	(0.004)	(0.004)	(0.023)	(0.025)	(0.033)
Amount recipient contributed below average (β_2)	0.033***	0.048***	0.063***	0.104***	0.128***	0.064***
	(0.002)	(0.004)	(0.004)	(0.023)	(0.030)	(0.019)
Amount recipient contributed below average (β_3)	0.006***	0.009***	-0.002	0.027	0.039*	0.009
	(0.002)	(0.003)	(0.004)	(0.020)	(0.022)	(0.031)
Period (β_4)	-0.006***	-0.015***	-0.012***			
	(0.002)	(0.002)	(0.003)			

***1% significance level, **5% significance level, *10% significance level
Source: Authors

Our results concerning the first regression (see Table III) are mostly in line with those of the original experiment. Negative coefficient on β_1 (significant at *p < 0.05* in Revenge Only and No Revenge treatments and bordering on the significance within the Full Information treatment) indicates that the higher others' average contribution is, the less subject *i* sanctions her co-players. The coefficient is not that significant in the receiving of data only for first period of a game (i.e. *t=1*) in any period. The positive coefficient on β_2 (significant in all three treatments at *p < 0.01* for both the data for all periods and for separate period one) means that *i* punishes *j* more, the less *j* contributed to the group account relative to other players in the group. However, the results show at the

[10] For more technical details on regression functions see Denant-Boèmont et al. (2007). We present here only results of our analysis without explaining in detail the derivation of regression equations.

same time that the more *j* contributes to the group account relative to other co-players the more he gets punished, as it indicates a positive coefficient on β_3 the significance within Full Information and Revenge Only treatments for data for all periods. This finding is in contrary to results of original experiment where β_3 was positive (which seems logical). In case of our subjects it holds that the more one deviates from the average of a group (no matter in which sense) the more he gets punished.

Equation 6 refers to the motivations of sanction behavior in the third stage (i.e. second stage of punishment). In addition to the punishment of low contributors (as it was encapsulated by equation 5) it records the possibility of sanction enforcement and counter-punishment.

$$p_i^{j3t} = \beta_0 + \beta_1 p_j^{i2t} + \beta_2\left\{\left(\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t}\right)/2\right\} + \beta_3 c_{-i}^{-t} + \beta_4 max\left\{0, \sum_{k\neq i} p_j^{k,2t} - \left(\left(\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t}\right)/2\right)\right\} + \beta_5 max\left\{0, \left(\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t}\right)/2 - \sum_{k\neq i} p_j^{k,2t}\right\} + \beta_6 max\left\{0, c_{-j}^{-t} - c_j^t\right\} + \beta_7 max\left\{c_j^t - c_{-j}^{-t}\right\} + \beta_8 t \qquad (6)$$

p_i^{j3t} represents the punishment points assigned by subject *i* to *j* in the third stage of period *t*. p_j^{i2t} are punishment points assigned in second stage by player *j* to player *i*, $\left(\sum_{m\neq j}\sum_{k\neq i,j} p_m^{k2t}\right)/2$ is average number of punishment points assigned to players other than *i* and *j* in second stage, $\sum_{k\neq i} p_j^{k,2t}$ is total number of punishment points assigned by *j* to other individuals than *i*. The variables do comprise the average punishment of third parties and *j*'s deviation from this average are not included in the analysis for Revenge Only treatment, because the individuals don't have information to hand.

Table IV. Motivations for third stage punishment

	All periods			First period		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant (β_0)	0.187***	0.141***	0.230***	0.211	-0.127	0.917*
	(0.056)	(0.047)	(0.063)	(0.238)	(0.136)	(0.535)
Points j assigned to i in 2nd stage (β_1)	-0.008	0.052***	-0.007	-0.023	0.004	0.046
	(0.020)	(0.019)	(0.014)	(0.055)	(0.050)	(0.085)

Others' average punishment in 2nd stage (β_2)	-0.031		-0.041**	-0.069		-0.128
	(0.034)		(0.017)	(0.086)		(0.098)
Others' average contribution (β_3)	-0.010***	-0.006	-0.012***	-0.010	0.023*	-0.051*
	(0.003)	(0.004)	(0.003)	(0.014)	(0.012)	(0.029)
Positive deviation of recipient from average punishment in 2nd stage (β_4)	0.004		0.032	0.031		0.433
	(0.072)		(0.039)	(0.155)		(0.394)
Negative deviation of recipient from average punishment in 2nd stage (β_5)	0.017		0.054**	0.032		0.178
	(0.038)		(0.021)	(0.112)		(0.107)
Amount recipient contributed below the average (β_6)	0.033***	0.011***	0.004	0.009	0.007	-0.003
	(0.003)	(0.004)	(0.003)	(0.013)	(0.015)	(0.016)
Amount recipient contributed above the average (β_7)	0.003	-0.005	0.004	-0.004	-0.005	0.018
	(0.003)	(0.003)	(0.003)	(0.012)	(0.011)	(0.026)
Period (β_8)	-0.010***		-0.003			
	(0.003)		(0.002)			

Source: Authors

According to Denant-Boèmont et al. (2007) there are several motivations of the third stage punishment: individuals may wait until the third stage to punish low contributors, they may enforce sanctions of a first stage or counter-punish. Table IV records estimates from the regression. A significantly (at $p < 0.01$) positive coefficient on β_1 in Revenge Only treatment indicates the existence of counter-punishment; the more subject *j* punished *i* in second stage the more *i* punishes *j* in third stage. This phenomena isn't present in the case of Full Information and No Revenge treatments where β_1 is negative but not significantly. Coefficient β_3 is negative and significant in No Revenge and Full Information treatments which demonstrates the same tendency for third stage punishments as for second stage sanctions. Positive coefficient on β_5 indicating the existence of sanction enforcement is significant only in No Revenge treatment. (This means that the fewer points *j* assigns relative to the average punishment of third parties in second stage, the more *i* sanctions *j* in next stage.) Positive coefficient on β_6 (significant in Full Information and Revenge Only treatments) means that low contributors are sanctioned even in third stage. In accordance with the original experiment, our data reveals that in the second punishment stage sanction enforcement as well as counter-punishment and (delayed) punishment of low contributors occur.

Equations 7 and 8 encapsulate the effects of sanctions on individual contributions (equation 7) and sanction behavior (equation 8).

$$c_i^{t+1} - c_i^t = \beta_0 + \beta_1 \sum_k p_k^{i2t} + \beta_2 \sum_k p_k^{i3t} + \beta_3 (c_i^t - c_{-i}^{-t}) \qquad (7)$$

Dependent variable $c_i^{t+1} - c_i^t$ represents the difference in contribution levels of individual *i* between period *t* and *t+1*. $\sum_k p_k^{i2t}$ is the total number of punishment points assigned to player *i* in the second stage of period *t*, whilst $\sum_k p_k^{i3t}$ is a sum of punishment points assigned to *i* in the third stage. The last variable represents the deviation of *i*'s contribution from the others' average contribution.

Table V. Effect of sanctions on change in contribution

	Low contributors (all periods)			High contributors (all periods)		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant (β_0)	0.000	0.145	-1.522***	-0.223	0.939*	-1.256**
	(0.000)	(0.410)	(0.505)	(0.530)	(0.518)	(0.519)
Points received in second stage of period *t* (β_1)	0.500***	0.711**	0.570**	1.291	0.594	0.038
	(0.000)	(0.284)	(0.264)	(3.066)	(0.832)	(1.393)
Points received in third stage of period *t* (β_2)	2.000***	-0.507	0.976**	-0.975	0.603	-2.782
	(0.000)	(0.363)	(0.474)	(3.307)	(1.001)	(2.257)
Deviation from others' average contribution in period *t* (β_3)	-1.500***	-0.137	-0.351***	-0.510***	-0.708***	-0.219**
	(0.000)	(0.095)	(0.090)	(0.087)	(0.080)	(0.091)

Source: Authors

In accordance with Denant-Boemont's et al. (2007) model we conducted a separate analysis for high contributors (those who contribute more than the average in given period) and low contributors (who contribute below group average). The estimates are recorded in Table V. Positive coefficient on β_1 for low contributors (significant in all three treatments) indicates that the more punishment points subjects obtain in period *t* (in which they contributed below the average), the more they raise their contributions in period *t+1* (in relation to *t*). However, this is not the case for high contributors for which β_1 is not significant in any treatment. Coefficient on β_2 is ambiguous in sign and shows no general behavioral pattern for either high or low contributors. Negative coefficient on β_3 (significant for low contributors in the No Revenge and Full Information treatments and for high contributors in all three treatments at $p < 0.05$) indicates the existence of regression to the mean in contributions (independent of the number of sanctions received): this means that *"the higher one's contribution relative to the average, the stronger the tendency is to lower it in the following period"* (Denant-Boèmont et al., 2007). Observed reactions in contribution behavior on received sanctions were again consistent with those of the original experiment.

The last equation (9) records changes in sanction behavior depending on received punishments.

$$\sum_k p_i^{k,2,t+1} - \sum_k p_i^{k,2,t} = \beta_0 + \beta_1 \sum_k p_k^{i3t} + \beta_2\left(\sum_k p_i^{k2t} - \overline{\sum_k p_j^{2t}}\right) \qquad (8)$$

Dependency variable $\sum_k p_i^{k,2,t+1}$ represents the difference in total number of punishment points assigned by individual *i* in the second stage between period *t* and *t+1*. Expression $\sum_k p_i^{k2t} - \overline{\sum_k p_j^{2t}}$ means deviation of *i*'s sum of punishment points assigned from average number of points assigned by group players in second stage of period *t*. (This variable is not known by subjects in Revenge Only and thus is not included in analysis of this treatment.) The regressions are conducted separately for low and high punishers (relative to the average number of punishment points distributed in a group in given period).

Table VI. Effect of received sanctions on punishment in following period

	Low punishers (all periods)			High punishers (all periods)		
	Full information	Revenge only	No revenge	Full information	Revenge only	No revenge
Constant (β_0)	-0.038	0.013	0.158	0.268	-1.308***	0.107
	(0.067)	(0.051)	(0.142)	(0.314)	(0.275)	(0.314)
Points received in 3rd stage of period *t* (β_1)	0.074*	0.242***	0.016	-0.045	0.319	-0.426
	(0.042)	(0.085)	(0.111)	(0.266)	(0.236)	(0.642)
Deviation from average punishment in 2nd stage of period *t* (β_2)	-0.120		-0.229	-1.768***		-1.318***
	(0,123)		(0.156)	(0.341)		(0.241)

Source: Authors

Our data reveals no general pattern concerning the coefficient β_1. In the case of low punishers, the coefficient is positive but significant (at $p < 0.01$) only for data obtained under Revenge Only treatment (and bordering significant in Full Information). As in Revenge Only treatment sanctions assigned in third stage are interpreted only as counter-punishment, this would mean that the more a subject gets (counter) punished in third stage of period *t*, the more he sanctions in second stage of following period. This phenomenon is contrary to the main findings of experiments studying the effects of counter-punishment, concluding that the possibility to avenge sanctions leads to lower willingness to engage in sanctioning (e.g. Nikiforakis, 2008). While for low punishers the negative coefficient on β_2 is not significant in any treatment, it is significant (at $p < 0.01$) in both Full Information and No Revenge treatments for high punishers. This means that the more individuals punish in excess of the average number of punishment points in the third stage of given period, the less they sanction in next period. This finding is in accordance with results of the original experiment.

The data acquired confirmed our hypothesis related to different matching types. Subjects contributed and punished less under the stranger matching treatment than under partner matching used in the original experiment. Another strong difference in relation to the original experiment was that in our experiment the average contributions tended to zero, i.e. to the theoretically predicted equilibrium.

Concerning impacts of counter-punishment and sanction enforcement on contribution behavior and on individual welfare, as well as the sanction behavior of subjects engaging in sanctions, our results were mostly in accordance with those of the original experiment. We observed only two important differences included in our results. 1) Our subjects punished the more not only the less their co-players contributed relative to the average of a group, but also the more their colleagues contributed relative to the average. One would say that these players punished every conduct which was not average. 2) The second difference was *"curios"* reactions in sanction behavior of players in Revenge Only treatment in response to received counter-punishments. One would say that the threat of counter-punishment wasn't strong enough to influence the willingness to engage in first-stage sanctioning (or influenced it in inverse sense). However, contributions and individual earnings under Revenge Only treatment were the lowest observed among all treatments which would support the assumption of credibility of counter-punishment threat (in accordance with Nikiforakis, 2008).

6. Limits of concept of decentralized punishment in VCM and issues for further research

The following part of the paper introduces limits and potential issues for further research in area of decentralized punishment. It is partly inspired by a recent study dealing with this problem, published by Guala (2012).

The extent to which experimental data may be extrapolated in order to explain economic reality is referred to in the **external validity of experimental evidence**. The related question is whether (and to which extent) behavior of experimental subjects within a simplistic model situation (“artificially” created by the experimenters) represents the real individual behavior outside laboratory. This is the source of the primary critics of the experimental method that has accompanied it from the very beginning. Guala (2012) in his article refers to “wide” reading of experimental evidence. (The “narrow” reading of it is related to the concept of internal validity of experimental results.) According to him, “*the wide interpretation can only be tested using a combination of laboratory data and evidence about cooperation in the wild.”* Combining laboratory results with field data could solve the problem of external validity and respond adequately to related critic.

6.1 Problem of non-occurrence of costly punishments “in the wild”

Guala (2012) criticizes on the non-existence of studies investigating related behavior in a natural setting. However, there are difficulties in obtaining field evidence in the area of decentralized punishment. At first, as Gächter (2012) states, “*in equilibrium punishment will be rare and therefore may be hard to observe in the field”*. If the threat of punishment is strong enough, potential free-riders will be discouraged from defecting, they will cooperate and, as an effect, there will be no punishment carried out. In other words, the absence of frequent punishment indicates its effectiveness at fostering cooperation (Johnson, 2012, Gintis and Fehr, 2012). Gintis and Fehr provide an example “from the wild” concerning drivers: that while most of them receive several traffic citations during their lives, many drivers adjust their driving in order to prevent citations. Johnson stresses (in reaction to Guala, 2012) that at least one natural field experiment provides evidence on influence of costly punishment on cooperation. He presents a field study of voter turnout by Gerber at al. (2008) indicating that the mere possibility of costly punishment increased considerably the co-operative voter turnout (for more see Johnson, 2012). Another natural field experiment was carried out by Balafoutas and Nikiforakis (2011). The goal of the experiment was to examine whether civilians punish normal violators (presented by authors theirselves). The rate of altruistic punishment was low overall.

Despite these difficulties in obtaining data “in the wild”, Gächter (2012) emphasizes that “*behavioral logic uncovered by lab experiments is not that fundamentally different from the behavioral logic of cooperation in the field*”. There exist several limitations related to laboratory results (due mainly to the simplification of a studied situation) which have to be taken account when extrapolating to real situations. In the laboratory, subjects are forced to interact with others whilst having minimal options available (e.g. to contribute vs. to keep money, to punish vs. not to punish) and having little control over the information flow (Casari, 2012). In the field people may employ multiple strategies. However, as stated by Nikiforakis (2012), “*field data will prove insufficient in some cases to explain the determinants of cooperation by itself*”. This is due to difficult (or impossible) measuring of the relevant variables and limited control over the environment by an experimenter. This is where laboratory results may “fill the gap” and in this logic, field and experimental data complement each other.

6.2 Problem of quantifying punisher’s cost

A costly punishment means that the act of punishment brings a cost to the sanctioning subject. In experiments this cost is usually presented by some material loss (income reduction). However, in reality, sanctioning of peers may cause more complex costs than only material ones. These costs shouldn’t be overlooked when searching for evidence in the wild. Adams and Mullen (2012) mention for example that decreased social status and psychological well-being which both represent social and psychological costs experienced by the punisher. Van den Berg et al. (2012) argue that in real interactions, costs of punishment do not correspond to direct payments or payoff deductions, but they arise from the repercussions that punishment has on social networks and of the future interactions among subjects. Guala (2012) states that in small societies studied by anthropologists, economic cooperation is usually supported by lo-cost or no-cost mechanisms. These mechanisms such as gossip, verbal criticism, ostracism, public ridicule etc. are much more current (and effective) than costly punishment. However, if the term “cost” is extended by other than material types of cost, even these mechanisms may be very costly (for detailed discussion see van den Berg et al., 2012). In this sense, laboratory experiments should focus more on the non-direct (non-material) costs of punishment.

6.3 Problem of the limited effectiveness of decentralized punishment

The last group of limits of the concept of decentralized punishment we mention here is related to its limited effectiveness. As it was already stressed within this paper, decentralized punishment has proven to be an effective mechanism at enhancing cooperation in the public goods game. However, this effectiveness is conditioned by several facts.

The positive effect of punishment is maximized when players receive a unique opportunity to impose sanctions on free-riders while those who are not aware of who sanctioned them. This finding is in accordance with our results under the Baseline treatment (compared to three others). However, such (artificial) limitation doesn’t enable the subject to engage in various punishment strategies like retaliation, sanction enforcement etc. The experimental setting also prevents the prospect of feuds. This moves the laboratory situation away from the real one in which people have many strategies available and try to optimize them. Experiments enabling multiple punishment opportunities, counter-punishment or feuds show that the former positive effect is often outweighed by these strategic considerations.

Decentralized punishment is effective only under a specific cost-impact ratio. Experimental results (see e.g. Egas and Riedl, 2004) have shown that the subjects, in deciding whether to engage in altruistic punishment or not take into account the costs and effects of their actions.

Decentralized punishment is more effective if combined with other cooperative-enhancing mechanisms. As Guala (2012) states, “*costly punishment alone does not seem to be an efficient solution to social dilemmas in the laboratory*”. It has been shown (e.g. in Ostrom et al. 1992 or Bochet et al., 2006), decentralized punishment is more effective if combined with the possibility of (verbal) communication.

There are further parameters that need to be taken account when studying the effects of decentralized punishment. Ferguson and Corr (2012) highlight other evolutionary parameters which may influence the willingness to engage in sanctioning: resource holding, status and sex and legitimacy of free-riding. The authors stress that the first three parameters may result in different levels of aggression and that in the real world there may exist legitimate reasons for non-cooperation, such as illness or other.

7. Bibliography

Adams, G., S., Mullen, E., 2012. The social and psychological costs of punishing. In Guala, F., 2012. Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. *Behavioral and Brain Sciences* 35, 15–16.

Andreoni, J., 1988. Why Free Ride? – Strategies and Learning in Public Goods Experiments. *Journal of Public Economics* 37, 291-304.

Balafoutas, L. & Nikiforakis, N. (2011) Norm enforcement in the city: A natural field experiment. *Mimeo*.

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

Casari, M., 2012. Weak reciprocity alone cannot explain peer punishment. In Guala, F., 2012. Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. *Behavioral and Brain Sciences* 35, 21–22.

Denant-Boemont, L., Masclet, D., Noussair, C., 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 33, 145–167.

Dawes, R. M., 1975. Formal Models of Dilemmas in Social Decision Making. In *Human Jugement and Decision Processes*, ed. Martin F. Kaplan and Steven Schwartz. New York: Academic.

Dawes, R. M., 1980. Social Dilemmas. *Annual Review of Psychology* 31 : 169-93.

Egas, M. & Riedl, A., 2008. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences* 275(1637):871–78.

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.

Ferguson, E., Corr, P., 2012. Blood, sex, personality, power, and altruism: Factors influencing the validity of strong reciprocity. In Guala, F., 2012.

Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. *Behavioral and Brain Sciences* 35, 25–26.

Fischbacher, U., 2007. Z-tree: zurich toolbox for readymade economic experiments. *Experimental Economics* 10, 171–178.

Guala, F., 2012. Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. Behavioral and Brain Sciences 35, 1–59.

Gerber, A. S., Green, D. P. & Larimer, C. W. (2008) Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review* 102(1):33–48.

Gintis, H., Fehr, E., 2012. The social structure of cooperation and punishment. In Guala, F., 2012. Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. *Behavioral and Brain Sciences* 35, 28–29.

Hobbes, T., 1960. *Leviathan*. Oxford: Basil Blackwell.

Johnson, T., 2012. The strategic logic of costly punishment necessitates natural field experiments, and at least one such experiment exists. In Guala, F., 2012. Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. *Behavioral and Brain Sciences* 35, 31–32.

Ledyard, J., 1995. Public Goods: A Survey of Experimental Research. In *Handbook of Experimental Economics*, ed. by Kagel, J. and Roth, A., Princeton University Press.

Nikiforakis, N., 2008. Punishment and counter-punishment in public good games: can we really govern ourselves? *Journal of Public Economics* 92, 91–112.

Nikiforakis, N., 2012. Altruistic punishment: What field data can (and cannot) demonstrate. In Guala, F., 2012. Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. *Behavioral and Brain Sciences* 35, 32–33.

Nikiforakis, N., Normann, H.T., 2008. A Comparative statics analysis of punishment in public-good experiments. *Experimental Economics* 11, 358–369.

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

Van den Berg, P., Molleman, L., Weissing, F., J., 2012. The social costs of punishment. In Guala, F., 2012. Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate. *Behavioral and Brain Sciences* 35, 42–43.