CEDEX
CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham

Discussion Paper No. 2010-06

Sergio Sousa

April 2010

Cooperation and Punishment under Uncertain Enforcement

CeDEx Discussion Paper Series
ISSN 1749 - 3293

Cooperation and Punishment under Uncertain Enforcement[*]

Sergio Sousa[†]

Jan-2010

Abstract

This paper investigates the efficacy of a punishment mechanism in promoting cooperative behaviour in a public goods game when enforcement of punishment is uncertain. Experimental studies have found that a sanctioning system can induce individuals to adopt behaviour deemed as socially acceptable. Yet, our experiment shows that a sanctioning system cannot promote cooperative behaviour if enforcement is a low-probability event and free-riding behaviour is not often punished. This supports the view that punishment needs to be exercised to be feared, otherwise the simple threat of it cannot be effective in promoting cooperation.

JEL Classication: C92, D81, H41

Keywords: uncertain enforcement; public good game; altruistic punishment; decision-making under uncertainty; cooperation.

[*] I thank Robin Cubitt and Martin Sefton for excellent supervision and comments. I am also grateful to Simon Gächter, Nikos Nikiforakis, Ana Neves, Hörst Zank and Elke Renner for useful suggestions, and participants at the 2010 ESA Asia-Pacific Conference and Cedex Seminar for comments. I gratefully acknowledge financial support from CAPES and School of Economics at the University of Nottingham.

[†] School of Economics, CeDex, University of Nottingham. E-mail: lexss12@nottingham.ac.uk

1 Introduction

This paper investigates experimentally the efficacy of a punishment mechanism in promoting cooperative behaviour when punishment enforcement is uncertain. We also investigate the role of uncertainty on punishment behaviour.

There has been a long-standing interest among social scientists and biologists in how self-interested individuals can be induced to cooperate in social dilemmas – situations where self-interested behaviour is at odds with collective interest[1]. Attempts to investigate this question have led to a number of experimental studies on how to increase cooperation in public good games. Many mechanisms have been investigated. Isaac & Walker (1988), Cinyabuguma *et al.* (2005), Guth *et al.* (2007) and Masclet *et al.* (2003) have showed that preplay communication, threat of expulsion, or even symbolic disapproval, respectively, can all boost cooperative behaviour. Alternative mechanisms such as giving subjects an opportunity to penalise others financially can also effectively increase and maintain high levels of cooperation in repeated public goods game (Fehr & Gachter, 2000). This monetary sanction system has been receiving increasing attention (see, e.g., Fehr & Gachter (2002), Sefton *et al.* (2007), Camera & Casari (2007), Nikiforakis (2008), Ertan *et al.* (2009)).

Subsequent studies have confirmed that subjects are willing to pay from their own earnings to punish defectors[2]; by doing so, they help to maintain contributions to the public good at high levels. Overall, they support the view that, at least under some circumstances, the existence of a sanctioning system can foster behaviour deemed as socially acceptable.

But despite the numerous studies on the subject, little is known about whether the existing evidence on cooperation with punishment opportunities is robust to changes in the institutional arrangement underlying the punishment mechanism. This is so because the punishment protocols in the experimental setting used in the vast majority of these studies abstract from some uncertainties involving the institution of punishment. We are particularly concerned with two types of uncertainty: uncertainty over observability (whether our actions are being watched by others) and uncertainty over enforcement (whether others' willingness to punish can be translated into actual punishment). Note that in the typical experimental design, contributions are disclosed in every period, after which punishment opportunities are given. Hence, there is certainty of being monitored all the time throughout the game. Also, there is no uncertainty about whether subject's demand for punishment will be satisfied: punishment decisions are always carried out. Therefore, the typical design implicitly assumes *perfect monitoring* and *perfect enforcement.*

However, most sanctioning systems in modern societies do not have these features. Often, there are hindrances to punishment enforcement. For instance, individuals tasked with enforc-

[1] See for example, in economics, Hardin (1968) and Axelrod (1984); in psychology Dawes (1980) and Messick & Brewer (1983); in biology Trivers (1971) and Boyd & Lorberbaum (1987); in sociology Kollock (1998) and Glance & Huberman (1993).

[2] See, e.g., Masclet *et al.* (2003), Anderson & Putterman (2006), Noussair & Tucker (2005), Bochet *et al.* (2006) and Gachter *et al.* (2008).

ing punishments can be corruptible, and anti-social behaviour, even if detected, could still not result in any penalty at all. Even when sanctions are decentralised and informal, individuals may have the willingness but not the ability to punish someone simply because an opportunity to do so will not arise. In either case, punishment is rarely perceived as certain. An experimental setting assuming perfect monitoring or perfect enforcement does not take into account those uncertainties, which could lead to a misleading assessment of the efficacy of punishment mechanisms in disciplining non-cooperators.

The aim of this study is then to relax one of those assumptions, isolating its effect on the deterrence force of punishment opportunities. We investigate, in particular, if a punishment mechanism can succeed in promoting cooperative behaviour in a public goods game when there is uncertain enforcement. While there exists a subjective element in individuals' perception of this uncertainty, assigning a probability distribution to enforcement of punishment could make the perceived uncertainty surrounding this event measurable (risk), controllable and, at least objectively, uniform across individuals.

Thus, to investigate the impact of uncertain enforcement on the common boosting effect of a punishment mechanism on cooperation, we designed an experiment that introduces measurable uncertainty into whether others' decision to punish a given player is actually carried out. To our knowledge only one previous study, by Walker & Halloran (2004), has investigated this. The authors compare cooperation in a one-shot two-stage punishment game in which imposition of sanctions is certain to a two-stage game in which imposition is uncertain. They find that uncertainty does not change the level of cooperation or the willingness to punish in a significant way. We took a second look at this issue by examining it in a repeated setting and including different levels of uncertainty. Specifically, the experiment sought to investigate how "high" and "low" enforcement probabilities affect cooperation in a repeated-play public goods game, comparing behaviour in such uncertain environments to behaviour in an environment in which punishment enforcement is certain.

While we are primarily interested in the effects of uncertain enforcement on cooperation[3], such uncertainty may also affect individuals' willingness to punish free-riders; and the effects of such uncertainty on punishment decisions are far from obvious. The reason is that backward- and forward-looking motives are likely to be driving punishment decisions. Experimental evidence suggests, for instance, that punishment is motivated by negative emotions triggered by past free-riding behaviour (Fehr & Gachter, 2002, p.139). There is also evidence that punishment tends to decrease over time (Nikiforakis, 2008, p.102), which suggests that the future matters: individuals presumably reason that, when there is a poor history of contributions, their ability to use punishment to enforce cooperation weakens as the game proceeds towards the end.

One can conjecture that uncertainty about punishment enforcement affects both backward- and forward-looking motives. Backward-looking motives because, as punishment may not be

[3] Because we want to isolate the effect of uncertainty over enforcement, costs of punishment are not incurred unless it is enforced. We come to this point later.

carried out, there may be many past episodes of free-riding behaviour that went unpunished not because of unwillingness to punish, but because "luck" got free-riders "off the hook". Hence, through its effect on backward-looking motives, enforcement uncertainty could cause an increase in punishment – reflecting a delayed outlet of accumulated negative emotions caused by past free-riding behaviour. Uncertainty could also affect forward-looking motives because the anticipation that punishment may not be enforced could weaken its strategic use. Hence, through its effect on forward-looking motives, enforcement uncertainty could cause a decrease in punishment. Given these countervailing forces, it is not clear whether punishment would be less effective even when there is no certainty over enforcement. So whether and how uncertainty affects individuals' willingness to punish free-riding behaviour is also an empirical question that remains open to investigation, providing additional motivation for this study.

Further, this study can be seen as extending the current body of research on the "robustness" of the punishment mechanism used by Fehr & Gachter (2000, 2002). Recent papers have provided evidence that punishment may not help to maintain cooperation. Even when there is certainty over enforcement, the effectiveness of punishment in promoting cooperation is sensitive to (i) its price (Anderson & Putterman, 2006), (ii) its payoff impact *per unit* of punishment (Egas & Riedl, 2008), (iii) whether individuals are given counter-punishment opportunities (Nikiforakis, 2008), (iv) feedback format (Nikiforakis, 2010) and, (v) cultural differences regarding the strength of norms of civic cooperation (Herrmann *et al.* , 2008). The findings reported here add to this literature, furthering our understanding of under which circumstances a punishment mechanism can induce cooperation in social dilemmas.

The experiment has two major results. First, that the threat of punishment cannot raise and sustain high levels of contributions when punishment enforcement is perceived by the individuals as a low-probability event. The experimental results show that a relatively low probability of non-enforcement does not impair punishment to serve as an effective deterrent device, whereas a high probability of non-enforcement does. This indicates that there is more at work in sustaining cooperation than the simple existence of a sanction system. Second, that low contributors are more intensely punished when enforcement of punishment decisions is a low-probability event. Also, and curiously enough, punishment of free-riders and low contributors is generally more intense at the beginning and the end of the game. Thus, in contrast to Walker & Halloran (2004), we find that the existence of uncertainty over the imposition of sanctions has consistent implications on subjects' decision rules.

The remainder of this paper is organized as follows. Section 2 describes the experiment design. Section 3 presents the hypotheses to be tested. Section 4 reports the results. Section 5 concludes.

2 Experimental Design

The design consists of a public good experiment with punishment with three treatment conditions. In one treatment (P100) there is certain enforcement. This corresponds to the standard case in the literature, in which punishment decisions are always enforced. The remaining two treatments differ according to the probability of enforcement of punishment decisions: one treatment with "high" probability of enforcement (P80) and the other with "low" probability of enforcement (P20) in which punishment decisions are carried out with probability of 0.8 and 0.2, respectively.

In each session, sixteen subjects are randomly partitioned into groups of four people and the composition of groups remains unchanged throughout the game – the so-called *partner matching* protocol. They play a public good game for ten periods. In each session subjects are only exposed to one of the following three treatment conditions[4].

2.1 Certain Enforcement Treatment (P100)

This treatment builds on the standard design for the public goods game with punishment, as in the seminal work by Fehr & Gachter (2000), with three differences. First, while they frame contribution decisions as an investment into a group project, we frame them as investment into a Public Account. Second, they use a convex punishment cost function while we adopt a linear one. Third, in the current experiment, group members' contributions are identified by an ID number when disclosed on the computer screen; contributions are always listed in the same ID column position[5], rather than randomly reassigned every period. Of these, we believe this last feature is potentially a major distinction from the standard design; it allows participants to create a link between the actions of other group members across periods. There are two reasons for that. First, by allowing individualization, we reduce the possibility for indiscriminate punishment and make interpretation of data more transparent. Second, by allowing subjects to track group member's contributions, we can investigate the extent to which punishment decisions are influenced by contributions in previous rounds. We allow this feature in P100 too to avoid potential confounding effects of probability of punishment enforcement with informational differences.

At the beginning of each of the ten periods, each subject is endowed with a fixed amount of 20 Rubis (the experimental currency used). Each period unfolds in two stages. In the first stage, subjects are required to simultaneously decide how much of their endowment to invest in a Public Account, say c_i, and, consequently, how much of it to invest in a Private Account, $20 - c_i$. Each Rubi a player allocates to the Private Account has a return of 1 for that player. A Rubi allocated to the Public Account yields a return of 0.4 for *every* player in the group. At the end of the first-stage, each subject is informed of the group's total investment, her income from the Public Account and her first-stage earnings (π), which is given by:

[4] Instructions are available upon request.

[5] Although players could track a particular co-player's contribution record, they have no way of identifying that person. This matters (a) for ethical treatment of subjects and (b) for elimination of confounds as subjects might respond to information contained in the name (e.g gender, nationality, etc).

$$\pi_i^1 = 20 - c_i + 0.4 \sum_{i=1}^{4} c_i \tag{1}$$

Note that the total return of investment in the Public Account depends on the total investment made by the entire group. While each Rubi allocated to the Public Account yields a marginal private return of less than 1, by investing in the Public Account players in a group may obtain earnings that exceed those associated with full investment in the Private Account. Investments in the Public Account, given its non-rivalness and non-excludability, can be seen as contributions to a public good.

In the second stage, participants are informed of the investment decisions of their group members and given the opportunity to punish each group member by assigning "deduction" points. Each deduction point costs the punisher one Rubi and reduces the punished players' first-stage income by 3 Rubis. Each subject can assign up to 10 "deduction points" to each one of the other members of the group.

Additionally, it is imposed that a subject cannot have her first-stage income, π^1, reduced below zero as a result of the punishment given her by others. Nevertheless, as she always carries the cost of punishment she does, her period income may end up negative depending on the total number of "deduction" points received and assigned[6]. Subject i's end-of-period payoff is given by:

$$\pi^2 = \left\{ \begin{array}{l} \pi^1 - 3(P_{-i,i}) - P_{i,-i} \quad \text{if} \quad 3(P_{-i,i}) < \pi^1 \\ -P_{i,-i} \quad \text{if} \quad 3(P_{-i,i}) \geq \pi^1 \end{array} \right\} \tag{2}$$

where $P_{-i,i}$ stands for the number of deduction points imposed on subject i by the other group members, and $P_{i,-i}$ stands for the number of punishing points assigned by subject i to all other group members.

2.2 Uncertain Enforcement Treatments (P80 and P20)

The uncertain treatment conditions are equivalent to the P100 except that now one stage is added after the second stage, which we refer here to as the "enforcement" stage. Recall that in the second stage, subjects are informed of the contribution decision of each other group member and are given the opportunity to punish them. In the "uncertain enforcement" treatment conditions (P80 and P20), they do so with the understanding that their "punishment"[7] decisions may not be carried out; they will be so with a probability $p < 1$, which is the same for all 10 periods of the experiment.

Note that it is as if their punishment decisions were delegated to a central authority who, depending on the state of the nature, say S, may fail to implement their decisions. Thus, there

[6] As in Fehr & Gachter (2000), Nikiforakis (2008) and others, each subject is given a one-time lump-sum payment of 25 Rubis at the beginning of the experiment to pay for negative payoffs they might incur during the experiment.

[7] We did not use this terminology ("punishment") in the experiment.

were two states of the nature (s_1 and s_2) and punishment decisions are enforced only when $S = s_1$, where $P(S = s_1) = p$.

In each period, whether or not punishment decisions are enforced is decided, for each group, as follows: a ball is drawn from a bingo cage with replacement. The bingo cage has balls numbered from 1 to 10. For a group in the P80 (P20) condition, if the ball drawn is numbered 9 or 10 (3 to 10), then punishment decisions are not carried out. In these cases, a subject's end-of-period earnings are equal to her earnings in the first stage. Otherwise, punishment decisions are implemented and the final earnings in the period are given by the equation in (2).

To avoid there being any communication of disapproval when punishment is not enforced (i.e., nonmonetary forms of punishment, see Masclet *et al.* , 2003), punishment assigned to each individual in a given group is not disclosed unless it is enforced. So only when punishment decisions are actually implemented are subjects informed of the total punishment points they received from the group. In a similar fashion, assigning punishment points will not have any cost to subjects if punishment is not enforced. This could correspond to a case where an opportunity to punish, as opposed to willingness to, simply does not arise. More importantly, this feature of our design avoids one's profile of punishment decisions being "contaminated" by her unwillingness to pay for something that may not happen.

Thus, the information disclosed at the end of each period depends on the enforcement state: in case punishment is not enforced, subjects are shown their final earnings, which in this case is equal to their earnings from the first stage. In case punishment is enforced, they are shown (a) the total cost of the punishment points they assigned, (b) the punishment points they received in total from the group, and (c) the associated reduction in their earnings along with their final earnings in the period. All subjects are also informed of their own accumulated earnings, which are equal to the sum of earnings over all previous periods.

In all three treatments conditions, the parameters of the experiment (endowment, the return rate from the Public and Private Accounts, group size, payoff functions, number of rounds) are publicly announced to the participants.

2.3 Administration

There were ninety six subjects in this experiment. None of them had previously participated in a public good experiment at the University of Nottingham[8]. The subjects signed up for one of six sessions. At that point, they only knew that the experiment would take up to 90 minutes. Treatment conditions were randomly allocated to sessions, with two sessions per treatment condition.

Sixteen subjects took part in each session. Following their arrival, each subject received instructions explaining the experiment[9]. The instructions were read aloud while the students read them silently. To ensure subjects' understanding of the game's structure and payoff de-

[8] Participants were recruited using ORSEE (Greiner, 2004), a subject-recruitment software that allows us to exclude from the database individuals who have previously participated in public good experiments.

[9] Instructions are available upon request.

termination, each of them was asked to complete a control questionnaire. The experiment only proceeded when all subjects had answered it correctly. The experiment was conducted using the software z-Tree (Fischbacher, 2007) and sessions took around fifty minutes to be completed. At the end of the experiment, subjects were asked to complete a short questionnaire about themselves. Their earnings were converted into Sterling Pounds and they were then paid in cash. The exchange rate was 1 Rubi = 2.5 pence. Participants earned on average £8.51, which included a show-up fee of £2 and a one-time lump-sum payment of 25 Rubis.

3 Theory: Effects of uncertain enforcement on cooperation and punishment

We now present predictions for cooperative and punishment behaviour. We start with cooperative behaviour. First, we consider a standard game-theoretic case in which players are of the same type: they are all strictly concerned with their material payoff. Then, we consider a mixed case in which some of the players have fairness concerns.

3.1 Cooperation: Homogeneous players

Assuming that individuals are monetary payoff maximizers and that this is common knowledge among them, they should contribute nothing to the Public Account. In the presence of punishment opportunities this still holds true. The threat of punishment is non-credible as this is a payoff-reducing action. Therefore, subgame perfection dictates that individuals would always be better off by not punishing at all. It is straightforward to see that the equilibrium outcome regarding punishment does not change when punishment enforcement is risky. In this case, the actual infliction of punishment is conditioned on a probability distribution over a set of states of nature. Even so, costly punishment would not be a credible action by self-interested payoff maximizers regardless of the status of enforcement. Since individuals do not punish, one should contribute nothing just as in the "certain enforcement" case. Thus, within the standard game-theoretic framework, zero cooperation and zero punishment would be the subgame-perfect equilibrium strategies in all enforcement conditions.

There is plenty of evidence, however, that these standard predictions hardly describe cooperation and punishment decision in public good games. The evidence suggests that some individuals have other-regarding preferences. This brings us to the next case.

3.2 Cooperation: Heterogenous players

If individuals have other-regarding preferences and are motivated by more than their pecuniary payoffs then no punishment and full defection may not be an equilibrium outcome. Fehr & Schmidt (1999) show, for instance, that if some people care about payoff equity, full cooperation can be sustained as an equilibrium outcome in a public good game with punishment. As individuals who care about disadvantageous inequality will be willing to punish defectors despite it being costly to them. Such a threat, given the information set of players, would be credible enough to sustain cooperation.

A number of experimental studies have shown that individuals are indeed willing to pay to punish defectors, and that high levels of cooperation can indeed be sustained in the presence of punishment. But uncertainty over punishment enforcement may change the decision setting. To get an insight into this, we use a simple model for a two-player case.

Let us start with some preliminaries. Consider a game G played by two players. Each player has a type that determines the preferences she acts on. Player 1 is purely self-regarding ("selfish") with a utility function $u_1(.)$ defined on her own payoff, π_1, in the game. Player 2 is inequity-averse with a utility function $u_2(.)$ defined both on her own payoff and the other player's payoff, π_2 and π_1. $u_1(.)$ has a linear form defined by $u_1(\pi_1) = \pi_1$. Player 2's utility function, $u_2(.)$, has a Fehr-Schmidt functional form (see Fehr & Schmidt, 1999, p.822) defined by

$$u_2(\pi_1, \pi_2) = \pi_2 - \alpha \max\{\pi_2 - \pi_1, 0\} - \beta \max\{\pi_1 - \pi_2, 0\}$$

With that in mind, let G be the following complete information public good game with three stages. In the first stage, players decide simultaneously whether or not to contribute to the public good. Each player has an endowment of e, so that $c_i \in \{0, e\}$ $(i \in \{1, 2\})$ is the discrete set of strategies each player can employ. Payoff at the end of this stage is given by

$$\pi_{i,C}(c_i, c_j) = e - c_i + r(c_i + c_j) \ , \ r \in (1/2, 1)$$

where r is the return to each player from contributions to the public good. In the second stage, each player is informed about the other player's contribution and both decide simultaneously whether or not to impose a punishment on the other player. Let $p_i \in \{0, \rho\}$ $(i \in \{1, 2\})$ be the strategy each player can employ in this stage. This stage's payoff is given by

$$\pi_{i,P}(p_i, p_j) = -[p_i + lp_j] \ , \ l > 1$$

where $l > 1$ is the punishment impact rate, which indicates the first-stage payoff deduction when the other player chooses to punish. Finally, in the third stage "nature" chooses whether to enforce players' punishment decisions; "nature" enforces punishment with probability $q \in [0, 1]$. Note that players move in the second stage without knowing what is nature's choice. The monetary payoff of a player i is simply $\pi_i = \pi_{i,C} + \pi_{i,P}$.

What is the prediction for this game under the assumption that players are of different *types*? More specifically, how are decisions in the first stage affected by the probabilistic enforcement of punishment decisions? The prediction is summarized in the following proposition:

Proposition 1

(1.1) *If $q = 0$, then it is a dominant strategy for both players to choose $c_i = 0$.*

(1.2.) *If $q > 0$, then punishment can sustain an equilibrium in which $c_i > 0 \ \forall \ i$ when (i) $\alpha > \frac{1}{l-1}$ and (ii) $q > q^* = \frac{e(1-r)}{l\rho}$.*

Proof *See Appendix B*

The crucial implication of the above results is that the threat of punishment can only induce self-regarding players to contribute (hence, sustain full cooperation in the game) if the probability q of punishment enforcement is sufficiently high. Otherwise, if q is too low, the self-regarding type will free-ride because expected punishment is too low to deter defection. In this case, free-riding is also the best response of the inequity-averse player. Based on Proposition 1, we conjecture the following

Hypotheses 1 (Contribution) *The presence of punishment opportunities will not raise contributions if enforcement is perceived as "weak" (low-probability event) to a sufficiently high proportion of subjects. The lower q, the more free-ride types will be, breaking down prospects of sustained cooperation.*

What are the predicted effects of probability enforcement on punishment behaviour? We know of no formal hypothesis that has been put forward which would allow us to predict the direction of punishment enforcement probability effects in subjects' punishment behaviour; and our previous basic framework regards only cooperative behaviour. Yet, we conjectured in the introduction that the effect of imperfect enforcement on punishment is ambiguous, ultimately depending upon how backward- and forward-looking elements influence punishment decisions.

3.3 Punishment with forward-looking dominance

If forward-looking motives dominate punishment decisions, then we conjecture that the higher is the uncertainty over punishment enforcement, the less punishment will be observed. The intuition is that the anticipation that punishment may not be enforced would weaken its strategic use as an instrument to shape future interactions. If uncertainty creates a hindrance to individuals' ability to influence the future, then the higher the uncertainty, the less punishment would be exercised by individuals.

Hypotheses 2.1 (Punishment) *If punishment decisions are dominantly driven by forward-looking motives, then punishment points assigned (not necessarily implemented) to free-riders and low contributions will be higher in P100 and P80 than in P20.*

3.4 Punishment with backward-looking dominance

Conversely, if backward-looking motives dominate punishment decisions, then we conjecture that the more likely it is that free-riders can escape punishment due to enforcement failure, the more intense will be the willingness to punish them. The intuition here is that "bygones are not bygones" and players might get more intensely punished in a given round t for failing to contribute in t *and* in rounds prior to t. By punishment intensity we mean the number of punishment points assigned to a player per deviation of her contribution from others' average.

Note that with weak enforcement in a repeated setting, there will probably be players, especially in treatment P20, with a history of free-riding behaviour that went unpunished because other players' punishment decisions were not enforced. Thus, if punishment is dominantly backward-looking and mainly directed towards free-riders, then more punishment will be directed towards free-riders in P20 than in P100 and P80 treatments.

Hypotheses 2.2 (Punishment) *If punishment decisions are dominantly driven by backward-looking motives, then punishment points assigned (not necessarily implemented) to free-riders and low contributions will be higher in P20 than in P80 and P100.*

If that is the case, one may wonder whether this extra punishment would not compensate the low probability of enforcement and lead to high contributions. Note, though, that a more intense *willingness* to punish a free-rider may have no bearing on cooperation, as willingness to punish may not translate into actual punishment. Thus, even if others are willing to punish free-riders more intensely in P20 than, say, in P80, this may not necessarily induce cooperation levels in P20 as high as in P80.

4 Experimental Results

4.1 Cooperative Behaviour

We start by examining contribution patterns across treatments. Figure 1 presents box plots of contribution to the Public Account over the 10 periods for each treatment condition.

Each box plot describes key distributional features of the data. The median contribution value is shown as a line drawn across the box[10]. The variability in contribution is represented by first (lower hinge) and third (upper hinge) quartiles of the distribution in each period. Let this interquartile difference be H. The lower and upper adjacent values of the contribution in each period are shown as "whiskers" in the plot lines used to outline a box[11]. The upper adjacent contribution value represents the largest contribution between the upper hinge (uh) and the threshold value of $uh + 1.5 \times H$; the lower adjacent contribution value, in turn, represents the smallest contribution between the lower hinge (lh) and the threshold value of $lh - 1.5 \times H$. Dots outside the box plot identify contributions that lie unusually far from the main body of data[12].

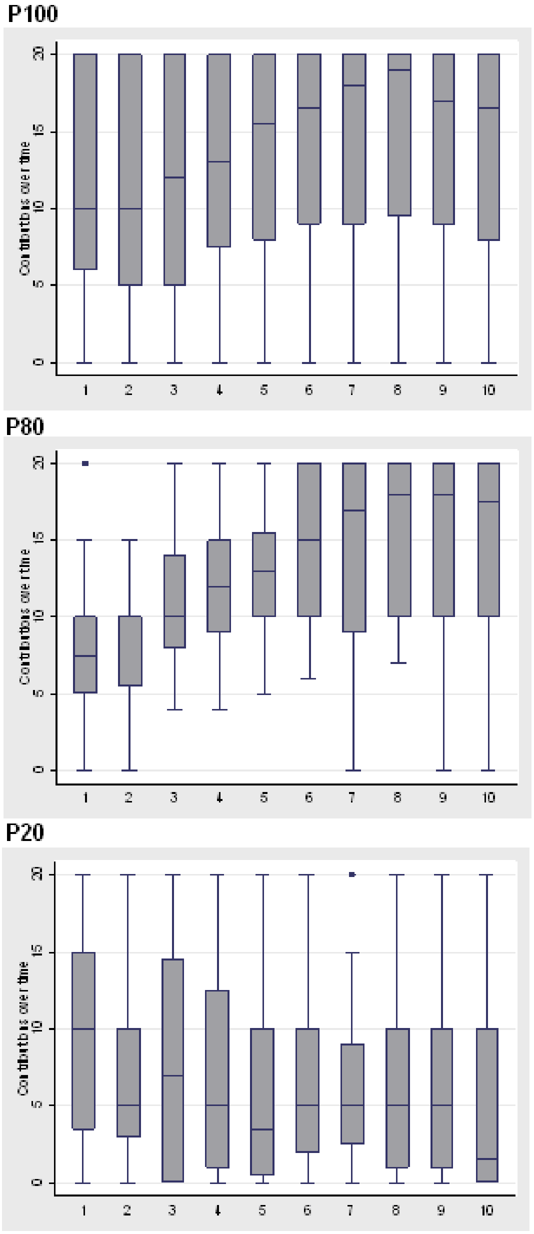
In the box plots for each treatment, by following the line drawn across the box at the median, one can see the evolution of median contribution to the Public Account over the ten periods. Contributions under the P100 condition, for instance, are in line with previous experimental findings: they start at roughly half of subjects' endowments and keep increasing over time. This result confirms that the existence of punishment can improve cooperation over

[10] In the second period of P80, the line representing the median contribution seems absent of the box plot. This is because it coincides with the upper quartile (10).

[11] For contribution data in some periods, these adjacent values coincide with the first (P20, periods 3 and 10) or third quartile (e.g. all periods in P100) of the distribution of contributions.

[12] Any contribution which lies more than three times the inter-quartile range, either lower than the first quartile or higher than the third quartile, falls into this category.

Figure 1: Average contribution, by enforcement condition

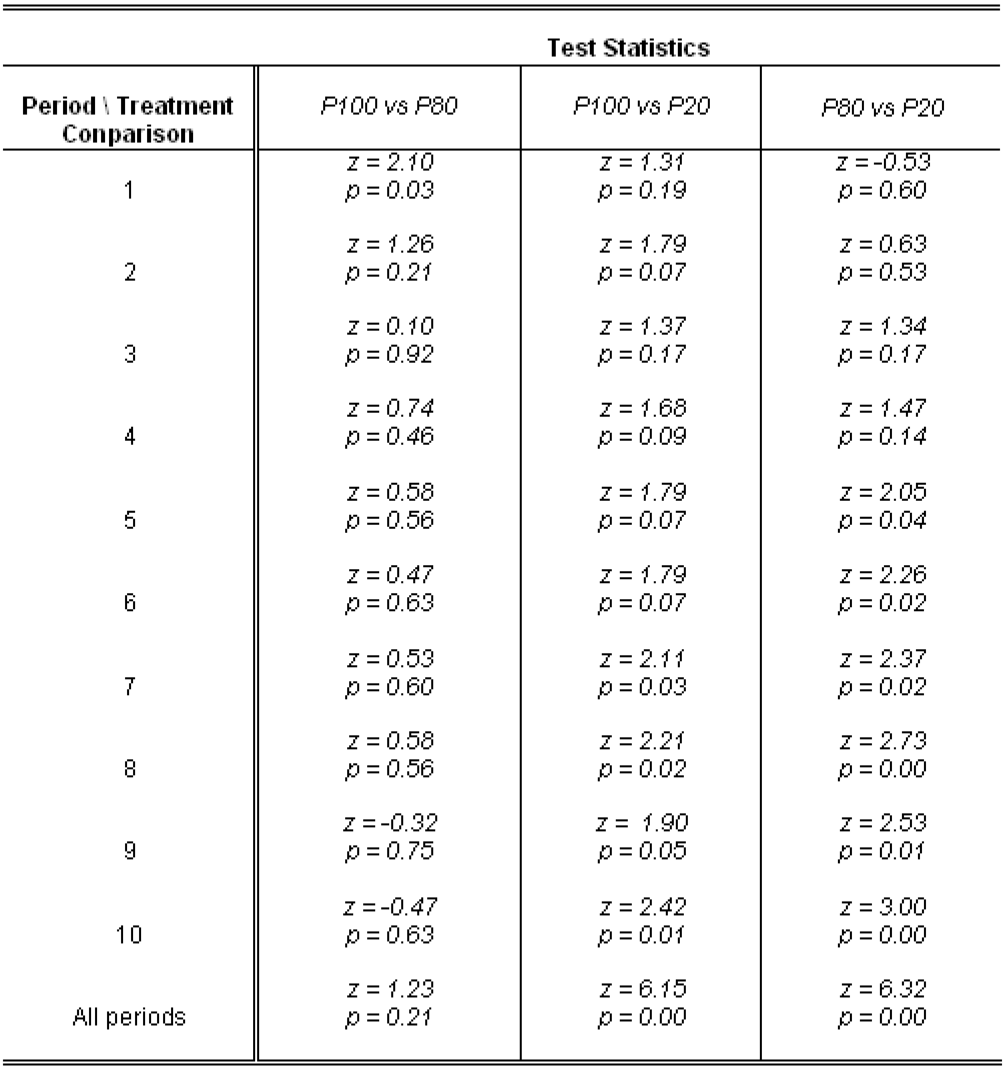


time. Additionally, and perhaps more importantly, it suggests that the ability of punishment to sustain cooperation is unaffected by knowledge of contribution histories.

There is, however, clear separation in contributions between the uncertain enforcement treatments. While median contributions in the P80 condition increase over time, closely following contributions in the certain enforcement condition, it is clear that contributions in the P20 condition are noticeably lower, and on a divergent path, than in the P80 condition. While median contributions in the P20 condition start higher than contributions in the P80 condition, they keep decreasing from the second period on, while contributions in the P80 treatment increase over time. This suggests that the existence of punishment opportunities is not effective in raising contributions if enforcement is perceived as a "low" probability event.

This result, based on visual inspection, is indeed confirmed by non-parametric tests. We conduct pairwise Mann-Whitney tests between treatments for each period at a time in order to test for equality of distribution of mean contribution of groups between all enforcement treatments[13]. Test statistics are reported in Table 1.

Table 1: Are groups' mean contribution different across enforcement treatments? Pairwise Mann-Whitney Tests



	Test Statistics		
Period \ Treatment Comparison	*P100 vs P80*	*P100 vs P20*	*P80 vs P20*
1	*z = 2.10* *p = 0.03*	*z = 1.31* *p = 0.19*	*z = -0.53* *p = 0.60*
2	*z = 1.26* *p = 0.21*	*z = 1.79* *p = 0.07*	*z = 0.63* *p = 0.53*
3	*z = 0.10* *p = 0.92*	*z = 1.37* *p = 0.17*	*z = 1.34* *p = 0.17*
4	*z = 0.74* *p = 0.46*	*z = 1.68* *p = 0.09*	*z = 1.47* *p = 0.14*
5	*z = 0.58* *p = 0.56*	*z = 1.79* *p = 0.07*	*z = 2.05* *p = 0.04*
6	*z = 0.47* *p = 0.63*	*z = 1.79* *p = 0.07*	*z = 2.26* *p = 0.02*
7	*z = 0.53* *p = 0.60*	*z = 2.11* *p = 0.03*	*z = 2.37* *p = 0.02*
8	*z = 0.58* *p = 0.56*	*z = 2.21* *p = 0.02*	*z = 2.73* *p = 0.00*
9	*z = -0.32* *p = 0.75*	*z = 1.90* *p = 0.05*	*z = 2.53* *p = 0.01*
10	*z = -0.47* *p = 0.63*	*z = 2.42* *p = 0.01*	*z = 3.00* *p = 0.00*
All periods	*z = 1.23* *p = 0.21*	*z = 6.15* *p = 0.00*	*z = 6.32* *p = 0.00*

Two features stand out in the test results. First, they show that, apart from the first period, there are no significant differences in mean contribution of groups in the P100 and P80 treatments. Second, they also show that there are statistically significant differences between mean contribution of groups in P20 and either P100 or P80 treatments after the initial periods of the game (in most periods, at the 5% level of significance). Both features are more salient when considering test results involving data from all periods pooled together.

We need, however, to examine the robustness of these results. The non-parametric tests do

[13] It is worth noting that the sample of observations from a given treatment is formed by the mean contribution of groups of players in a given treatment. This is so because individuals' contributions, while independent across samples, are not independent within treatments – which violates an assumption which the test relies on.

not capture intertemporal dependencies in group contributions and may confound treatment effects. We then turn to a more formal analysis of the data; we do so by running a regression of individual contributions on treatment and individual variables. The panel structure of the data allows us to handle some degree of individual heterogeneity and obtain more consistent estimates of treatment effects.

We estimate an empirical model relating contribution to individual and structural parameters of the game that largely follow a common specification in these studies (e.g. Anderson & Putterman, 2006; Nikiforakis, 2008). But our econometric specification also includes lagged variables that seek to capture recursive elements in contribution decisions. The underlying reason for this is hardly controversial: in repeatedly played games, individuals tend to reciprocate actions of other players; this produces behaviour that is largely reactive and influenced by past outcomes (see, e.g. Fischbacher *et al.* (2001), Frey & Meier (2004) and Gunnthorsdottir *et al.* (2007)). The model has then the following specification:

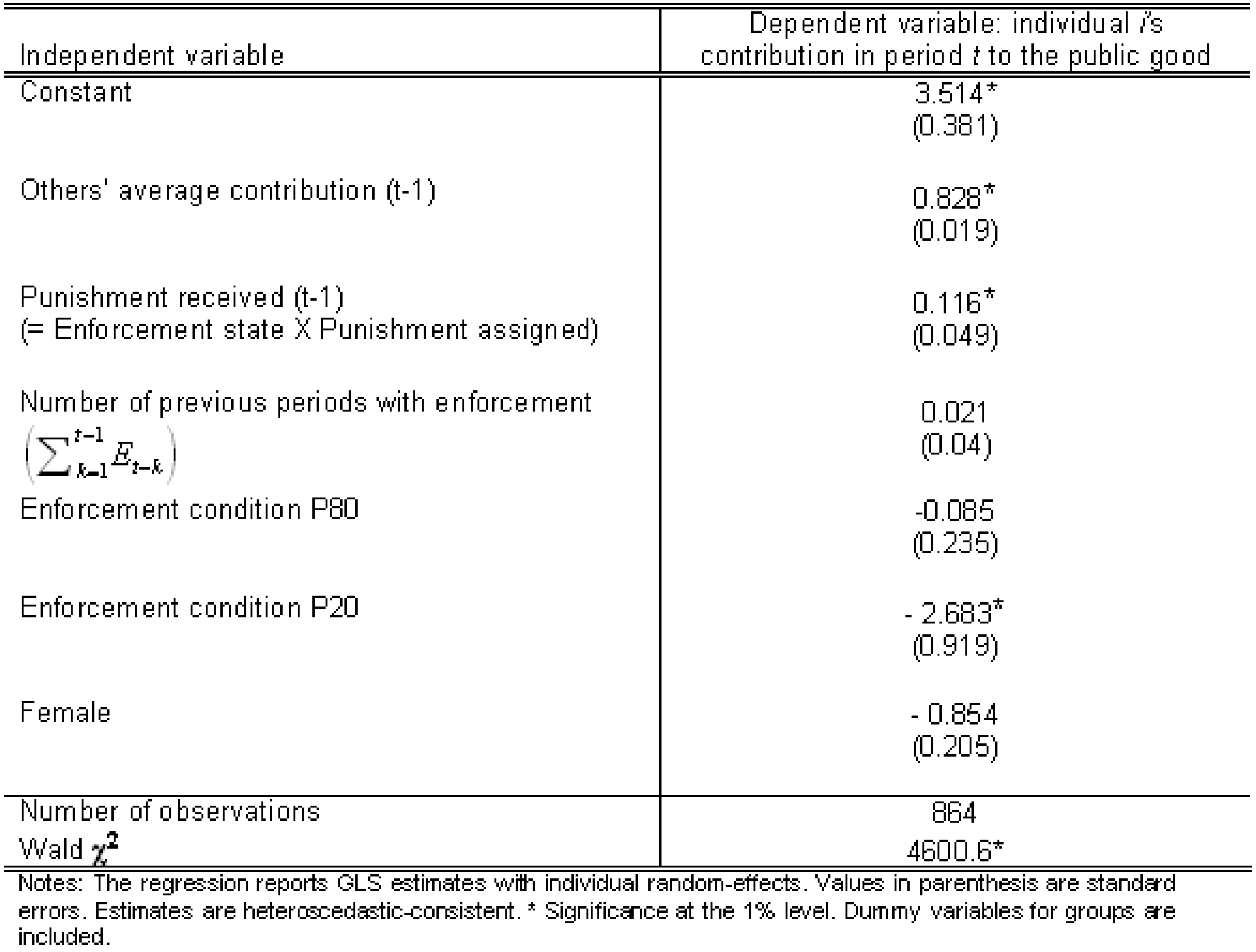
$$c_{i,t} = \beta_0 + \beta_1 \bar{c}_{-i,t-1} + \beta_2 \left(P^R_{i,t-1} \right) + \beta_3 \Sigma E_{i,t-1} + \beta_4 P80 + \beta_5 P20 + \mathbf{z_i'} \alpha + \mathbf{u_{i,t}} \qquad (3)$$

where the $\bar{c}_{-i,t}$ is the average contribution of the other group members in period t, $P^R_{i,t}$ is the total punishment points actually received by individual i in period t – which is 0 if punishment decisions were not enforced. $\Sigma E_{i,t}$ is the number of previous periods in which punishment was enforced in the group i belongs to; this is meant to capture the effects of the particular sequence of enforcement experienced by i. $P80$ and $P20$ are dummy variables that equal one if individual i is taking part in the "high" or "low" probability of enforcement condition, respectively. Components of $\mathbf{z}$ will control for the variation strictly related to some subject-specific attributes (gender, ethnicity, etc). Dummy variables to control for group effects are included. $u_{i,t}$ is a composite error term including a subject-specific random intercept and a purely random disturbance term which is assumed to be i.i.d. over i and t.

Table 2 reports the results of the generalized-least-squares regressions of the model in (3). Contributions are, on average, positively affected by retaliatory behaviour from others in the past: the actual number of punishment points received in the previous period as well as the number of periods in which punishment points were actually enforced have both significant and positive effect on contributions. Of interest in the results is the parameter in front of the dummy variable $P20$; it represents the estimated effect of the "low" probability of enforcement treatment effects on contribution decisions. Note that even after controlling for the different enforcement conditions and group effects (interaction and sequence of enforcement experienced by groups), one can see that contributions from subjects in the low-probability of enforcement treatment are lower than contributions in both certain and "high" probability of enforcement conditions. $P20$ is, in fact, the only enforcement treatment whose effect on contributions is statistically significant. Thus, parameter estimates of the model support the raw results depicted in Figure 4.1.

Therefore, as was apparent in Figure 1, there are significant differences in contribution estimates between "high" and "low" probability of enforcement conditions. The mere knowledge

Table 2: Do enforcement treatments affect contributions?



Independent variable	Dependent variable: individual *i*'s contribution in period *t* to the public good
Constant	3.514* (0.381)
Others' average contribution (t-1)	0.828* (0.019)
Punishment received (t-1) (= Enforcement state X Punishment assigned)	0.116* (0.049)
Number of previous periods with enforcement $\left(\sum_{k=1}^{t-1} E_{t-k}\right)$	0.021 (0.04)
Enforcement condition P80	-0.085 (0.235)
Enforcement condition P20	- 2.683* (0.919)
Female	- 0.854 (0.205)
Number of observations	864
Wald χ^2	4600.6*

Notes: The regression reports GLS estimates with individual random-effects. Values in parenthesis are standard errors. Estimates are heteroscedastic-consistent. * Significance at the 1% level. Dummy variables for groups are included.

that sanctions may be imposed to punish those regarded as free-riders cannot induce cooperative behaviour if punishment enforcement is viewed as "weak". Based on the non-parametric and regression analysis one can conclude that:

Result 1. *The threat of punishment can only promote cooperative behaviour if enforcement is* perceived *as a high-probability event.*

4.2 Punishment Behaviour

The next issue to be examined is whether and how subjects' willingness to punish is affected by the possibility of not having their punishment decisions enforced. To get an intuition on this, we begin with some descriptive statistics.

Table 3: Fraction of subjects who assign no punishment points

Period	P100	P80	P20
1	0.31	0.53	0.57
2	0.65	0.66	0.63
3	0.50	0.53	0.63
4	0.53	0.59	0.63
5	0.63	0.66	0.56
6	0.69	0.63	0.72
7	0.66	0.53	0.72
8	0.69	0.56	0.69
9	0.56	0.63	0.72
10	0.66	0.66	0.63

Table 3 presents the frequency of individuals who assign no punishment. Two things are worth noting: first, that there is a considerable amount of "free-riding" behaviour on punishment efforts across treatments. In most periods, the option to punish is exercised by less than half of the subjects. Second, that there is more punishing of individuals in the first period of the certain enforcement conditions than there is in the uncertain enforcement conditions.

Examining punishment points assigned to subjects, we find that individuals in the P100 condition were assigned more punishment points (2.19) on average than those who are in the P80 and P20 conditions (1.41 and 1.53). A Mann-Whitney test shows that this difference in punishment in the first period is significant at 5% level of significance (P100 versus P80: $p < 0.0224$; P100 versus P20: $p < 0.0507$). A natural question to ask is why subjects in P100 are imposing more sanctions relative to P80 and P20 conditions in the beginning of the game?

It is not that there is a great deal more of free-riding behaviour in the P100 relative to the other two conditions. In the first two periods, only one subject in the P100 contributes nothing to the Public Account, against five and four subjects in the P80 and P20 conditions, respectively. While the fraction of subjects who in the first period contribute less than the group average is slightly greater in P100 (59%) than it is in P80 and P20 conditions (56% and 51%), average contribution in P100 is actually higher (11.03) than it is in P80 and P20 treatments (7.78 and 9.03, respectively). Kolmogorov-Smirnov tests provide a second bit of evidence consistent with that: they show that one cannot reject the hypothesis that there is no significant differences in the distribution of deviations from others' average contribution between treatments in the first period (P100 versus P80: $p < 0.627$;P100 versus P20: $p < 0.964$; P80 versus P20: $p < 0.627$).

A possible interpretation of this first-period differences in punishment between treatments is that subjects in the certain enforcement condition are trying to discipline behaviour from the beginning by signalling "toughness" with free-riders and low-contributors. Yet, this strategic reputation building would be mitigated among subjects in P80 and P20 enforcement conditions. Because they know that their punishment decisions may fail to be enforced, they would be unwilling to accept the cost of enforced punishment as the potential "extra" cost of such strong signals early in the game may not be compensated by higher cooperation levels later in the game. This is likely to be the case of a forward-looking subject who believes that punishment will only work if it is enforced frequently, in which case it would be rational not to punish in P20 even though unenforced punishment is costless.

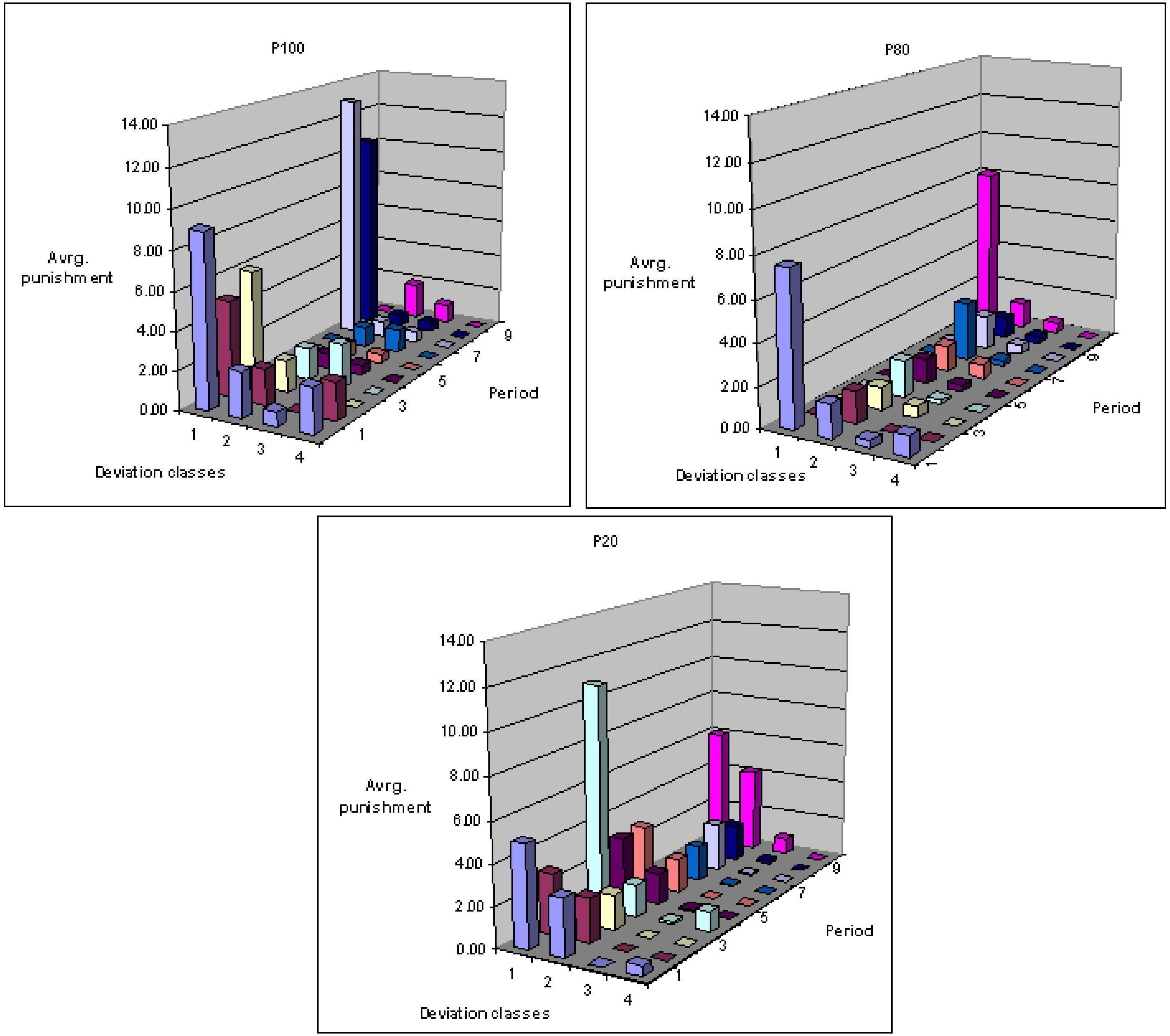
Graphics in Figure 2 show other interesting aspects of punishment behaviour in each enforcement condition. Each figure plots the average punishment points assigned to i by range of deviation from the others' average contribution over time. There are four classes of deviation: 1 (-20,-10], 2 (-10,0], 3(0,10], and 4 (10,20]. For example, the leftmost cluster of ten columns in each figure shows the average punishment assigned to individuals whose contribution is between 10 and 20 Rubis *less* than the average contribution from other group members (Deviation class 1). The other three ranges of deviation move towards a positive domain as one moves to the right-hand side of the "Deviation classes" axis of the graph.

Visual inspection of these plots suggests three things. First, that there are some sort of first- and last-period effects. Note that the willingness to punish free-riders and low contributors (those in deviation class 1) is stronger at the beginning and at the end of the game. Second, that there is no "anti-social punishment" when enforcement is weak: individuals whose contribution is above the average contribution from the other group members – deviation classes 3 and 4 – are barely punished in P20. Third, that "negative deviators", especially in the middle rounds, are more intensively punished in P20 than they are in P100 and P80: individuals whose contribution falls short of the average seem, on average, to have more punishment points assigned to them in P20 treatment than in the certain and P80 enforcement conditions.

We now perform an econometric analysis of treatment effects on punishment behaviour. We regress the amount of punishment assigned to a player on lagged contribution treatment and structural parameters of the game. The general empirical model has the following form:

$$\begin{aligned} P_{i,t} &= \beta_0 + \beta_1 \bar{c}_{-i,t} + \beta_2 POSDEV + \beta_3 NEGDEV + \\ &\quad + \beta_4 ANGER + \beta_5 P80 + \beta_6 P20 + \mathbf{z}'\alpha + \mathbf{u_{i,t}} \end{aligned} \tag{4}$$

Figure 2: Average punishment, by range of deviation from others' average contribution



where $P_{i,t}$ represents the number of punishment points assigned to subject i, $\bar{c}_{-i,t}$ is the average contribution from other group members, $POSDEV$ and $NEGDEV$ are the absolute values of the deviation of i's contribution from other group members' average. We follow here (Fehr & Gachter, 2000), including them as separate regressors. One of those variables is zero depending whether i's contribution is either above (or equal) or below the others contribution. $ANGER$ denotes all the punishment points assigned i that have not been actually enforced over the previous periods. $P80$ and $P20$ are dummy variables that are equal to 1 if i is in the $P80$ or $P20$ enforcement treatments and 0 otherwise. Due to the random assignment of participants to treatment conditions, those dummies allow us to isolate the effect of enforcement conditions on subjects' willingness to punish. $\mathbf{z}$ is a vector of other dummies and interaction terms between treatment conditions and deviation from i's contribution from other group members' average that try to capture different levels of intensity of punishment assignment in each treatment condition. We include, for instance, a dummy regressor for the last period to capture last period effects on punishment decisions. $u_{i,t}$ is the compound error term. Parameter estimates of model 4 are presented in Table 4 column (1).

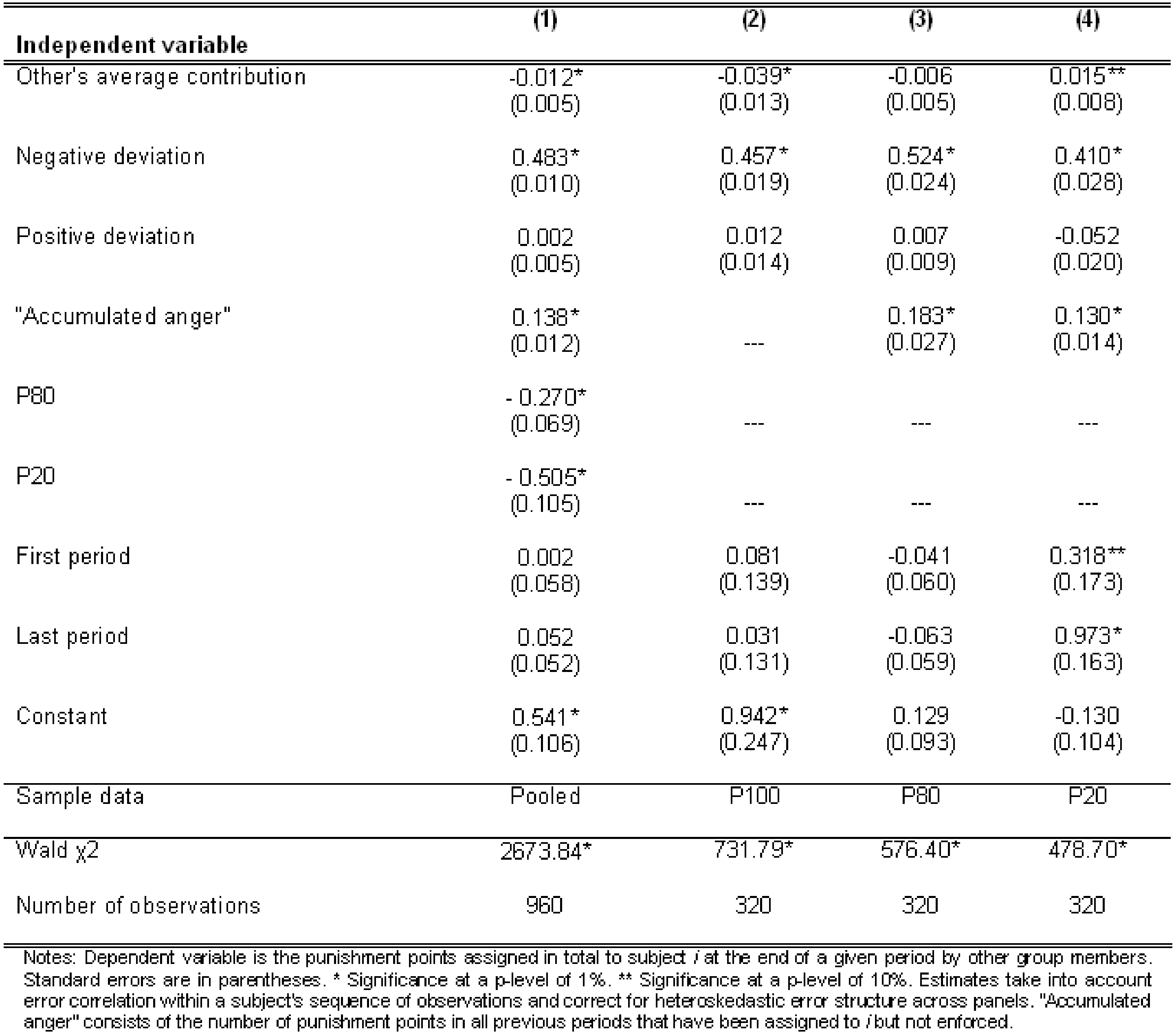
We also separate the data according to enforcement treatments and run separate regressions for each sub-sample of subjects. This allows us to examine our conjecture (see Section 3) that, because punishment is likely to be less frequent in P20 than in P80 and P100, subjects will assign punishment differently across enforcement treatments. These results are reported in Columns (2)-(4).

Beginning with the estimates of the general model in column (1), we notice that enforcement conditions do have an effect on punishment decisions: subjects in the uncertain enforcement conditions punish relatively less. We have conjectured that this effect has to do with the impact of uncertainty on the strategic value of punishment: players would be less inclined to punish if enforcement failure threats their ability to send a signal to free-riders on a consistent basis.

Looking across the treatments, there are other noticeable aspects influencing punishment decisions. First, we see that an increase in the group average contribution induces a reduction in punishment. This holds for all but the P80 treatment. Second, that punishment is mostly directed towards free-riders, those who contribute below the group average. These two results illustrate the elements of reciprocity in individuals' behaviour. Third, we find that "bygones are not bygones": the more punishment points towards a player ended up not being enforced in the history of the game – what we term "accumulated anger" –, the more punishment from others is directed to her. This can arguably indicate that punishment decisions are driven by emotions and not only by intertemporal concerns with material payoff.

All in all, these results seem to support the view that punishment is driven by a mix of backward- and forward-looking motives. On the one hand, the uncertainty over whether the willingness to punish will be materialised over the course of the game seems to weaken the strategic value of punishment in shape future behaviour; on the other hand, because of the history of free-riding that goes unpunished, it also creates frustration and increasing "anger"

Table 4: Do punishment decisions differ by treatment?



Independent variable	(1)	(2)	(3)	(4)
Other's average contribution	-0.012* (0.005)	-0.039* (0.013)	-0.006 (0.005)	0.015** (0.008)
Negative deviation	0.483* (0.010)	0.457* (0.019)	0.524* (0.024)	0.410* (0.028)
Positive deviation	0.002 (0.005)	0.012 (0.014)	0.007 (0.009)	-0.052 (0.020)
"Accumulated anger"	0.138* (0.012)	---	0.183* (0.027)	0.130* (0.014)
P80	- 0.270* (0.069)	---	---	---
P20	- 0.505* (0.105)	---	---	---
First period	0.002 (0.058)	0.081 (0.139)	-0.041 (0.060)	0.318** (0.173)
Last period	0.052 (0.052)	0.031 (0.131)	-0.063 (0.059)	0.973* (0.163)
Constant	0.541* (0.106)	0.942* (0.247)	0.129 (0.093)	-0.130 (0.104)
Sample data	Pooled	P100	P80	P20
Wald χ2	2673.84*	731.79*	576.40*	478.70*
Number of observations	960	320	320	320

Notes: Dependent variable is the punishment points assigned in total to subject *i* at the end of a given period by other group members. Standard errors are in parentheses. * Significance at a p-level of 1%. ** Significance at a p-level of 10%. Estimates take into account error correlation within a subject's sequence of observations and correct for heteroskedastic error structure across panels. "Accumulated anger" consists of the number of punishment points in all previous periods that have been assigned to *i* but not enforced.

towards those who have gotten “off the hook”.

It should not come as surprise, therefore, that punishment in the first and the last periods is statistically significantly different from punishment over the other periods of the game in the P20 treatment. Since there is no strategic incentive to punish relatively more at the end of the game, this seems to suggest that individuals are pursuing some revenge for something they deemed as unfair. Indeed, the last round is the only round in which i can punish other group members without any danger of repercussions.

Thus, the results from the regression suggest that the existence of uncertainty on whether punishment decisions will be carried out has statistically significant effect on punishment levels. The following result summarizes the findings of this section.

Result 2: *The willingness to punish free-riders is affected by the “uncertainty” over whether punishment will be actually enforced. In both uncertain treatments, individuals tend to punish less. There is a backward-looking element in punishment decisions as the more an individual has escaped being punished in the past, the more punishment is directed to her.*

4.3 Welfare Analysis

In addition to looking for differences in punishment behaviour across treatments, we now investigate how “uncertain” enforcement affects individuals’ welfare. The key difficulty in addressing this issue is that aspects that are likely to affect individuals’ utility in this experiment are not directly measured. For instance, there must be gains in utility from punishing a free-rider as much as there are losses from not being able to punish a free-rider because of an enforcement failure. We sidestep this problem for a while and following Nikiforakis (2008) we use individual earnings as a proxy measure for welfare. Using the certain enforcement treatment as a benchmark, we begin by examining whether earnings are increased in the “uncertain” enforcement conditions.

Table 5 provides an overview of how earnings look like in each enforcement condition. While average contributions are slightly lower in P80 than they are in P100, subjects in P80 have higher earnings on average than subjects in P100. We have seen that contribution levels are similar between P100 and P80 treatments, despite the fact that in P80 punishment decisions might not be enforced with a probability of 20%. While such possibility weakens the threat of punishment, the degree of enforcement was sufficient to lead to an increase in contributions over time. Since punishment assignment is costless in P80 if punishment is not enforced, it should make intuitive sense then that subjects in P80 could benefit from higher contributions without necessarily incurring punishment costs in every period. As a result subjects in P100 have lower earnings than subjects in P80.

Let us now look at what happens in P20. Compared to conditions where the sanction system in place has stronger enforcement, earnings in P20 are lower. As discussed before, for most groups in P20 punishment enforcement occurred in few occasions irregularly spaced over the ten-period sequence. This created a “disbelief” in the enforcement system, encouraging

Table 5: Earnings by enforcement treatment

Treatment	Average contribution	Average earnings after contribution	Punishment-associated costs[1]	Average earnings
P100	13.20	27.92	5.98	22.75
P80	12.67	27.60	3.11	24.54
P20	6.82	24.09	1.63	22.46

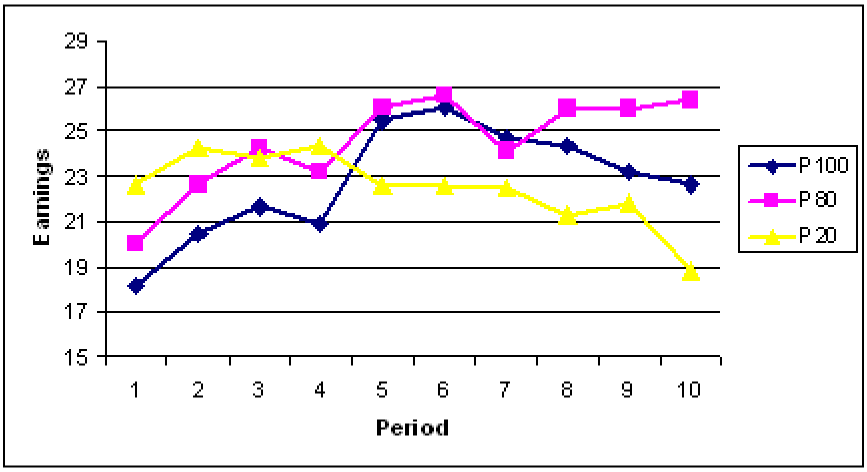
1 Average costs of punishment points given out (and enforced) to other group members and average deductions in first-stage payoff as a consequence of punishment received.

more free-riding behaviour and, consequently, leading to a decline in contributions over time. Yet, note that average earning differences between P20 and P100 are not great; this highlights, as showed by others (e.g. Fehr & Gachter (2002); Sefton *et al.* (2007)), that in public goods experiments with certain enforcement, the benefit of higher contributions may be outweighed by punishment costs.

We now turn to a formal assessment of the relation between individual earnings and enforcement conditions. Table 6 reports regressions which examine treatment effects on individual earnings. Estimated results are reported in columns (1) and (2).

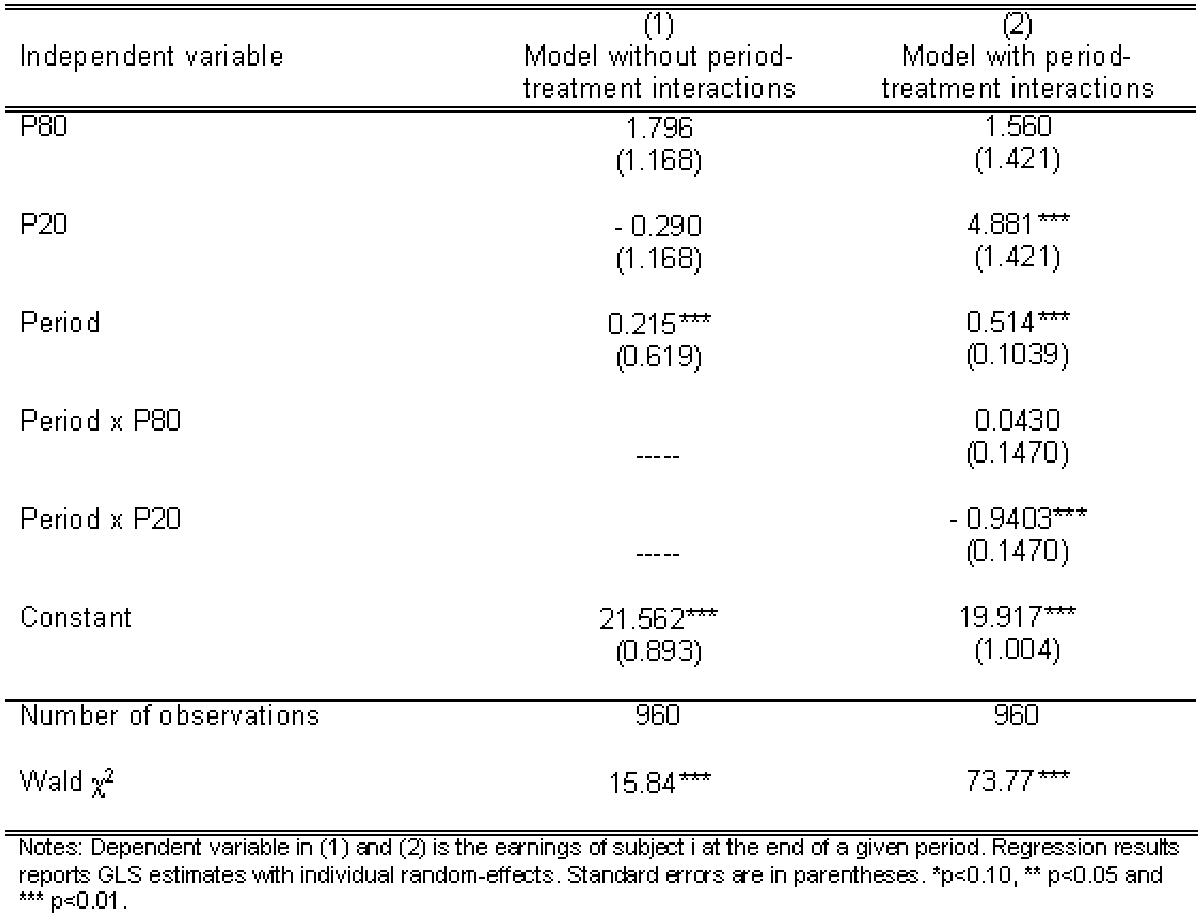
The first regression results show that none of the enforcement treatments have a significant effect on period earnings. This is in line with the general impression that, at least in the first half of the game, average earnings do not show much variation across treatments (see Figure 3).

Figure 3: Average earnings



In fact, P100 and P80 follow similar trends in terms of contribution and punishment-associated costs yielding similar earning levels throughout the game. In P20, however, average earnings closely follow the trends of P100 and P80 only at the first half of the experiment. They decrease in the second-half of the experiment as a result of the decay in contributions, while in P100 and

Table 6: Determinants of earnings (GLS Regressions)



Independent variable	(1) Model without period-treatment interactions	(2) Model with period-treatment interactions
P80	1.796 (1.168)	1.560 (1.421)
P20	- 0.290 (1.168)	4.881*** (1.421)
Period	0.215*** (0.619)	0.514*** (0.1039)
Period x P80	-----	0.0430 (0.1470)
Period x P20	-----	- 0.9403*** (0.1470)
Constant	21.562*** (0.893)	19.917*** (1.004)
Number of observations	960	960
Wald χ^2	15.84***	73.77***

Notes: Dependent variable in (1) and (2) is the earnings of subject i at the end of a given period. Regression results reports GLS estimates with individual random-effects. Standard errors are in parentheses. *p<0.10, ** p<0.05 and *** p<0.01.

P80 individuals are benefiting from higher contributions relative to the first half. The second regression adds interaction terms between period and enforcement treatments, $Period \times P80$ an $Period \times P20$. These variables try to capture time trends in P80 and P20 with respect to that in P100. The coefficient for the interaction term $Period \times P20$ is negative and significant. This just confirms what we have seen before: that there is continuous decrease in average period earnings over time in P20 as a consequence of decline in contributions, whereas earnings are kept at higher levels in P100. Note, however, that earnings in P80 are higher with respect to that in P100. This is the result of a more significant increase in contributions and that in some period individuals in P80 need not incur any punishment cost. Result 4 summarizes this finding.

Result 4: *The highest welfare level, measured by accumulated earnings, is found in P80. While punishment enforcement is not certain in P80 like it is in P100, individuals in P80 condition benefit from higher contributions without incurring punishment costs in every period.*

5 Conclusions

This paper reports an experiment examining the effects of uncertain enforcement of punishment on cooperative and punishment behaviour in a public good setting. In each period of the game, punishment decisions are enforced with a known probability p. The game is played under three treatment conditions, which differ only by the value of p (1, 0.8, or 0.2).

One of the findings is that punishment opportunities do not promote cooperative behaviour when enforcement is perceived as "weak". In this case, average contributions start at around half of subjects' endowment and keep declining over time. This contrasts with the levels of cooperative behaviour observed in the treatment where punishment enforcement is perceived as "strong": average contributions are raised and sustained at a high level. This result is somewhat comforting as it suggests that a sanctioning system with some degree of "imperfectness" can still induce cooperative behaviour in social dilemma situations. It also indicates that the deterrence effect of a sanctioning system operates through the perception it induces regarding either detection or enforcement likelihood. This result is in line, for example, with the evidence that income tax compliance increases when taxpayers are simply threatened to have their income reports "more closely examined" (see Slemrod *et al.* , 2001). Tax compliance, which is a form of cooperative behaviour, is promoted not by a threat of more severe punishment, but by inducing a change in the perceived likelihood of detection.

Another finding is that there is a backward-looking element in punishment decisions as the more an individual has escaped being punished in the past, the more punishment is directed to her. Furthermore, punishment of free-riders and low contributors in general is more intense at the beginning and the end of the game. While this could be rationalized as a compromise between strategic (reputation building) and emotional (vindictiveness) components of individual's decision making, it is still unclear how to interpret these phenomena within a rational framework. Such end-of game effects, in particular, may have implications for the theoretical study of iterated prisoner's dilemma type of games as they hint at the existence of path-dependencies in the play of the game.

It is also observed that individuals in the P80 condition benefit from higher contributions without incurring punishment costs in every period and, as a result, accumulated earnings in P80 are higher than in P100. Having said that, the P80 and P20 enforcement treatments have no significant effect on average earnings. This serves as an indication that, while failing to lead to high contributions, a punishment mechanism with "weak" enforcement might lead to similar earnings to the treatment where enforcement is certain.

The major finding of our experiment – that, put loosely, subject's perception of the likelihood of punishment enforcement matters – raises some interesting questions: in a repeated setting, can a strong threat of punishment deter individuals from deviating from a collective optimal course of action without it being ever "demonstrated"? Can a history of "punishment" itself sustain cooperation in social dilemmas without a strongly credible threat of "punishment"? Can a threat of "punishment" efficiently induce and sustain high levels of cooperation in social dilemmas without a periodic demonstration of "punishment"? In sum, which "mixes" of threat and punishment history can induce cooperation?

It is worth noting that from a theoretic standpoint a threat should suffice. In the theory of infinitely repeated games, it is a classic result that it is possible to achieve a subgame-perfect equilibrium in which players achieve the highest payoff of all existing equilibria (Friedman, 1971). The core idea underlying this result is that a given player is persuaded to follow such perfect equilibrium strategy by *threatening* her with the strongest credible punishment. Punishment may not necessarily be history-dependent (Abreu, 1988). But while the perfect equilibrium strategy profile specifies punishment for deviations, the equilibrium path ends up not involving any imposition of punishment – the simple *threat* of punishment has a deterrence effect. Some may view results from Fehr & Gachter (2000) and many other studies as lending support to this: looking at first-period data, when there is no history of play, one can see that contributions are significantly higher in punishment treatments relative to no-punishment treatments. While in an experimental setting there is some degree of uncertainty over the size of punishment in terms of payoff, the simple threat of punishment often encourages pro-social behaviour.

But our experiment raises some questions. Its results suggest that the incentive constraints implicit in such punishment schemes may not rely only on credibility, but also on what we term here "punishment demonstration" – that punishment must be exercised upon subjects. We find that an "imperfect" sanction system (in terms of enforcement) can achieve higher levels of pro-social behaviour by simply changing, through probability manipulation, subject's *perception* about the likelihood of sanction enforcement. It is unknown, however, to what extent the efficacy of punishment in inducing cooperative behaviour depends on perceived credibility of punishment threat (probability) and the factual history of the game. We view this as of theoretical and empirical relevance. In our experiment, like in all other experimental studies on cooperation with sanction systems, threat and demonstration of punishment are entangled. To investigate the influence of these factors is a topic for further research.

References

Abreu, Dilip. 1988. On the Theory of Infinitely Repeated Games with Discounting. *Econometrica*, **56**(2), 383–96.

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

Axelrod, R. A. 1984. *The evolution of cooperation*. New York: Basic Books.

Bochet, Olivier, Page, Talbot, & Putterman, Louis. 2006. Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, **60**(1), 11–26.

Boyd, R., & Lorberbaum, J. P. 1987. No Pure Strategy is Stable in the Repeated Prisoner's Dilemma Game. *Nature*, 58–59.

Camera, Gabriele, & Casari, Marco. 2007. Cooperation among strangers: an experiment with indefinite interaction. *American Economic Review*, **99**(3), 979–1005.

Cinyabuguma, Matthias, Page, Talbot, & Putterman, Louis. 2005. Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, **89**(8), 1421–1435.

Dawes, Robyn M. 1980. Social Dilemmas. *Annual Review of Psychology*, **31**, 169–193.

Egas, Martijn, & Riedl, Arno. 2008. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of The Royal Society B-Biological Sciences*, **275**(1637), 871–878.

Ertan, Arhan, Page, Talbot, & Putterman, Louis. 2009. Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, **53**(5), 495–511.

Fehr, Ernst, & Gachter, Simon. 2000. Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, **90**(4), 980–994.

Fehr, Ernst, & Gachter, Simon. 2002. Altrustic Punishment in Humans. *Nature*, **415**(10), 137–140.

Fehr, Ernst, & Schmidt, Klaus M. 1999. A Theory Of Fairness, Competition, And Cooperation. *The Quarterly Journal of Economics*, **114**(3), 817–868.

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

Fischbacher, Urs, Gachter, Simon, & Fehr, Ernst. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, **71**(3), 397–404.

Frey, Bruno S., & Meier, Stephan. 2004. Social Comparisons and Pro-social Behavior: Testing Conditional Cooperation in a Field Experiment. *American Economic Review*, **94**(5), 1717–1722.

Friedman, James W. 1971. A Non-cooperative Equilibrium for Supergames. *Review of Economic Studies*, **38**(113), 1–12.

Gachter, Simon, Renner, Elke, & Sefton, Martin. 2008. The Long-Run Benefits of Punishment. *Science*, **322**(5907), 1510.

Glance, N.S., & Huberman, B. A. 1993. The outbreak of cooperation. *Journal of Mathematical sociology*, **17**(4), 281–302.

Greiner, Ben. 2004. *Forschung und wissenschaftliches Rechnen 2003*. Gottingen: Gesellschaft fur Wissenschaftliche Datenverarbeitung. Chap. An Online Recruitment System for Economic Experiments, pages 79–93.

Gunnthorsdottir, Anna, Houser, Daniel, & McCabe, Kevin. 2007. Disposition, history and contributions in public goods experiments. *Journal of Economic Behavior & Organization*, **62**(2), 304–315.

Guth, Werner, Levati, M. Vittoria, Sutter, Matthias, & van der Heijden, Eline. 2007. Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics*, **91**(5-6), 1023–1042.

Hardin, G. 1968. Tragedy of commons. *Science*, **162**(3859), 1243–1248.

Herrmann, Benedikt, Thoni, Christian, & Gachter, Simon. 2008. Antisocial Punishment Across Societies. *Science*, **319**, 1362–1367.

Isaac, R Mark, & Walker, James M. 1988. Communication and Free-Riding Behavior: The Voluntary Contribution Mechanism. *Economic Inquiry*, **26**(4), 585–608.

Kollock, Peter. 1998. Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, **24**(1), 183–214.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.C. 2003. Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review*, **93**(1), 366–380.

Messick, D. M., & Brewer, M. B. 1983. Solving social dilemmas: A review. *Review of personality and social psychology*, **4**, 11–44.

Nikiforakis, Nikos. 2008. Punishment and counter-punishment in public good games: Can we really govern ourselves. *Journal of Public Economics*, **92**(1-2), 91–112.

Nikiforakis, Nikos. 2010. Feedback, Punishment and Cooperation in Public Good Experiments. *Games and Economic Behavior*, **68**(2), 689–702.

Noussair, Charles, & Tucker, Steven. 2005. Combining Monetary and Social Sanctions to Promote Cooperation. *Economic Inquiry*, **43**(3), 649–660.

Sefton, Martin, Shupp, Robert, & Walker, James M. 2007. The Effect Of Rewards And Sanctions In Provision Of Public Goods. *Economic Inquiry*, **45**(4), 671–690.

Slemrod, Joel, Blumenthal, Marsha, & Christian, Charles. 2001. Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota. *Journal of Public Economics*, **79**(3), 455–483.

Trivers, R.L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology*, **46**, 35–57.

Walker, James M., & Halloran, Matthew A. 2004. Rewards and Sanctions and the Provision of Public Goods in One-Shot Settings. *Experimental Economics*, **7**(3), 235–247.

Appendix A - Proofs

Proof of **Proposition 1: Part 1.1.**

If $q = 0$, then only the first-stage is payoff-wise relevant. In this stage, the strategy $c_i = e$ is strictly dominated by $c_i = 0$, since $\pi_{i,C}(c_i = 0, c_-i = 0) > \pi_{i,C}(c_i = e, c_-i = 0)$ for both players. The game has then a unique Nash equilibrium in which both players defect: $(c_1 = 0, c_2 = 0)$, given their payoff outcomes, is the pair of strategies that maximises the utility of the self-regarding player, $u_1(.)$, and the utility of the inequity-averse player, $u_2(.)$, for all $\alpha > 0$.

Proof of **Proposition 1: Part 1.2.**

If $q > 0$, then actions in the punishment stage may have payoff consequences for both players. For the self-regarding player, imposing no punishment, i.e. $p_1 = 0$, is a dominant strategy as $p_1 = 0 = \arg\max u_1(\pi_{i,C}(c_i, c_-i) + \pi_{i,P}(p_1, p_2)$ for all cooperative and punishment strategies of the inequity-averse player. Since this is common knowledge, it is easy to see that the inequity averse player will choose no punishment, i.e. $p_2 = 0$, at the second-stage of the game. Now, the inequity-averse player will also choose no punishment, $p_1 = 0$, when the profile of actions chosen in the first stage are $\{c_1 = e, c_2 = e\}$, $\{c_1 = 0, c_2 = 0\}$, $\{c_1 = e, c_2 = 0\}$; in all these cases, the inequity-averse player cannot be better off by choosing to punish, i.e. $p_2 = \rho$. For instance, choosing $p_2 = \rho$ following $\{c_1 = e, c_2 = e\}$ is dominated by $p_2 = 0$ since $r2e > r2e - (\rho + \beta(\rho + l\rho))$ for all $\beta > 0$. But the inequity averse player will punish following the first-stage pair of strategies $\{c_1 = 0, c_2 = e\}$ if the final payoff when she assigns punishment to the self-regarding player is larger than the final payoff of not doing so, that is, if $\pi_1(c_2 = e, p_2 = \rho) > \pi_1(c_2 = e, p_2 = 0)$. This amounts to the following condition

$$(re - \rho) - \alpha max\{(e(1 + r) - l\rho) - (re - \rho), 0\} > re - \alpha max\{(e(1 + r)) - (re), 0\}$$

which holds only if $\alpha > \frac{1}{l-1}$. Note that in this case the threat of punishment can only induce the self-regarding type to cooperate, and she will have no incentive to deviate from that, if the final payoff from cooperating is larger than the expected final payoff from free riding in the first stage and getting punished in the second stage

$$r(2e) > (1 - q)[e(1 + r)] + q[e(1 + r) - l\rho]$$

which holds only if $q > q^* = \frac{e(1-r)}{l\rho}$. This completes the proof.