Journal of Behavioral and Experimental Economics 106 (2023) 102063

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

Punishment credibility and cooperation in public good games

Sergio Almeida

Department of Economics, University of Sao Paulo, Sao Paulo, SP, 05508-010, Brazil

ARTICLE INFO

Keywords:
Public good experiments
Punishment
Enforcement
Social norms
Decision-making under risk

ABSTRACT

This paper investigates the efficacy of a punishment mechanism in promoting cooperative behaviour in a public goods game when punishment enforcement is risky. Experimental studies have found that a sanctioning system can induce individuals to adopt behaviour deemed socially acceptable in such games. Our experiment qualifies this result by showing that a sanctioning system can only promote cooperative behaviour if subjects perceive punishment enforcement as a high-probability event. This result supports the view that sanctioning systems can only induce people to comply with social norms that enhance efficiency if such systems are sufficiently credible. We also find that the more punishment points towards a player were not being enforced in the history of the game, the more punishment from others was directed to them. This suggests that bygones are not bygones and that punishment behaviour attempts to compensate for the history of free-riding that goes unpunished.

1. Introduction

There has been a long-standing interest among social scientists and biologists regarding how self-interested individuals can be induced to cooperate in social dilemmas—situations where self-interested behaviour is at odds with collective interest.[1] Investigation of this question has generated a large number of experimental studies on how to increase cooperation in public good games. Many mechanisms have been investigated. Cinyabuguma et al. (2005), Guth et al. (2007), Isaac and Walker (1988) and Masclet et al. (2003) have shown that pre-play communication, threat of expulsion, or even symbolic disapproval can all boost cooperative behaviour. Since the seminal work of Fehr and Gachter (2000), a particular mechanism that has received considerable attention is enabling subjects to financially penalise others. Several studies have shown that this monetary sanction system is significantly effective in increasing and maintaining contributions (Ambrus & Greiner, 2012; Anderson & Putterman, 2006; Bochet et al., 2006; Camera & Casari, 2007; Ertan et al., 2009; Gachter et al., 2008; Masclet et al., 2003; Nikiforakis, 2008; Noussair & Tucker, 2005; Sefton et al., 2007).

In this paper, we investigate how efficaciously such a punishment mechanism promotes cooperation in a more realistic setting where punishment may not necessarily be enforced. In particular, we look into a public good game in which the same group of subjects make their contribution decisions knowing that others' decision to punish them will not be enforced with a probability $p > 0$. We compare how high and low enforcement probabilities affect cooperation and punishment behaviour in such risky environments with behaviour in an environment in which punishment enforcement is certain.[2]

This design corresponds to institutional punishment situations in which individuals who have violated social norms may still go unpunished because of the sanctioning system's imperfections. After all, across all societies, imperfect monitoring, corruption, and economic and cultural factors may, to some extent, undermine the effectiveness of such a sanctioning system (see, e.g., Gachter and Schulz (2016)). This design also serves to investigate the role of enforcement in attaining cooperative outcomes in environments where there are strong reputation concerns. The folk theorem in repeated games states that the mere threat to punish noncooperators can maintain cooperation in prisoner's dilemma situations. This result suggests that the incentive constraints implicit in such punishment schemes rely only on punishment credibility, given the material incentives for cooperation, and not on whether and how punishment is exercised upon players. Our experiment will test this theoretical prediction.

Our design closely follows that of Fehr and Gachter (2000). In particular, we examine 10-period public good games with punishment opportunities, and we adopt the same payoff structure. There are, however, two different aspects in our experiment. First, Fehr and

E-mail address: sergio.almeida@usp.br.

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

[2] Free riding is always risky in the presence of punishment opportunities as others may or may not punish such behaviour. Adding risk to punishment *enforcement* adds an extra layer of risk to free-riding.

https://doi.org/10.1016/j.socec.2023.102063
Received 21 December 2021; Received in revised form 25 June 2023; Accepted 29 June 2023
Available online 10 July 2023
2214-8043/© 2023 Elsevier Inc. All rights reserved.

Gachter used a convex punishment cost function while, following Page et al. (2005), Bochet and Putterman (2009) and Sefton et al. (2007), we adopt a linear one.[3] Second, in our experiment group, members' contributions are identified by an ID number when disclosed on the computer screen and are always listed in the same ID column position, rather than randomly reassigned every period. This was implemented to allow subjects to track group members' contributions more easily. This should mitigate indiscriminate punishment and reduce noise in the data. This design feature also facilitates investigation of the extent to which punishment decisions are influenced by contributions in previous rounds. A major dimension in which we varied the design was the credibility of the sanctioning system: we employed (i) a low credibility condition, in which the probability of punishment decisions actually being carried out is 0.2; (ii) a high credibility condition, in which the probability of punishment decisions actually being carried out is 0.8; (iii) a benchmark certain punishment enforcement condition used in Fehr and Gachter (2000) and several other experimental papers.

We have two major findings. First, when punishment enforcement is perceived by the individuals as a low-probability event, the threat of punishment cannot raise and sustain high levels of cooperation. But, as expected, when punishment enforcement is certain or imperfect but perceived as a high-probability event, punishment opportunities serve as an effective deterrent, raising and sustaining high levels of contributions throughout the game. Second, low contributors are more intensely punished when enforcement of punishment decisions is a low-probability event, with punishment of free-riders and low contributors being generally more intense at the beginning and end of the game.

Our results suggest that the credibility of enforcement is a key issue: a sanctioning system with some degree of imperfection can still induce cooperative behaviour in a social dilemma situation as long as the perceived probability of punishment is high and sustained by a relatively high number of actual punishments. They also suggest 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 them.

Our paper complements the current body of research on the "robustness" of the punishment mechanism used by Fehr and Gachter (2000, 2002). Recent papers have demonstrated that punishment may not help 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 the circumstances under which a punishment mechanism can induce cooperation in social dilemmas.

The remainder of this paper is organised as follows: Section 2 describes the experiment design. Section 3 reports the results. Section 4 concludes.

2. Experimental design

2.1. Basic design

The overall design consists of a public good experiment, comprising punishment opportunities with three treatment conditions (see Table 1). In one treatment (P100), punishment decisions are enforced with certainty. This treatment builds on the standard design for the public goods game with punishment, as in the seminal work by Fehr and Gachter (2000). The remaining two treatments differ according to the probability of enforcement of punishment decisions: one treatment with a "high" probability of enforcement (P80) and the other with a "low" probability of enforcement (P20) in which punishment decisions are carried out with a probability of 0.8 and 0.2, respectively.

Table 1
Treatment conditions.

	Enforcement of punishment decisions with probability p		
	P100 (p = 1)	P80 (p = 0.8)	P20 (p = 0.2)
PGG with punishment opportunities (10 periods)	8 groups of size 4	8 groups of size 4	8 groups of size 4

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 (PGG) with punishment for ten periods. We randomly assigned treatment conditions to sessions, so that all subjects in the same session face the same treatment condition.

2.2. Payoffs

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 or three stages depending on the treatment assigned to the session.

In the first stage, subjects are required to simultaneously decide how much of their endowment to invest in the public account, say c_i, and, consequently, how much of it to invest in the private account, $20 - c_i$. Each Rubi allocated to the private account has a return of 1 for the depositing 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, their income from the public account and their first-stage earnings (π), which is given by:

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

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-rivalry and non-excludability, can be seen as contributions to a public good.

In the second stage, participants are informed of their group members' contribution decisions 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.[4]

[3] We do so because we want to make the impact-to-cost ratio effectiveness of punishment constant with the severity of punishment. Also, a variable marginal cost of punishment may make payoff calculation more difficult to be performed.

[4] A subject cannot have their first-stage income, π^1, reduced below zero as a result of the punishment given by others. Nevertheless, as they always carry the cost of punishment incurred, their period income may end up negative depending on the total number of "deduction" points received and assigned. As in Fehr and 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.

In the P100 condition, punishment decisions are carried out with certainty. In this case, subject i's end-of-period payoff is given by:

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

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.

In the P80 and P20 conditions, subjects make their punishment decisions at the end of the second stage with the understanding that such decisions may not necessarily be carried out; they will be so with a probability $p < 1$, with $p = 0.8$ if P80 and $p = 0.2$ if P20. This probability is the same for all 10 periods of the experiment's session.[5] When punishment decisions are implemented, a subject's final earnings in the period are calculated by the equation in (2). When they are not carried out, a subject's end-of-period earnings are equal to their earnings in the first stage.

Punishment is neither disclosed nor costly unless it is enforced. Thus, the information disclosed at the end of each period depends on the enforcement state: if the punishment is not enforced, subjects are shown their final earnings, which are equal to their earnings from the first stage. In case the punishment is enforced, they are shown (a) the total cost of the punishment points assigned, (b) the punishment points 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 96 subjects in this experiment, recruited from an email pool of undergraduate students at the University of Nottingham. None of them had previously participated in a public good experiment. There were six sessions with 16 participants each. Upon arrival, subjects received an ID number and were assigned to a desktop computer.

To ensure subjects' understanding of the game's structure and payoff determination, 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 were 50 min long.

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. Experimental results

3.1. Cooperative behaviour

We start by examining contribution patterns across treatments.[6] Fig. 1 presents a line graph of median contribution to the Public Account over the 10 periods for each treatment condition.

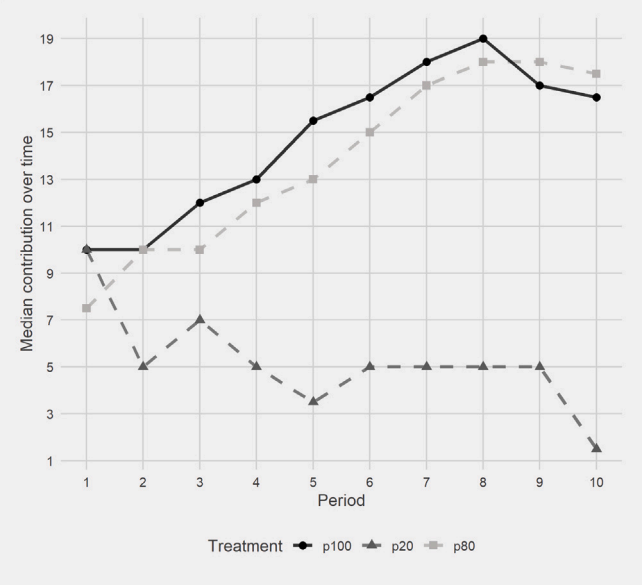

Fig. 1. Median contribution, by punishment enforcement treatment.

Contributions under the P100 condition, for instance, were in line with previous experimental findings: they started at roughly half of subjects' endowments and kept increasing over time. This result confirms that the existence of punishment can improve cooperation over time. Additionally, and perhaps more importantly, it suggests that the ability of punishment to sustain cooperation is unaffected by knowledge of contribution histories.

There was, however, a clear separation in contributions between the uncertain enforcement treatments. While median contributions in the P80 condition increased over time, closely following contributions in the certain enforcement condition, contributions in the P20 condition were noticeably lower, and on a divergent path, than in the P80 condition. While median contributions in the P20 condition started higher than contributions in the P80 condition, they kept decreasing from the second period on, while contributions in the P80 treatment increased 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 a non-parametric test. A pairwise Mann–Whitney test between treatments for all periods shows us two things. First, that there are no significant differences in mean contribution of groups in the P100 and P80 treatments (P100 vs P80: $Z(160) = 1.23, P = 0.21$). Second, that there are statistically significant differences between mean contribution of groups in P20 and either P100 or P80 treatments (P100 vs P20: $Z(160) = 6.15, P < 0.01$, P80 vs P20: $Z(160) = 6.32, P < 0.01$).

We now turn to a more formal analysis of the data by running a regression of individual contributions on treatment and individual variables. The panel structure of the data allowed us to handle some degree of individual heterogeneity and obtain more consistent estimates of treatment effects.

We estimated an empirical model relating each contribution to the individual and parameters of the game that largely follow a common specification in these studies (e.g. Anderson & Putterman, 2006; Lohse & Waichman, 2020; Nikiforakis, 2008). However, our econometric specification also included lagged variables that sought to capture recursive elements in contribution decisions.[7] The baseline model then

[5] A physical bag containing 10 balls, numbered from 1 to 10, was presented to subjects in a given session after they read the instructions so they could inspect the fairness of the balls. At the end of the second stage in *each* period of the P20 and P80 treatment conditions, a subject in each group was invited to draw a ball from the bag to determine whether punishment decisions were enforced. Subjects in the P20 (P80) treatment were told that punishment decisions would be carried out if a ball numbered 9 or 10 (3 through 10) was drawn.

[6] The dataset is available at Almeida (2021).

[7] The underlying reason for this is hardly controversial: in repeatedly played games, individuals tend to reciprocate actions of other players; this

Table 2
Determinants of public good contributions.

	(1)	(2)	(3)	(4)
Others avg contribution ($t-1$)		0.961***	0.961***	0.942***
		(0.015)	(0.016)	(0.016)
Punishment received ($t-1$)		0.183***	0.185***	0.167**
		(0.060)	(0.059)	(0.085)
Sum of enforced punish.		0.064	0.062	−0.017
		(0.352)	(0.351)	(0.397)
P20	−6.381**	−0.809*	−0.773	−2.046**
	(2.645)	(0.435)	(0.491)	(0.878)
P80	−0.528	0.558*	0.566*	−0.885
	(2.358)	(0.323)	(0.331)	(0.826)
Accumulated UPP			−0.007	−0.011
			(0.034)	(0.028)
Constant	13.197***	0.818	0.810	2.564***
	(1.976)	(0.741)	(0.758)	(0.819)
Controls	No	Yes	Yes	Yes
Observations	960	960	960	960

Note: Each column presents results from an OLS regression with robust standard errors clustered at the group level reported in parentheses. Dependent variable is contribution at the individual level. Estimates are heteroscedastic-consistent. Dummy variables for groups and sex are included. * $p < 0.1$;** $p < 0.05$;*** $p < 0.01$.

had the following specification:

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

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 the individual 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 time trends and the variation related to some subject-specific attributes (gender, ethnicity etc.).

Table 2 reports the results of the OLS regressions of several specifications of the model in (3). Column (1) is just a regression of contribution on treatments dummies to get pure enforcement treatment effects. Column (2) is the baseline model. Columns (3) and (4) are specifications that add accumulated unenforced punishment points and additional controls for time trends and their interaction with treatments.

Contributions were, on average, positively affected by retaliatory behaviour from others in the past: the number of punishment points actually received in the previous period as well as in the history of the game had both significant and positive effects on contributions. Of interest in the results was the parameter in front of the dummy variable $P20$; it represented 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 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, the model's parameter estimates support the raw results depicted in Fig. 1.

Therefore, as was apparent in Fig. 1, there were significant differences in contribution estimates between "high" and "low" probability of enforcement conditions. The mere knowledge that sanctions may be imposed to punish those regarded as free-riders may not induce cooperative behaviour if punishment enforcement is viewed as "weak". Based on the non-parametric and regression analysis one can conclude that:

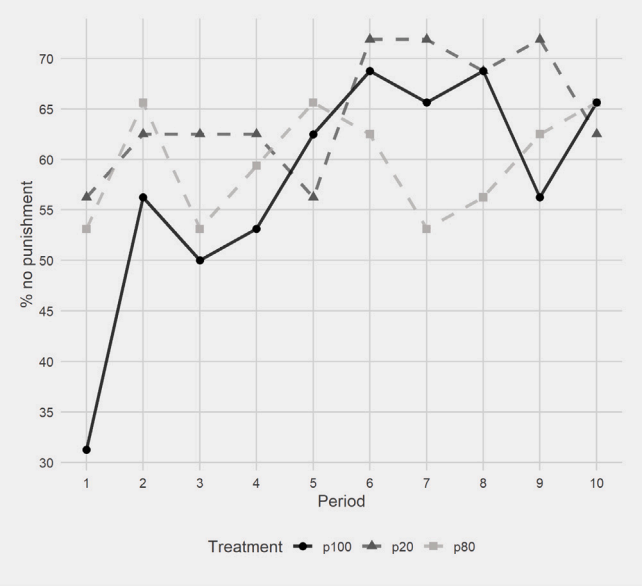

Fig. 2. Fraction of subjects who assign zero punishment points, by punishment enforcement treatment.

Result 1. *The threat of punishment does not promote cooperative behaviour if enforcement is perceived as a low-probability event.*

3.2. 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.

Fig. 2 presents the frequency of individuals who assign no punishment. Two things are worth noting: first, that there was a considerable amount of "free-riding" behaviour on punishment efforts across treatments. In most periods, less than half of the subjects exercised the option to punish. Second, there was more punishment of individuals in the first period of the certain enforcement condition than in the risky enforcement conditions.

produces behaviour that is largely reactive and influenced by past outcomes (see, e.g. Fischbacher et al. (2001), Frey and Meier (2004) and Gunnthorsdottir et al. (2007)).

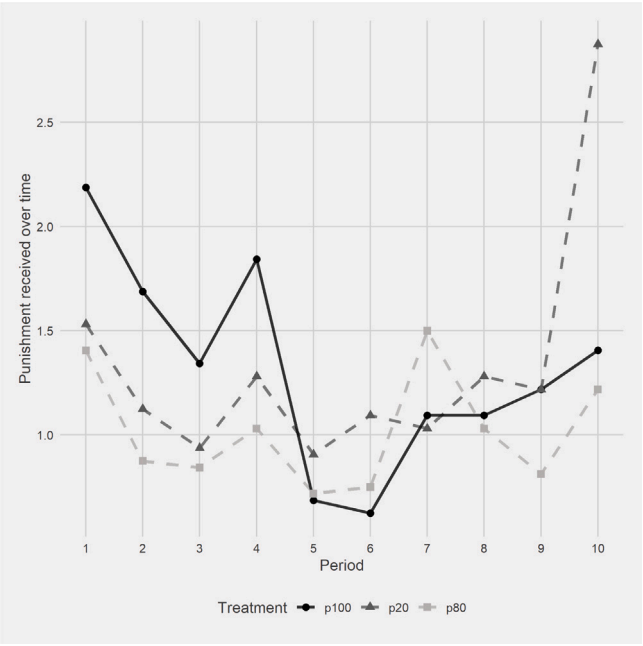

Fig. 3. Average punishment received by treatment condition, per period.

Table 3
Determinants of punishment behaviour.

	(1)	(2)
Others average contribution	−0.013	−0.010
	(0.012)	(0.011)
Positive deviation	−0.008	−0.010
	(0.019)	(0.020)
Negative deviation (abs)	0.543***	0.540***
	(0.048)	(0.049)
Accumulated unenforced punish. points	0.108***	0.101***
	(0.024)	(0.024)
P80	−0.215	−0.223*
	(0.135)	(0.134)
P20	−0.538***	−0.620***
	(0.206)	(0.203)
P20 × Last period		1.246
		(0.769)
P80 × Last period		0.132
		(0.589)
Constant	0.660***	0.625***
	(0.237)	(0.225)
Controls	Yes	Yes
Observations	960	960

Note: Each column presents results from an OLS regression with robust standard errors clustered at the group level reported in parentheses. Dependent variable is the punishment points assigned in total to a subject at the end of the period by other group members. * $p < 0.1$;** $p < 0.05$;*** $p < 0.01$.

A possible interpretation of this first-period difference in punishment between treatment conditions is that subjects in the certain enforcement condition tried to discipline behaviour from the beginning by signalling "toughness" with free-riders and low-contributors. This strategic reputation building would be mitigated among subjects in P80 and P20 enforcement conditions. One reason for that is that knowing that their punishment decisions may fail to be enforced, they might have believed that the cost of stronger signals early in the game might not be compensated later by higher cooperation levels. This is likely to be the case of a forward-looking subject who believed that punishment would only work if it was enforced frequently, in which case it would be rational not to punish in P20 even though unenforced punishment was costless.

The line graph of average punishment points in Fig. 3 show other interesting aspects of punishment behaviour in each enforcement condition. Each dot on the line plots the punishment points assigned in a given period.

Visual inspection of this figure suggests two things. First, that there are some sort of first- and last-period effects. Note that the willingness to punish (mostly directed to free-riders and low contributors) is stronger at the beginning and at the end of the game. Second, that free riders are most of the time more intensively punished in P20 than they are in P80 (and P100 too, in the second half of the game).

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 parameters of the game. The baseline empirical model has the following form:

$$P_{i,t} = \beta_0 + \beta_1 \bar{c}_{-i,t} + \beta_2 POSDEV + \beta_3 NEGDEV + \\ + \beta_4 UPP + \beta_5 P80 + \beta_6 P20 + \mathbf{z}'\alpha + u_{i,t} \quad (4)$$

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. One of those variables is zero depending on whether i's contribution is either above (or equal) or below the other's contribution. UPP denotes all the unenforced punishment points assigned i 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 tries 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. Parameter estimates of the model (4) are presented in Table 3.

Beginning with the estimates of the general model in column (1), we notice that not all enforcement conditions do have an effect on punishment decisions: only subjects in the riskier enforcement conditions (P20) punished relatively less. We have conjectured that this effect has to do with the impact of risk on the strategic value of punishment: players would be less inclined to punish if enforcement failure threatens their ability to consistently send a signal to free-riders, which is clearly the case of the P20 treatment condition.

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, but the effect is not significant. 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 were not being enforced in the history of the game, the more punishment from others was directed to them. This might be simply picking up an adjustment in the severity of punishment to compensate for previous unsuccessful attempts to punish others. Alternatively, this can also be capturing other group members' sentiment of anger for those who escaped being punished, arguably indicating that punishment decisions are driven by emotions and not only by intertemporal concerns with material payoff.

While it is beyond the scope of this paper to examine theoretically how the enforcement probability causally affects punishment behaviour, we conjecture that the severity of punishment decreases with the probability of enforcement. The mechanism underlying this behaviour pattern would have to do with the interplay between

backward- and forward-looking motives driving punishment decisions. On one hand, the higher is the probability of punishment enforcement, the stronger is the threat credibility of punishment to discipline future behaviour. In this case, there is little need to exercise punishment upon others to enforce cooperation as the game proceeds towards the end. On the other hand, when the probability of punishment enforcement is low, free-riding behaviour is more likely to go unpunished—not because of unwillingness to punish, but because "luck" got free-riders "off the hook". Punishment, in this case, would be harsher to rectify an accumulated history of unpunished free-riding. Hence, when the probability of enforcement is low, backward-looking motives would dominate forward-looking motives, causing an increase in punishment. The opposite would happen when the probability of enforcement is high.

Our results indeed support the view that a history of free-riding that goes unpunished creates frustration towards those who have gotten "off the hook". It should not come as a 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 unfair. Indeed, the last round is the only round in which a subject can punish other group members without any danger of repercussions.

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

Result 2. *The willingness to punish free-riders is affected by the "uncertainty" over whether punishment will actually be enforced. In the more uncertain treatment, 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 them*

4. Concluding remarks

This paper reports an experiment examining the effects of risky enforcement of punishment on cooperative and punishment behaviour in a public good game.

One of the findings is that punishment opportunities do not promote cooperative behaviour when punishment enforcement is perceived as weakly credible. 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 strongly credible: 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 aligns, for example, with the evidence that income tax compliance increases when taxpayers are simply threatened to have their income reports "more closely examined" (see, e.g., Kleven et al., 2011; 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 them. 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 rationalised as a compromise between strategic (reputation building) and emotional (vindictiveness) components of individual's decision-making, it is still unclear how these phenomena can be interpreted 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.

The major finding of our experiment – that the subject's perception of the likelihood of punishment enforcement matters both for cooperation and punishment behaviour – raises some interesting questions. For instance, to what extent the efficacy of punishment in inducing cooperative behaviour depends on perceived credibility of punishment threat (probability) and the factual history of punishment over 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.

Data availability

Data will be made available on request.

Acknowledgements

I thank Robin Cubitt and Martin Sefton for their helpful comments. I gratefully acknowledge financial support from CAPES (Brazilian Federal Agency under the Ministry of Education for support and evaluation of graduate education) and The Centre for Decision Research and Experimental Economics at the University of Nottingham. I would like to thank Iago Vieitez for his excellent research assistance. I would like to thank JBEE's Associate Editor Angela Sutan for the excellent feedback on this paper. Finally, I am very grateful to two anonymous reviewers for their careful reviews and constructive suggestions that improved the quality of the paper. All remaining errors are my own.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socec.2023.102063.

References

Almeida, S. (2021). Punishment credibility and cooperation in public good games. http://dx.doi.org/10.6084/m9.figshare.23680974, Figshare Repository [Dataset], https://figshare.com/s/305faa87ffadc2c31de9.

Ambrus, Attila, & Greiner, Ben (2012). Imperfect public monitoring with costly punishment: An experimental study. *American Economic Review*, *102*(7), 3317–3332.

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 Behaviour and Organization*, *60*(1), 11–26.

Bochet, Olivier, & Putterman, Louis (2009). Not just babble: Opening the black box of communication in a vooluntary contribution experiment. *European Economic Review*, *53*(3), 309–326.

Boyd, R., & Lorberbaum, J. P. (1987). No pure strategy is stable in the repeated prisoner's dilemma game. *Nature*, (327), 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.

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.

Gachter, Simon, Renner, Elke, & Sefton, Martin (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510.

Gachter, Simon, & Schulz, Jonathan F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, *531*(7595), 496–499. http://dx.doi.org/10.1038/nature17160.

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

Gunnthorsdottir, Anna, Houser, Daniel, & McCabe, Kevin (2007). Disposition, history and contributions in public goods experiments. *Journal of Economic Behaviour and 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.

Kleven, Henrik Jacobsen, Knudsen, Martin B., Kreiner, Claus Thustrup, Pedersen, Søren, & Saez, Emmanuel (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica*, *79*(3), 651–692. http://dx.doi.org/10.3982/ECTA9113.

Kollock, Peter (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, *24*(1), 183–214.

Lohse, J., & Waichman, I. (2020). The effects of contemporaneous peer punishment on cooperation with the future. *Nature Communications*, *11*(1), 1815.

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.

Page, Talbot, Putterman, Louis, & Unel, Bulent (2005). Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency*. *The Economic Journal*, *115*(506), 1032–1053. http://dx.doi.org/10.1111/j.1468-0297.2005.01031.x.

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.