

UNIVERSITY OF  
BIRMINGHAM

Department of Economics

# The Behavioural Consequences of Unfair Punishment

Department of Economics Discussion Paper 10-34

**Michalis Drouvelis**

# The behavioural consequences of unfair punishment

Michalis Drouvelis, University of Birmingham \*

November 2010

## Abstract

Experimental evidence from public good games with punishment suggests that punishment works when subjects assign it fairly by sanctioning non-cooperators. This paper reports an experiment in which punishment is assigned unfairly in the sense that it is not linked to individual behaviour and is meted out to all group members (irrespective of their prior behaviour). We test whether unfair punishment generates different contribution and punishment behaviour relative to the standard punishment game. Our findings suggest different dynamics of average contributions in the presence of unfair punishment relative to the standard punishment game. Contribution levels are significantly different only when subjects have obtained experience from both games. We also find that, although the assignment of punishment is unaffected after the experience of an environment with unfair punishment, a history of unfair punishment makes a difference regarding reactions to alleviation, reward and punishment received.

*Keywords:* Reciprocity; Unfair punishment; Public good experiments

*JEL codes:* C92, D01, H41.

**Acknowledgements:** I owe thanks to Ananish Chaudhuri, James Cox, Robin Cubitt, Eamonn Ferguson, Simon Gächter, and John Hey for helpful suggestions and discussions. The paper has also benefited from comments by seminar participants at the University of Nottingham and the ESA Meetings in Innsbruck (2009) and in Copenhagen (2010). Financial support from the ESRC (PTA-030-2005-00608) and the University of Nottingham is gratefully acknowledged.

---

\* Department of Economics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom. Email: [m.drouvelis@bham.ac.uk](mailto:m.drouvelis@bham.ac.uk)

## 1. Introduction

A central theme in the behavioural sciences is the examination of the ability of punishment to sustain high cooperation rates and to regulate behaviour in social dilemma games. Laboratory experiments on public goods have shown that punishment is a successful norm enforcement mechanism that fosters cooperation (for excellent reviews, see Chaudhuri, 2007; Gächter and Herrmann, 2009). Yet, the cooperation enhancing effect of punishment has been found to be sensitive to a number of factors such as the cost and effectiveness of punishment (see Anderson and Putterman, 2006; Carpenter, 2007, Egas and Riedl, 2008; Nikiforakis and Normann, 2008), second-round punishment opportunities (see Cinyabuguma et al., 2006; Denant-Boemont et al., 2007; Nikiforakis, 2008), and antisocial punishment (see Herrmann et al., 2008; Gächter and Herrmann, 2009, 2010).

The aim of this paper is to explore experimentally the behavioural effects triggered by the assignment of unfair punishment and its impact on people's willingness to punish. Experimental research suggests that one condition for punishment to work is that individuals assign it fairly by sanctioning non-cooperators only (e.g., Fehr and Gächter, 2000; Fehr and Gächter, 2002; Masclet et al., 2003; Herrmann et al., 2008). These findings are interpreted as evidence that individuals punish non-cooperators because they violate a norm of, or a predisposition towards, reciprocity. In a recent experiment, Falk et al. (2005) suggest the violation of fairness principles is a major driving force of sanctions. Motivated by these observations, we extend this line of investigation by exploring whether and if so, how an unfair environment that violates such principles may affect subjects' willingness to use punishment.

Previous evidence from bargaining and public good games (e.g., Henrich et al., 2004 and Herrmann et al., 2008, respectively) indicates that people's everyday experiences are reflected in their observed experimental behaviour. However, in these experiments, experiences have been shaped exogenously, outside of the lab. Our experiment explores how a history of an unfair environment experienced *in the lab* affects individuals' expectations of how punishment might work. To generate an environment where punishment is assigned unfairly, we propose a variant of the standard punishment game (Fehr and Gächter, 2000), which we refer to as the default punishment game. In this game, members of a group participate in a two-stage game. In the first stage, they are engaged in a standard linear public goods game, in which

they have to decide how much of their initial endowment they are willing to contribute to the public good (see Ledyard, 1995). At the beginning of the second stage, we introduce a new element: all group members are exogenously sanctioned by having an automatic penalty imposed on them – the default punishment.<sup>1</sup> This implies that all group members unconditionally receive a decrease in their monetary income, irrespective of their first stage behaviour. During the second stage, after contribution decisions have been anonymously revealed, subjects are given the opportunity to alleviate the exogenous default punishment of others at some cost to themselves. As the default punishment is unrelated to first stage behaviour and is not tailored to fit individual misbehaviour, we assume this procedure to be unfair.<sup>2</sup> Our punishment scheme has the interesting feature that punishment does not depend on the individual behaviour and thus, automatic penalty is unjust in the sense that it cannot identify individual defectors or cooperators. This makes automatic penalty not being any more social since it is targeted at all group members.<sup>3</sup>

A noteworthy aspect of our default punishment game is that it resembles the reward game previously studied in the experimental literature (see Sefton et al., 2007; Rand et al., 2009; Sutter et al., 2010). Specifically, helping behaviour (i.e. reward via alleviation of the automatic penalty) in the default punishment game is tantamount to rewarding in the reward game. Therefore, the main substantive difference between the default punishment game and the reward game is the presence of the automatic penalty in the former but not in the latter game. This implies that behaviour in the default punishment game can be affected either by the existence of the default

---

<sup>1</sup> This sort of punishment has parallels in real world situations, such as environmental or natural disasters. Take for example an earthquake or a hurricane, which are exogenously implemented and hit all members of a community, regardless of their prior behaviour. Arguably, these circumstances can be perceived as being unfair, since this sort of “punishment” is not linked to individual behaviour and reduces social welfare.

<sup>2</sup> To investigate the extent to which the automatic penalty was indeed perceived as unfair by subjects, in a debriefing post-experimental questionnaire, we asked subjects to rate on a 7-point scale (1 = “Strongly disagree”, ..., 7 = “Strongly agree”) the degree with which they agreed with the statement: “I perceive the automatic penalty unfair”. According to their answers, 55.3% of subjects maintain that the exogenously imposed automatic penalty is unfair, while approximately half of this percentage (28.9%) disagree with this statement. The remaining percentage (15.8%) had no opinion.

<sup>3</sup> The effects of blind punishment on public good provision have been recently studied by Fatas et al. (2010). In particular, they use a punishment mechanism based on random exclusions. Yet, their design is distinctive to ours in two main respects concerning the rules governing the implementation of punishment. First, their punishment scheme is still social and pursues a collective goal. Good teams were never punished under their sanctioning system. Second, unlike our experiment where subjects are given the opportunity to alleviate the automatic penalty, in their design there is no second stage in which individuals can correct the unfairness of the blind punishment. Their findings suggest that random exclusions generate more public good provision (compared to a standard public good game without punishment) and promote efficiency in a significant way.

punishment which is allocated to each group member *or* by the opportunity given to group members to reciprocate positively by helping each other. In order to disentangle these effects, we also explore subjects' behaviour in the reward game where they can only reciprocate positively via rewards. Note that, in the reward game, there is no automatic penalty, but the reward dimension remains unchanged relative to the default punishment game. Thus, investigation of the default punishment game automatically extends to the investigation of the reward game as well.

In sum, this paper addresses two main research questions. First, does behaviour in the default punishment game differ from behaviour in the standard punishment game? Second, how do subjects behave in the standard punishment game when they have previously experienced the default punishment game? We are specifically interested in these two research questions as they directly test the behavioural effects of unfair punishment on contribution and punishment attitudes. Our sanctioning scheme renders punishment not being social any more and thus, aims to complement findings from earlier studies which suggest antisocial punishment to remove the cooperation-enhancing effect of punishment (Herrmann et al., 2008).

Our findings suggest that the time profile of average contribution levels under the default punishment game is different relative to the standard punishment game. Contribution levels are on average similar between the three games, but are significantly different when same subjects participate both in the default and the standard punishment game. We find that assigned alleviation and reward are not significantly different and that the assignment of punishment is unaffected by the previous experience of unfair punishment. Yet, a history of unfair punishment makes a difference regarding reactions to alleviation, reward and punishment received.

The remainder of the paper is organized as follows. Section 2 presents the design and the procedures of the experiment. Section 3 reports the results and Section 4 concludes.

## **2. Experimental Design and Procedures**

### *2.1 Experimental design*

Our experimental design consists of three conditions: the “default punishment condition” (D-condition), the “standard punishment condition” (S-condition) and the “reward condition” (R-condition). In all three conditions, subjects were involved in a

two-stage game. The first stage of the game was common to all three conditions. Yet, regarding the second stage, some of its features were common, while others varied.

To begin with, the first stage involves a voluntary contributions mechanism game with linear payoffs. For this stage, subjects, being randomly assigned to a four-person group, are privately endowed with 20 tokens each and have to decide how many of these to keep for themselves and how many to contribute to a public good (described to subjects as “project”). For each token kept, each subject earns 1 Money Unit; whereas, for each token contributed the return is equal to 0.5 Money Units, resulting in a total of 2 Money Units for the whole group. Subjects make their decisions in private and at the end of the first stage they are informed about the sum of the contributions to the public good made by the whole group and about their own first stage income.

After the first stage has finished, a second stage begins. In each of our three conditions, the common characteristics of the second stage are as follows. At the beginning of the second stage, subjects can see the profile of contributions of the other three group members. However, no subject could identify the particular contribution of any other subject, since the order of contributions shown in each screenshot randomly changed from period to period. More specifically, each subject’s own contribution is always listed in the first column of his computer screen and the remaining three subjects’ contributions are randomly listed in the second, third or fourth column, respectively. Therefore, subject-specific reputations cannot develop across periods, since subject  $i$  does not have the information to construct a link between individual contributions of subject  $j$  across periods. After subjects become aware of the whole vector of individual contributions in their group from the first stage, each subject could assign adjustment points to other group members. Since subject-specific reputations cannot build up, the possibility that player  $i$  assigns adjustment points to player  $j$  in period  $t$  for contribution decisions made in a previous period from  $t$  is ruled out. Subjects could assign between 0 and 2 adjustment points to each other group member. Assignment of adjustment points is always costly, with each adjustment point having a cost of one Money Unit per token. In addition, assigning an adjustment point has an impact on the payoff of the receiver. The absolute magnitude of this impact is equal to three in all conditions. How these adjustment points can be used depends on each condition, but in any case, subjects are given a message/suggestion about how they might use their adjustment points.

Finally, at the end of the second stage, subjects were informed about their own cost of assigning adjustment points, the total number of adjustment points assigned to them, and their earnings. No information about the number of adjustment points received by each group member was available.

Depending on the condition, there are three features that vary: (i) the presence or absence of an automatic penalty; (ii) the sign of the impact of an adjustment point on the assignee; and (iii) the message regarding the use of the adjustment points.

In particular, under the D-condition, all subjects incurred an automatic penalty irrespective of their first stage contributions. We refer to this penalty as default punishment and assume it to be unfair since it was unrelated to subjects' past behaviour. In this stage, the message subjects were given was that they can reward other group members by alleviating their automatic penalty, which was costly for the alleviator, but beneficial for the person receiving the adjustment points. Note that if a subject received more alleviation points than the automatic penalty, their income did not increase by this extra amount. In our experiment, the automatic penalty was set equal to 10 Money Units. We did so for two reasons. First, complete alleviation of the automatic penalty was possible only if the majority of the group members decided to assign adjustment points. Recall that each group member can assign up to 2 adjustment points, with each point decreasing the automatic penalty by 3 Money Units. This essentially implies that the automatic penalty is fully alleviated only if two or more group members assign the total amount of points they control. Second, we did not want to create a situation where subjects would be very likely to end up with substantial losses due to the automatic penalty at the end of the experiment. In the case that this could occur, subjects would receive a large lump sum payment to cover possible losses in the D-condition. However, since the lump sum payment has to be kept constant across conditions, an unnecessarily high level of it might affect behaviour, which we wanted to avoid.

Contrary to the D-condition, the S-condition does not include any automatic penalty. Subjects are now given the opportunity to decrease other group members' income. In other words, the message sent to subjects in this condition was that they can penalize each other group member. Assignment of adjustment points is costly both for the punisher and the recipient of the punishment.

In order to assess the extent to which behaviour differs in a situation where punishment is assigned unfairly and exogenously and a situation in which it is not

assigned in such a way, we compare contribution behaviour between the D- and the S-condition. However, by making a comparison between these two conditions, it turns out that such a difference could be due to either the sign of the impact of the adjustment points or the automatic penalty components. It is therefore crucial to include a condition which allows us to disentangle these two effects. This is done by the inclusion of the R-condition, in which subjects are given the opportunity to increase the earnings of each other group member. In other words, under the R-condition, subjects were given the message that they can reward their counterparts. As in the D-condition, assignment of adjustment points is costly for the donor, but benefits its recipient. The cost-to-impact ratio was identical to the one used in the D-condition; namely, 1:3. Clearly, the inclusion of the R-condition allows us to identify the source of a potential difference between the S-condition and the D-condition. If it turns out that the R-condition is the same as the S-condition, the difference between the S-condition and the D-condition will be due to the automatic penalty. However, if it turns out that the R-condition is the same as the D-condition, the difference between the S-condition and the D-condition will be due to the sign of adjustment points. Table 1 presents the differences between our three conditions.

**Table 1.** Differences among conditions

	S-condition	D-condition	R-condition
Automatic penalty	No	Yes	No
Sign of impact of adjustment points	- 3	+ 3	+ 3
Message	“You can penalize the other group members.”	“You can reward the other group members. This can alleviate the automatic penalty of 10 Money Units.”	“You can reward the other group members.”

To answer our research questions, we implement a within-subjects design under which the same individual participates in a sequence, consisting of two conditions. Each condition has 10 periods. In total, we have three sequences: the DS sequence, in which the D-condition is followed by the S-condition; the SD sequence, in which the S-condition is followed by the D-condition; and the RS sequence, in which the R-condition is followed by the S-condition. Each of the three sequences described above was conducted twice, yielding a total of 6 sessions. We implemented a Partners' matching protocol meaning that the group composition remained the same across all 20 periods within a sequence. In the DS sequence 40 subjects participated, resulting in 10 independent observations; while in the SD and RS sequences 36 subjects participated separately, resulting in 9 independent observations per sequence. At the beginning of each sequence, subjects were informed that the session consists of two conditions in order to reduce the possibility for having wrong expectations about the nature of the experiment. However, they were not told what will happen in the second condition.

In sum, our experimental design provides direct answers to our research questions. First, it enables us to investigate the pure behavioural effects of each of the three conditions in terms of contributions. For instance, comparison of the D-condition of the DS sequence with the S-condition of the SD sequence allows us to assess whether there is any difference in contribution levels between these two environments as subjects experience them both for the first time. This is also the case for the comparison between the R-condition of the RS sequence with the S-condition of the SD sequence. Second, our design allows us to assess the robustness of both contribution and punishment behaviour after the experience of either the D-condition or the R-condition. Recall that we chose a cost-to-impact ratio equal to 1:3. We did so in the light of previous experimental findings on public good experiments with punishment (see Nikiforakis and Normann, 2008; Egas and Riedl, 2008) which demonstrate that these ratios can induce high and stable contribution levels. Based on this observation, we can test whether the effectiveness of the S-condition is still maintained after the experience of the D- and R-conditions. Thus, in order to investigate the robustness of behaviour in the S-condition of the SD sequence, we compare it with the S-condition of the DS sequence and the S-condition of the RS sequence.

## 2.2 Procedures

All sessions took place in April and May 2008 in the Centre for Decision Research and Experimental Economics (CeDEx) lab. Recruitment was conducted via the software ORSEE (Greiner, 2004) at the University of Nottingham using subjects from a university-wide pool of registered students. All conditions were computerized and programmed with the software z-Tree (Fischbacher, 2007). At the beginning of each sequence subjects received instructions for the first condition and at the end of it for the second condition.<sup>4</sup> All participants answered several test questions, concerning the calculation of payoffs for various hypothetical configurations of behaviour. None of the conditions proceeded until every subject had answered these questions correctly. At the end of a sequence, subjects were privately paid according to their accumulated earnings from all 20 periods, using an exchange rate of £0.015 per Money Unit. Average earnings per sequence were as follows: £10.45 for the DS sequence; £9.48 for the SD sequence; £11.65 for the RS sequence. Sessions lasted, on average, 75 minutes, with no session taking longer than 90 minutes.

## 3. Results

### 3.1 Contribution levels

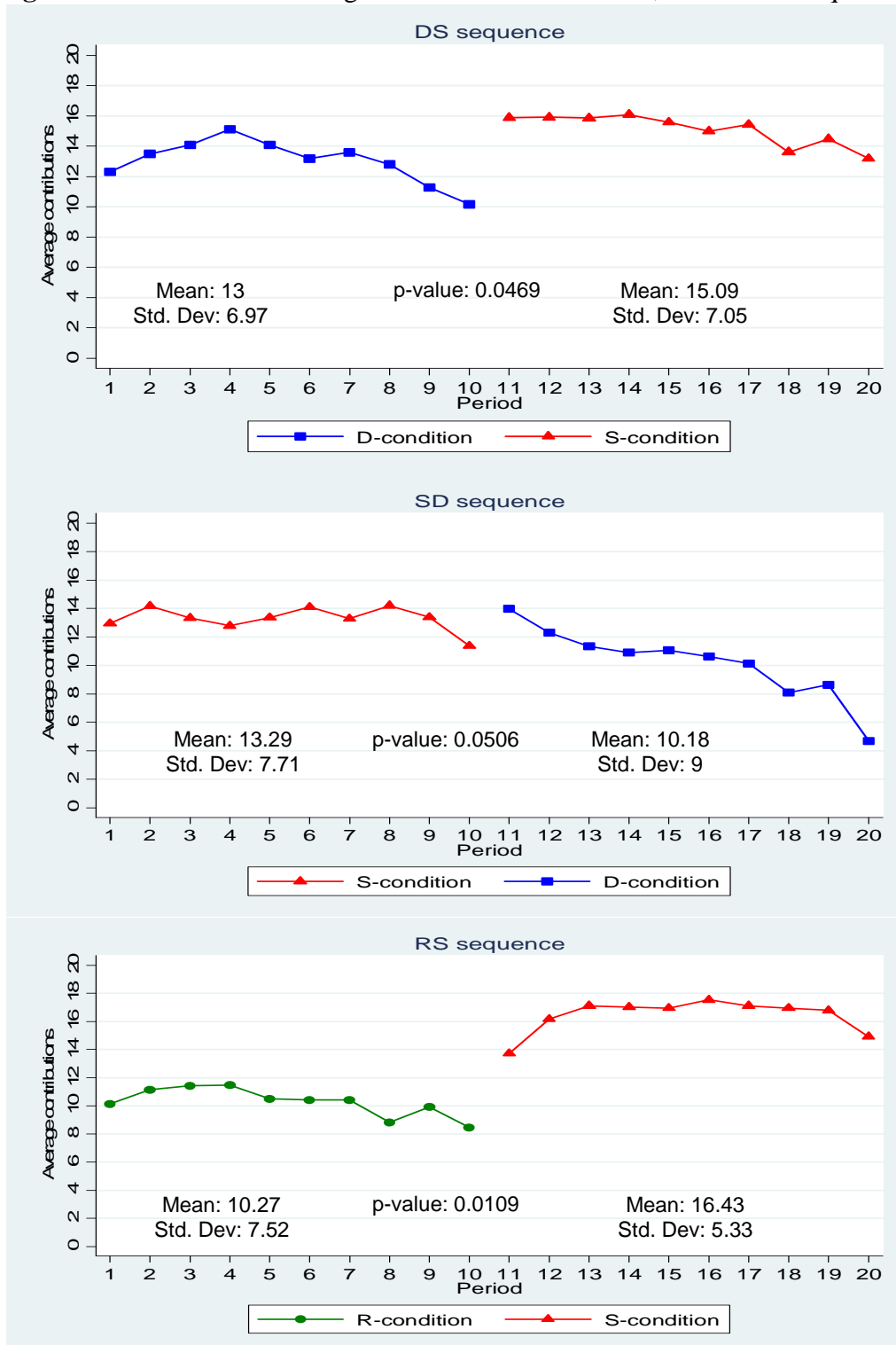
We begin our data analysis by looking at how contribution levels evolved in each of our three sequences. Data are presented as the amount of tokens contributed to the group account. Starting with the DS sequence and averaging across all ten periods, we find that subjects' mean contributions were 13 and 15.09 tokens for the D- and the S-conditions, respectively. Regarding the SD sequence, average contributions across all ten periods were 13.29 tokens for the S-condition and 10.18 tokens for the D-condition; while in the RS sequence, the corresponding mean contribution levels were 10.27 tokens for the R-condition and 16.43 tokens for the S-condition. The average contribution levels as a function of periods, for each sequence separately, are illustrated in Figure 1. In each panel, the mean contributions across periods for each condition, the corresponding standard deviation and the p-values from a Wilcoxon signed-rank test for within sequence comparisons are also shown. Since subjects did

---

<sup>4</sup> A copy of the instructions is available upon request from the author.

not change groups across all periods, each matching group is treated as an independent observation.

**Figure 1.** Time series of average contributions for the DS, SD and RS sequences



To assess the extent to which contributions differ between our three conditions, we contrast the average contributions for the following three comparisons: (i) the S-condition of the SD sequence versus the D-condition of the DS sequence, (ii) the R-condition of the RS sequence versus the S-condition of the SD sequence, and (iii) the R-condition of the RS sequence versus the D-condition of the DS sequence. Recall that if we find a difference between the S- and D-conditions, then the R-condition will help us understand such a difference. That is, if the R-condition is like the S-condition, the difference between the D- versus the S-conditions is due to the automatic penalty; whereas, if the R-condition is like the D-condition, the difference between the D- and S-conditions is due to the sign of the adjustment points.

To identify whether there are any differences between the pair-wise comparisons of our interest, we perform a Wilcoxon rank-sum test. We find that contributions are not statistically significantly different between any of the three pair-wise comparisons. More specifically, we find that comparing the S-condition (SD sequence) with the D-condition (DS sequence) yields a p-value of 0.775; comparing the R-condition (RS sequence) with the S-condition (SD sequence) yields a p-value of 0.2332; and comparing the R-condition (RS sequence) with the D-condition (DS sequence) yields a p-value of 0.1651.

Our econometric analysis also corroborates these findings, suggesting no significant difference in contribution levels relative to our three conditions. We estimate three OLS models,<sup>5</sup> in which the dependent variable is the contribution to the group account and the independent variable comprises four dummies, where each one corresponds to a block of two periods and an additional dummy variable called “condition” that captures the possible difference between our pair-wise condition comparisons. The dummy variable “condition” equals 1 for the first condition in each comparison. For instance, for the first comparison of Table 2, “condition” equals 1 for the S-condition of the SD sequence and 0 for the D-condition of the DS sequence. Our regression results are presented in Table 2. Robust standard errors are reported in parentheses.

---

<sup>5</sup> Since contribution levels are censored at 0 and 20, we also estimated Tobit regression models. The results obtained using this econometric specification is qualitatively similar as in the OLS regressions. The reason why we present the OLS coefficients hinges on the fact that they can be interpreted more easily.

**Table 2.** Condition comparisons in contribution levels

<i>Dependent variable: Contribution</i>			
	S-condition (SD sequence) vs. D-condition (DS sequence)	R-condition (RS sequence) vs. S-condition (SD sequence)	R-condition (RS sequence) vs. D-condition (DS sequence)
Periods 1&2	1.704 (1.171)	1.319 (1.066)	1.836 (1.164)
Periods 3&4	2.362** (0.895)	1.472* (0.775)	3.112*** (1.016)
Periods 5&6	2.178** (0.796)	1.319* (0.656)	2.138** (0.863)
Periods 7&8	1.954*** (0.573)	0.896* (0.426)	1.514** (0.672)
Condition	0.289 (2.378)	-3.025 (2.496)	-2.736 (2.015)
Constant	11.363*** (1.529)	12.290*** (2.141)	11.283*** (1.547)
Obs.	760	720	760

*Note: OLS estimates with robust standard errors (clustered on independent matching groups) presented in parentheses. For the first comparison, the dummy variable “condition” equals 1 for the S-condition in the SD sequence and 0 otherwise. For the second comparison, the dummy variable “condition” equals 1 for the R-condition in the RS sequence and 0 otherwise. For the third comparison, the dummy variable “condition” equals 1 for the R-condition in the RS sequence and 0 otherwise. Periods 9 & 10 are the baseline. \* denotes significance at the 10-percent level, \*\* denotes significance at the 5-percent level and \*\*\* denotes significance at the 1-percent level.*

Regression coefficients from Table 2 suggest that subjects do not contribute differently on average in any of the three conditions we examine, when subjects experience each of these conditions for the first time. In other words, the default punishment game does not yield significantly different contribution levels from the standard punishment game and the reward game.

However, observing Figure 1, we notice that the time profile of average contributions seems to differ between the three conditions. We check formally whether there are any period effects affecting the observed contribution patterns illustrated in each panel above. To do so, we run three OLS regression models in

which the dependent variable is the contribution to the group account. To control for period effects, we include four dummies, where each one corresponds to a block of two periods (as in Table 2).<sup>6</sup> The regression results from this analysis are given in Table 3. Robust standard errors are also presented in parentheses.

**Table 3.** Period effects in the D-, S- and R-conditions

<i>Dependent variable: Contribution</i>			
	D-condition (DS sequence)	S-condition (SD sequence)	R-condition (RS sequence)
Periods 1&2	2.175 (1.795)	1.181 (1.564)	1.458 (1.548)
Periods 3&4	3.875** (1.482)	0.681 (0.627)	2.264 (1.418)
Periods 5&6	2.913* (1.327)	1.361 (0.814)	1.278 (1.083)
Periods 7&8	2.488** (1.026)	1.361*** (0.406)	0.431 (0.746)
Constant	10.713*** (1.821)	12.375*** (2.336)	9.181*** (1.863)
Obs.	400	360	360

*Note: OLS estimates with robust standard errors (clustered on independent matching groups) presented in parentheses. Periods 9 & 10 are the baseline. \* denotes significance at the 10-percent level, \*\* denotes significance at the 5-percent level and \*\*\* denotes significance at the 1-percent level.*

Table 3 reveals that period effects are significant for the D-condition, as suggested by the first panel of Figure 1. Specifically, we observe that the contribution patterns for the D-condition follow a hump-shaped pattern. Contributions increase up to a certain level (first half of the game), but after that they start declining till the end of the game. In contrast, for both other conditions (S- and R-conditions), contribution levels are not strongly affected by periods and show a rather stable pattern across

<sup>6</sup> In our regression model, Periods 9 & 10 is the baseline category. Since in the D-condition subjects interact in an unfair environment, it is likely that at the beginning of this condition they are not sure what norms will prevail. This is the reason why we chose the last two periods as the comparison group as by that time subjects will have enough experience of other group member's behaviour.

time.<sup>7</sup> This last observation is in line with existing experimental evidence suggesting stable contribution levels over time at least for a cost-to-impact ratio of 1:3 in the presence of punishment opportunities (e.g., Egas and Riedl, 2008; Nikiforakis and Normann, 2008). Interestingly, when rewards are available, a cost-to-impact ratio of 1:3 also generates contribution stable patterns. This complements previous experimental evidence on reward games suggesting that a ratio of 1:1 is not able to sustain cooperation (e.g., Walker and Halloran, 2004; Sefton et al., 2007; Sutter et al., 2010).

Having documented strong period effects in the default punishment game, we further investigate whether these can affect the effectiveness of the S-condition. The reason why this might be the case is that unravelling of cooperation can be seen as an indicator of the perceived unfairness of the D-condition which may impact on the use of punishment in the S-condition. To assess whether the ability of the S-condition to sustain contributions can survive after subjects have experienced the D-condition and the R-condition, we compare: (i) contribution behaviour in the S-condition when it is played first versus contribution behaviour in the S-condition when the D-condition has preceded it, and (ii) contribution behaviour in the S-condition when it is played first versus contribution behaviour in the S-condition when the R-condition has preceded it as well.

A non-parametric Wilcoxon rank sum test reveals no significant differences from either comparison.<sup>8</sup> This finding is corroborated by our formal econometric analysis (see Appendix A, Table A.1). Following similar econometric methodology as previously, our regression results suggest that the ability of the punishment game to generate high contribution levels is not affected even after contribution levels collapse after the second half of the D-condition. Not surprisingly, a history of the R-condition, in which the contribution pattern is rather stable over time, fails to affect the ability of the punishment game to sustain contributions.

---

<sup>7</sup> This observation is also supported by fitting a quadratic function in the regression, with “Period” and “Period squared” as independent variables. We find that, for the D-condition, the coefficient for “Period” is positive and significant and the coefficient for “Period squared” is negative and significant, confirming the hump-shaped pattern of contributions. However, for the S- and R-conditions, these two coefficients are not statistically significant.

<sup>8</sup> The corresponding p-values are 0.3691 and 0.2004, respectively.

### 3.2 Within-subjects comparisons in contribution levels

This subsection explores whether subjects' willingness to contribute differs within a given sequence. This comparison is important as it takes into account the behaviour of the subjects who experienced two conditions and we can therefore identify whether there is any impact of a history of one condition on another. As indicated in Figure 1, performing a Wilcoxon signed-rank test, we find that significant differences on average contributions in all three sequences, with the S-conditions yielding always higher contribution levels. This evidence is corroborated by our formal econometric analysis presented in Table 4.

**Table 4.** Contribution differences within a given sequence

<i>Dependent variable: Contribution</i>			
	DS sequence	SD sequence	RS sequence
Periods 1&2	2.119 (1.513)	3.826*** (0.775)	0.278 (0.937)
Periods 3&4	3.006* (1.423)	2.569*** (0.751)	1.743** (0.673)
Periods 5&6	2.181 (1.325)	2.764*** (0.602)	1.333** (0.483)
Periods 7&8	1.588* (0.804)	1.903** (0.690)	0.806 (0.488)
Condition	-2.09** (0.835)	3.117** (1.066)	-6.158*** (1.395)
Constant	13.314*** (2.168)	7.963*** (2.151)	15.593*** (1.185)
Obs.	800	720	720

*Note: OLS estimates with robust standard errors (clustered on independent matching groups) presented in parentheses. For the first comparison, the dummy variable "condition" equals 1 for the D-condition in the DS sequence and 0 otherwise. For the second comparison, the dummy variable "condition" equals 1 for the S-condition in the SD sequence and 0 otherwise. For the third comparison, the dummy variable "condition" equals 1 for the R-condition in the RS sequence and 0 otherwise. Periods 9 & 10 are the baseline. \* denotes significance at the 10-percent level, \*\* denotes significance at the 5-percent level and \*\*\* denotes significance at the 1-percent level.*

Regression results from Table 4 suggest that contribution levels in the S-condition are always higher than those in either the D-condition or the R-condition when subjects experience both conditions. We also observe significant and strong period effects with respect to the SD sequence. It is worth mentioning that these period effects are due to the D-condition of this sequence. Specifically, running a similar regression as in Table 3, we notice that the coefficients of the block dummies are positive in decreasing order and statistically significant, indicating a clear downward trend in contribution levels for the D-condition. This observation implies that subjects react faster to the unfairness of the automatic penalty by decreasing their contributions, when they have already experienced the S-condition, which is arguably a fairer condition. However, when there is no previous experience of another condition and thus, no other means of comparison, cooperation rates decline only after the second half of the game.

In sum, our analysis of contribution levels suggest that, in the absence of any previous history, unfair punishment causes different contribution patterns, but, on average, contribution levels are not significantly different between the three conditions. Yet, different contribution levels are found when subjects have obtained experience by participating in one condition. It turns out that in all three sequences, the S-condition generates higher contribution levels compared to either the D- or the R-condition.

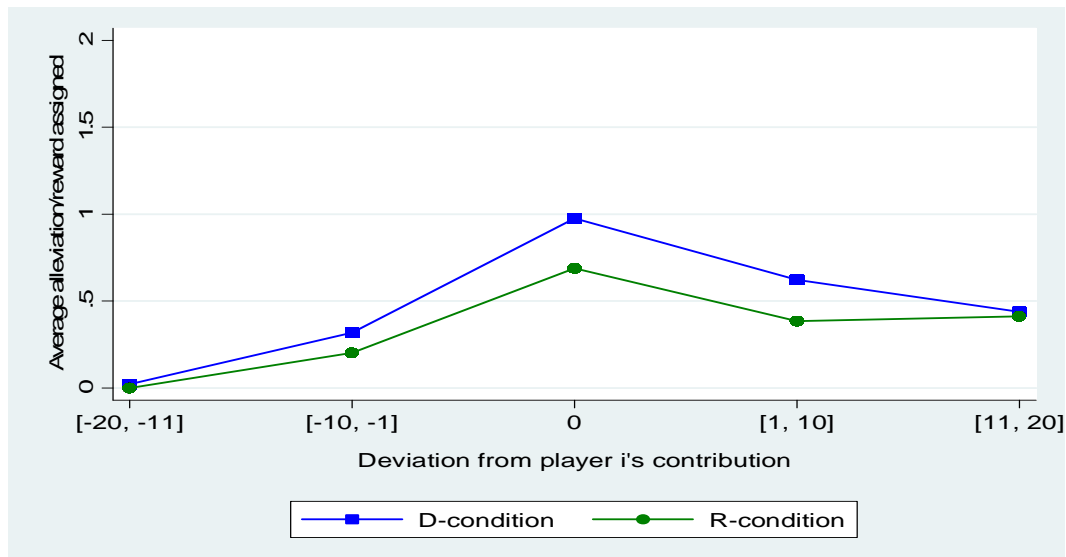
### *3.2 Alleviation vs. Reward*

We next turn our attention to the behavioural consequences that unfair punishment has on subjects' second stage behaviour. We begin by exploring whether second stage behaviour differs between the D- and the R-condition. In both conditions, assignment of an adjustment point reduces that group member's earnings by 1 Money Unit, but increases the recipient's earnings by 3 Money Units. Comparison of assigned alleviation and rewards is possible, since the cost-to-impact ratio and the sign of the adjustment points is identical across both conditions and thus, rewarding is tantamount to alleviating. This comparison is a test for whether the automatic penalty affects helping behaviour (i.e. either alleviation or reward).

Figure 2 provides a graphical illustration of how subjects alleviated and rewarded as a function of the recipient's deviation from the alleviator's/donor's contribution.

The vertical axis indicates the average alleviation and reward assigned to a group member by player  $i$ . The horizontal axis indicates the deviation in discrete intervals of the recipient's contribution from the contribution of the alleviator/donor (player  $i$ ). We refer to the solid lines of Figure 2 as the "alleviation function" or the "reward function" depending on the condition. From a visual inspection of this figure, we observe similar patterns both with respect to the level and the slope of the alleviation and the reward function. Specifically, both functions are positively sloped for negative deviations, suggesting that the less a group member contributes relative to the alleviator/donor, the less alleviation/reward is assigned to him. For positive deviations, the slope of the function is negatively sloped, indicating that higher contributions from the alleviator's/donor's contribution trigger less alleviation/reward. The intuition behind the negative slope of both functions is that in the positive deviation intervals those who assign the adjustment points are low contributors and not willing to incur costs in order to reward high contributors. Consequently, as we move further down to the right of the horizontal axis, low contributors are less and less willing to give up some of their earnings for the sake of costly alleviation and reward of other group members.<sup>9</sup>

**Figure 2.** Average alleviation/reward assigned



<sup>9</sup> Performing a Wilcoxon rank sum test fails to reject the null hypothesis that assigned alleviation is equal to assigned reward across conditions (p-values > 0.142).

For a formal analysis, we estimate an ordered probit regression model (see Appendix, Table A.2), controlling for factors that are likely to affect alleviation/reward behaviour, such as the recipient's (player  $j$ ) contribution, the absolute negative (positive) deviation from the alleviator's/donor's contribution, and a dummy variable capturing level differences between the D- and R-condition. Corroborating our statistical analysis, we find that alleviation and reward as a function of both positive and negative deviations are not statistically significant at conventional levels. Yet, as suggested by the graph above, we find that alleviation and reward functions have a positive slope for the negative deviation interval (the sign of the "absolute negative deviation" is negative) and a negative slope for the positive deviation interval (the sign of the "positive deviation" is negative as well). Additionally, assigned alleviation and reward are both found to be higher the more the recipient contributes.

Motivated by the significant time trends relative to subjects' contributions in the D-condition, but not in the R-condition, we examine how a subject reacted who got alleviated and rewarded. In our econometric model, the dependent variable is the change in the recipient's contribution between period  $t$  and period  $t + 1$ . The independent variables are the amount of alleviation/reward received from the other three group members in period  $t$  and variables measuring the time trend ("Period" and "Final period"). We estimate this model either for the cases in which a group member contributed less than or at least as much as the average contribution of the other three group members. Table 5 reports our regression results for each case separately.

**Table 5.** Reactions to alleviation/reward received

	<i>Change in contribution if a subject contributed less than the average of the other 3 group members</i>		<i>Change in contribution if a subject contributed more than the average of the other 3 group members or the same</i>	
	D-condition (DS sequence)	R-condition (RS sequence)	D-condition (DS sequence)	R-condition (RS sequence)
Received alleviation/reward in period t-1	-1.395* (0.748)	-0.983 (0.705)	0.484 (0.300)	-0.578*** (0.171)
Period	-0.342 (0.280)	0.022 (0.170)	-0.304* (0.152)	-0.246 (0.203)
Final period	1.921 (2.500)	-0.581 (2.941)	-0.339 (1.534)	-0.723 (1.396)
Constant	5.785** (1.893)	2.291** (0.933)	-1.591 (1.177)	0.646 (1.012)
Obs.	138	140	222	184

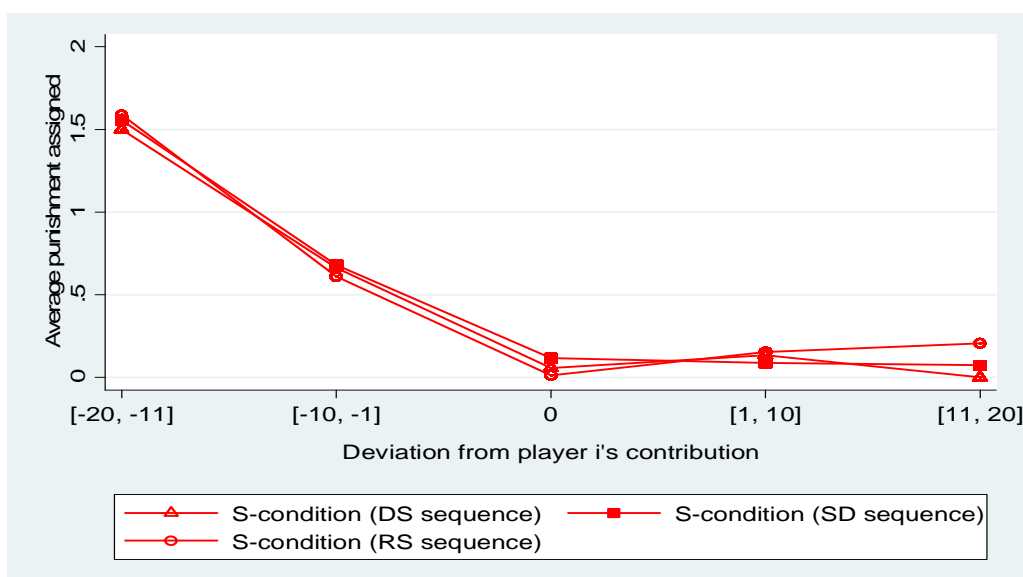
*Note: OLS estimates with robust standard errors (clustered on independent matching groups) presented in parentheses. \* denotes significance at the 10-percent level, \*\* denotes significance at the 5-percent level and \*\*\* denotes significance at the 1-percent level.*

Table 5 suggests that subjects who have contributed less than the group average do not change their contributions significantly with the received reward, whereas they lowered their contributions at least weakly significantly with the received alleviation. On the other hand, for the D-condition of the DS sequence, those subjects who have contributed at least as much as the group average do not significantly change their contributions per alleviation point received. Second, contrary to the D-condition, we find that for the R-condition of the RS sequence, those subjects who have contributed at least as much as the group average significantly decreased their contributions per reward point received.

### 3.3 Punishment

In this subsection, we turn our attention to punishment behaviour. In particular, we contrast second stage behaviour in the S-condition of the SD condition versus the S-condition of the DS sequence. Recall that helping behaviour in the D-condition is tantamount to rewarding behaviour in the D-condition. We thus isolate the effects of the rewarding element (in the D-condition) by examining subjects' willingness to punish after they experience the R-condition where there are only rewards. Figure 3 below shows the assignment of punishment in each of the three sequences.

**Figure 3.** Punishment assigned in each sequence



In the above figure, the horizontal axis indicates the deviation in discrete intervals of the recipient's contribution from the contribution of the punisher. The vertical axis shows the average punishment assigned by the punisher. We refer to the solid lines of Figure 3 as the "punishment function", which gives the average punishment points assigned by the punisher as a function of the recipient's deviation from the punisher's contribution. As previous literature would suggest, looking at the negative deviation intervals, the punishment function is negatively sloped, indicating that the more an individual negatively deviates from the punisher's contribution the higher the

punishment assigned to him. We also observe some antisocial punishment targeted at high contributors.

Our statistical analysis suggests no significant differences on average punishment assigned among the three conditions (Wilcoxon rank-sum test;  $p$ -values  $> 0.13$ ). This finding is supported by our formal econometric analysis presented in the Appendix (see Table A.3). In sum, we find that a history of either the D- or the R-condition does not have a significant impact on punishment assigned compared to an environment where there is no history at all. Not surprisingly, we observe that absolute negative deviation from the punisher's contribution is a significant determinant of punishment assigned.

Having recorded significant period effects with respect to how subjects contributed when unfair punishment is present, we further investigate how a subject reacted who got punished for a contribution above (or equal to) and for a contribution below the average contribution of the other three group members. In our econometric model, the dependent variable is the change in the recipient's contribution between period  $t$  and period  $t + 1$ . The inclusion of dependent and independent variables follows similar reasoning as in Table 6.

**Table 6.** Reactions to punishment received

	<i>Change in contribution if a subject contributed less than the average of the other 3 other group members</i>			<i>Change in contribution if a subject contributed more than the average of the 3 other group members or the same</i>		
	S- condition (DS sequence)	S- condition (SD sequence)	S- condition (RS sequence)	S- condition (DS sequence)	S- condition (SD sequence)	S- condition (RS sequence)
Received punishment in period t-1	0.288 (0.359)	1.341* (0.615)	1.179* (0.619)	-0.055 (0.370)	0.681 (0.382)	-0.529 (0.849)
Period	0.124 (0.185)	0.198 (0.220)	-0.617** (0.239)	-0.157 (0.212)	-0.040 (0.105)	-0.022 (0.067)
Final period	-3.010** (1.165)	-3.557 (3.331)	2.751 (2.241)	-0.085 (1.335)	-1.079 (1.019)	-2.404 (1.352)
Constant	1.376 (1.231)	-1.438 (1.340)	2.765 (1.560)	-0.448 (0.967)	-1.456 (0.894)	-0.270 (0.498)
Obs.	104	109	82	256	215	242

*Note: OLS estimates with robust standard errors (clustered on independent matching groups) presented in parentheses. \* denotes significance at the 10-percent level, \*\* denotes significance at the 5-percent level and \*\*\* denotes significance at the 1-percent level.*

Table 6 suggests that a history of either the D-condition or the R-condition generates different reactions with respect to punishment received when a subject contributes less than the average of the other three group members. Specifically, when there is a history of the R-condition the estimated coefficient of “Received punishment” is statistically positive, indicating that subjects who contributed less than the average contribution of the other three group members increased their contributions per punishment point received. This also occurs when there was no history before the S-condition. Yet, a history of the D-condition renders the relationship between change in contributions and punishment received insignificant, implying that subjects with such an experience did not change their contributions significantly in the S-condition that followed the D-condition. For those subjects who

contributed at least as much as the average of the other three members, it turns out that in any of the three comparisons subjects did not change their contributions per punishment point received.

#### **4. Concluding remarks**

Previous research on public good experiments with punishment suggests that punishment works when subjects assign it fairly by sanctioning non-cooperators. In this paper, we report an experiment in which punishment is assigned unfairly in the sense that punishment does not depend on the individual behaviour. Specifically, in our experiment, punishment is meted out exogenously to all group members (unconditionally and by default), irrespective of their prior behaviour. We tested whether an environment with unfair punishment generates a difference relative to the standard punishment game, both in terms of contribution behaviour and use of punishment. Our default punishment game has also a reward element incorporated in its structure as subjects can alleviate the exogenously assigned punishment. As an auxiliary condition, we therefore included a condition in which group members are only given the opportunity to reward their fellow group members, without having been exogenously punished.

Our findings suggest that contributions do not differ significantly between the default punishment game, the standard punishment game and the reward game. However, we find that the contribution pattern in the default punishment game is characterized by strong period effects, which are not present either in the standard punishment game or in the reward game. In addition, a history of an unfair environment does not affect the ability of the standard punishment game to sustain high levels of contributions; whilst, a history of the standard punishment game causes contribution levels in the default punishment game to collapse. This interesting feature of our data suggests that the perceived unfairness of the automatic penalty is more pronounced when subjects have experience of a fairer punishment scheme which identifies individual misbehaviour and can, thus, pursue a collective goal.

The behavioural effects triggered by the existence of the automatic penalty are also evident in the way individuals react to alleviation, reward and punishment received. We provide evidence that those subjects who contributed at least as much as the group average decreased their contributions per reward point received (but did not

change their contributions in the default punishment game); while those who contributed less than the group average decreased their contributions in the default punishment game (but did not change their contributions in the reward game). These findings suggest two interesting observations. On the one hand, reactions of high contributors in the reward game imply that choosing too small a reward may lead to opposite effects than those intended. On the other hand, we observe that alleviation targeted at low contributors does not appear to have a positive impact on their behaviour, as they seem to exploit it even more by decreasing their contributions (at least weakly significantly) per point received.

We also observe different reactions with respect to punishment received due to the automatic penalty. We find that subjects who contributed less than the group average increased their contributions per punishment point received after a history of the reward game or after no history at all. This observation is in line with the evidence provided by Herrmann et al. (2008). More specifically, in their cross-cultural experiment, they find that in most of their subject pools (such as Boston, Nottingham, Copenhagen, Bonn, Zurich, St. Gallen, Seoul, Chengdu and Melbourne) in which people do not punish anti-socially, those who contributed less than the group average increased their contributions per punishment point received. On the other hand, similar to our findings, in some of their subject pools with high levels of antisocial punishment (e.g., Athens, Dnipropetrovs'k, Minsk, Muscat and Riyadh), it is observed that those who contributed less than the group average did not change their contributions per punishment point received. These observations provide evidence that, at least for those who contribute less than the group average, our unfair punishment scheme (which renders punishment as not being social any more) generates similar reactions as antisocial punishment does, with respect to punishment received.

We see two avenues for future research. First, investigating the perception of automatic penalty in subject pools with high levels of antisocial punishment may lead to interesting insights on the limits of unfair punishment in breaking cooperation norms. Second, in the light of recent studies (e.g., Anderson and Putterman, 2006; Carpenter, 2007) suggesting that if norm adherence or enforcement becomes more costly, norms are more likely to collapse, it would be of interest to analyse whether and if so, how the size of automatic penalty (i.e. the cost of norm enforcement) impacts on the sustainability of social norms.

## References

- Anderson, C., and Putterman, L., 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior* 54, 1-24.
- Bochet, O., Page, T., and Putterman, L., 2006. Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior and Organization* 60, 11-26.
- Carpenter, J. 2007. The demand for Punishment. *Journal of Economic Behavior and Organization* 62, 522-542.
- Chaudhuri, A., 2007. Conditional Cooperation and Social Norms in Public Goods Experiments: A Survey of the Literature. Mimeo, University of Auckland.
- Cinyabuguma, M., Page, T., and Putterman, L., 2006. Can second-order punishment deter perverse punishment? *Experimental Economics* 9, 265-279.
- Denant-Boemont, L., Masclet, D., and Noussair, C., 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 33, 145-167.
- Egas, M., and Riedl, A., 2008. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B Biological Sciences* 275, 871-878.
- Falk, A., Fehr, E., and Fischbacher, U., 2005. Driving Forces behind Informal Sanctions, *Econometrica* 73, 2017 – 2030.
- Fatas, E., Morales, A. J., and Ubeda, P., 2010. Blind Justice: An experimental analysis of random punishment in team production. *Journal of Economic Psychology* 31, 358-373.
- Fehr, E., and Gächter, S., 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90, 980-994.
- Fehr, E., and Gächter, S., 2002. Altruistic Punishment in Humans, *Nature* 415, 137-140.
- Fischbacher, U., 2007. z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics* 10, 171-178.
- Gächter, S., and Herrmann, B., 2010. The Limits of Self-Governance when Cooperators Get Punished: Experimental Evidence from Urban and Rural Russia. *European Economic Review* (in press).

- Gächter, S. and Herrmann, B., 2009. Reciprocity, culture and human cooperation: Previous insights and a new cultural experiment. *Philosophical Transactions of the Royal Society B – Biological Sciences* 364, 761-806.
- Greiner, B., 2004. An Online Recruitment System for Economic Experiments. In: Kurt Kremer and Volker Macho (Eds.): *Forschung und wissenschaftliches Rechnen 2003, GWDG Bericht 63*. Göttingen: Gesellschaft für Wissenschaftliche Datenverarbeitung, 79-93.
- Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., and Gintis, H., 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford University Press.
- Herrmann, B., Thöni, C., and Gächter, S., 2008. Antisocial Punishment Across Societies. *Science* 319, 1362-1367.
- Ledyard, J., 1995. Public Goods: A Survey of Experimental Research. In J. H. Kagel and A. E. Roth, eds., *Handbook of Experimental Economics*, Princeton University Press, 111-194.
- Masclet, D., Noussair, C., Tucker, S., and Villeval, M-C., 2003. Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review* 93, 366-380.
- Nikiforakis, N., 2008. Punishment and Counter-Punishment in Public Goods Game: Can we still govern ourselves? *Journal of Public Economics* 92, 91-112.
- Nikiforakis, N., and Normann, H.-T., 2008. A Comparative Static Analysis of Punishment in Public-Good Experiment. *Experimental Economics* 11, 358-369.
- Noussair, C., and Tucker, S., 2005. Combining Monetary and Social Sanctions to Promote Cooperation. *Economic Inquiry* 43, 649-660.
- Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., and Nowak, M. A., 2009. Weighing Reward and Punishment: Response. *Science* 326, 1632 – 1633.
- Sefton, M., Shupp, R., and Walker, J. M., 2007. The effects of rewards and sanctions in provision of public goods. *Economic Inquiry* 45, 679-690.
- Sutter, M., Haigner, S., and Kocher, M.G., 2010. Choosing the Carrot or the Stick? – Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies* 77, 1540-1566.
- Walker, J. M., and Halloran, W. A., 2004. Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics* 7, 235–247.

## Appendix A – Regression results

**Table A.1.** Robustness of contribution behaviour in the S-condition after a history of the D-condition and the R-condition

<i>Dependent variable: Contribution</i>		
	S-condition (SD sequence) vs. S-condition (DS sequence)	S-condition (SD sequence) vs. S-condition (RS sequence)
Periods 1&2	1.645 (1.052)	0.139 (1.187)
Periods 3&4	1.447 (0.864)	0.951 (0.666)
Periods 5&6	1.408 (0.876)	1.375** (0.581)
Periods 7&8	1.007* (0.558)	1.271*** (0.322)
Condition	-1.801 (2.521)	-3.133 (2.216)
Constant	13.991*** (1.908)	15.678*** (1.291)
Obs.	760	720

*Note: OLS estimates with robust standard errors (clustered on independent matching groups) presented in parentheses. For the first comparison, the dummy variable “condition” equals 1 for the S-condition in the SD sequence and 0 otherwise. For the second comparison, the dummy variable “condition” equals 1 for the S-condition in the SD sequence and 0 otherwise. Periods 9 & 10 are the baseline. \* denotes significance at the 10-percent level, \*\* denotes significance at the 5-percent level and \*\*\* denotes significance at the 1-percent level.*

**Table A.2** Assigned alleviation for negative and positive deviations

<i>Dependent Variable: Alleviation/Reward assigned by player i</i>	
D-condition (DS sequence) vs. R-condition (RS sequence)	
Player <i>j</i> 's contribution	0.128*** (0.012)
Absolute negative deviation	-0.080*** (0.024)
Positive deviation	-0.059*** (0.016)
Condition	0.141 (0.193)
Condition × Absolute negative deviation	0.023 (0.031)
Condition × Positive deviation	-0.002 (0.020)
Obs.	2,280

*Notes: Ordered probit estimates. Standard errors presented in parentheses (clustered on independent matching groups). The variable "absolute negative deviation" is the absolute value of the actual deviation of subject *j*'s contribution from subject *i*'s contribution, when subject *j*'s contribution is below subject *i*'s contribution; and zero otherwise. The variable "positive deviation" is constructed in an analogous way. The dummy variable "condition" equals 1 for the D-condition in the DS sequence and 0 otherwise. \* denotes significance at the 10-percent level, \*\* at the 5-percent level, and \*\*\* at the 1-percent level.*

**Table A.3** Punishment assigned after a history of D- and R-condition

<i>Dependent Variable: Punishment assigned by player <math>i</math></i>		
	S-condition in the SD sequence vs. S-condition in the DS sequence	S-condition in the SD sequence vs. S-condition in the RS sequence
Player $j$ 's contribution	-0.032 (0.012)	-0.063*** (0.011)
Absolute negative deviation	0.125*** (0.025)	0.146*** (0.023)
Positive deviation	-0.025 (0.021)	0.034 (0.021)
Condition	0.101 (0.233)	0.146 (0.133)
Condition $\times$ Absolute negative deviation	0.006 (0.026)	-0.035 (0.023)
Condition $\times$ Positive deviation	-0.004 (0.025)	-0.056** (0.027)
Obs.	2,280	2,160

*Notes: Ordered probit estimates. Standard errors presented in parentheses (clustered on independent matching groups). The variable "absolute negative deviation" is the absolute value of the actual deviation of subject  $j$ 's contribution from subject  $i$ 's contribution, when subject  $j$ 's contribution is below subject  $i$ 's contribution; and zero otherwise. The variable "positive deviation" is constructed in an analogous way. For the first comparison, the variable "condition" takes the value 1 for the S-condition of the SD sequence and 0 otherwise. For the second comparison, the variable "condition" takes the value 1 for the S-condition of the SD sequence and 0 otherwise. \* denotes significance at the 10-percent level, \*\* at the 5-percent level, and \*\*\* at the 1-percent level.*